

Statistical methods for NHS incident reporting data

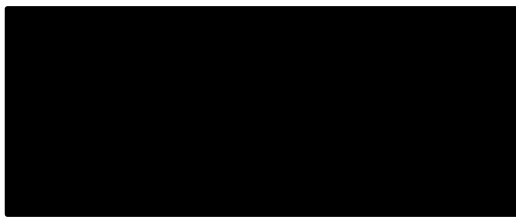
Christopher Paul Mainey

Thesis submitted for the degree of
Doctor of Philosophy

University College London (UCL)

I, Christopher Paul Mainey, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed

A large black rectangular box used to redact the signature of Christopher Paul Mainey.

29/03/2019

For Claire, Evelyn & Sammy. You can do all things through Him who gives you strength.

Abstract

The National Reporting and Learning System (NRLS) is the English and Welsh NHS' national repository of incident reports from healthcare. It aims to capture details of incident reports, at national level, and facilitate clinical review and learning to improve patient safety. These incident reports range from minor 'near-misses' to critical incidents that may lead to severe harm or death. NRLS data are currently reported as crude counts and proportions, but their major use is clinical review of the free-text descriptions of incidents. There are few well-developed quantitative analysis approaches for NRLS, and this thesis investigates these methods.

A literature review revealed a wealth of clinical detail, but also systematic constraints of NRLS' structure, including non-mandatory reporting, missing data and misclassification. Summary statistics for reports from 2010/11 – 2016/17 supported this and suggest NRLS was not suitable for statistical modelling in isolation.

Modelling methods were advanced by creating a hybrid dataset using other sources of hospital casemix data from Hospital Episode Statistics (HES). A theoretical model was established, based on 'exposure' variables (using casemix proxies), and 'culture' as a random-effect.

The initial modelling approach examined Poisson regression, mixture and multilevel models. Overdispersion was significant, generated mainly by clustering and aggregation in the hybrid dataset, but models were chosen to reflect these structures. Further modelling approaches were examined, using Generalized Additive Models to smooth predictor variables, regression tree-based models including Random Forests, and Artificial Neural Networks. Models were also extended to examine a subset of death and severe harm incidents, exploring how sparse counts affect models.

Text mining techniques were examined for analysis of incident descriptions and showed how term frequency might be used. Terms were used to generate latent topics models used, in-turn, to predict the harm level of incidents.

Model outputs were used to create a 'Standardised Incident Reporting Ratio' (SIRR) and cast this in the mould of current regulatory frameworks, using process control techniques such as funnel plots and cusum charts. A prototype online reporting tool was developed to allow NHS organisations to examine their SIRRs, provide supporting analyses, and link data points back to individual incident reports.

Impact statement

Medical error and unsafe care have been recognised globally as leading causes of harm (Jha et al., 2013). Estimates from the NHS suggested that between 6,000 and 25,500 NHS patients may suffer harm each year, as a direct result of healthcare interventions (Donaldson, 2002, Hogan et al., 2012). Learning from error is a cornerstone of patient safety and commonly aims to defend against future errors. Measurement of error is a major issue, with a reliance on clinical review and audit, but incident reporting has been recognised as a key process to help organisations learn from error (Donaldson, 2002, Berwick, 2013).

The National Reporting and Learning System (NRLS), the English and Welsh NHS' repository for incident reports, has been examined in this thesis. Despite its short-comings (Pham et al., 2013), the system has provided a wealth of information for the NHS and led to, or supported, many changes in practice (Panesar et al., 2009). Case-note review methods are well established, but they are resource intensive and can review only a small fraction of reports, leaving the majority unused. Current reporting methods use comparatively crude counts or rates, and this thesis advances methods for casemix-adjustment and more accurate prediction/monitoring.

The results of casemix models have been cast in the framework currently used by NHS regulator the Care Quality Commission (CQC), and with the completion of this thesis, will be presented to CQC for their use. Models have been developed with advice from NHS Improvement (NHSI) and are due to be presented to them during 2019 to aid their regulatory and publishing approaches. The validity of these models will also be investigated by working with NHSI to determine if models identify organisations of specific interest.

Methodological conclusion from statistical methods in this project have also been shared with NHS Digital as feedback on their mortality indicator development.

Publications have not been pursued during this work, due to time constraints and information governance arrangements. After submission, and with stakeholder approval, papers detailing methods for statistical modelling, the identification of outliers and adjustments for overdispersion, and text mining techniques are planned for publication in peer-reviewed journals and via Arxiv.

Outputs have been presented in the form of conference posters for the Royal Statistical Society and Patient Safety Congress, as well as numerous local events at UCL and University Hospitals Birmingham NHS Foundation Trust (UHB). I have published the funnel plot methods as an R package: 'FunnelPlotR' available from the CRAN repository, with 3722 downloads at time of submission.

This work was sponsored by UHB, with the intention of developing reporting tools, and spreading learning throughout the Informatics department. These tools are now ready for launch, pending approval from NHSI, and will allow approximately 50 NHS organisations to examine their data and aid learning. Dissemination of skills and knowledge within the Informatics department is progressing well, and I have used the skills I've developed to deliver training, develop statistical literacy in the team, write training material and examples, and lead on the adoption of R. It has also provided a route to access the growing NHS R community and contribute to national NHS learning.

Acknowledgements and authorship statement

I have conducted all stages of the work presented in this thesis including literature review, data processing, SQL coding, statistical programming in R, SAS, Python, and in module development. As a student sponsored and employed by an NHS department, my work interfaces with that of other colleagues. Any reuse of data processes, data flows/warehouses, or code is explicitly acknowledged below.

Supervision: Professor Nick Freemantle (Primary Care and Population Health/Comprehensive Clinical Trials Unit - UCL), Dr Milena Falcaro (Primary Care and Population Health - UCL), and during the initial months, Professor Daniel Ray (NHS Digital).

I am so grateful for the faith you had in me, the guidance and support you've provided over the last few years. I'm particularly grateful for how you've helped me meet commitments to my study, my work and my family.

PhD adviser and Clinical Lead: Professor Simon Ball (University Hospitals Birmingham NHS Foundation Trust)

Sponsorship: This project has been generously sponsored by University Hospitals Birmingham NHS Foundation Trust (UHB) Health Informatics department, and I am grateful for UHB's support, infrastructure, governance and expertise that has made the project possible.

I would like to express my thanks to those who've provided comments and feedback, given advice, or helped me with proof-reading drafts: my supervisors, Dr David McNulty, Dr Idunn Alpaar and Michelle Lockett (all UHB), Dr Frances Healy, Dr Julia Abernathy, and Noreen Gul (NHS Improvement), Dr Helen Hogan (London School of Hygiene and Tropical Medicine), and Barbara Mainey.

Data sharing agreements and the information governance arrangements with NHS Improvement and NHS Digital were arranged and managed by Michelle Lockett and Jessica Dickinson (UHB).

Data sources other than NRLS were access from data warehousing processes performed by UHB's monthly data processes. These duties are performed by various members of the team, but are managed by Dr Sarah Wang and Jessica Dickinson.

Python code for overdispersed funnel plot limits, for use in the online reporting tool (Chapter 10), was adapted from original code by Dr Ping Sun (AstraZenica). This tool also featured encryption processes adapted from processes and code written by Jonathan Lundy (UHB).

I would like to express my thanks, in particular, to my colleague, mentor and unofficial supervisor Dr David McNulty. Your support and encouragement have been catalysts for my development as a statistician. You've gone above and beyond the call of duty with reviewing my drafts, and I am very grateful for it.

To my family: my wonderful wife Claire, and my children Evelyn and Sammy. Your love, constant support, and sacrifices over the last few years have spurred me on to finish, and we have achieved it together. I hope, in the years to come, it might inspire you to pursue the challenges and opportunities in front of you. Be assured of my love and support in the same measure that you have given me. I love you beyond all measure. Thank you.

Finally, to my Lord and Saviour Jesus Christ. It does not appear popular, in the parts of academia that I've experienced, to profess to faith. I do not seek to evangelise, but I do recognise this opportunity for the gift that this PhD is, and I thank Him for it. It arrived immediately after asking for it in prayer, and when finances and the ability to support my family should have made it impossible, my needs were met. Jehovah Jireh!

Glossary of terms abbreviations, and conventions

All abbreviations are defined on their first use in the text and are occasionally restated in full to remind the reader of the terms where it seems appropriate.

Adverse events	A broader term than ‘incident,’ referring to unintended events. There are various definitions of adverse events depending on setting and context.
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
CQC	Care Quality Commission
Cusum	Cumulative Summary (control charts)
DPSIMS	Development of Patient Safety Incident Management Systems
GAM	Generalized Additive Model
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
HES	Hospital Episode Statistics
HSMR	Hospital Standardised Mortality Ratio
Incident	Any unintended or unexpected event that could have, or did, lead to harm for one or more patients or staff receiving NHS-funded healthcare (Sari et al., 2007). This includes the potential for incidents that might be described as a ‘near-miss.’ The term will commonly refer to incidents reported through incident reporting systems in this thesis.
LDA	Latent Dirichlet Allocation (also used in other statistical contexts to mean ‘linear discriminant analysis,’ that is not used in this thesis)
LOS	Length of stay in hospital
MAE	Mean Absolute Error
NHS	National Health Service
NHSE	NHS England
NHSI	NHS Improvement
NPSA	National Patient Safety Agency
NRLS	National Reporting and Learning System
SIRR	Standardised Incident Reporting Ratio
SHMI	Summary Hospital-level Mortality Indicator

The statistical programming in this thesis was conducted in the statistical programming language R (R Core Team, 2016). R is commonly extended by third party ‘packages’ written by researcher and programmers for specific purposes. Where R code, functions or package names are mentioned, they occur in Courier New font as: `package name`.

Contents

Abstract	4
Impact statement	5
Acknowledgements and authorship statement	7
Glossary of terms abbreviations, and conventions	9
Contents	10
List of figures	15
List of tables.....	17
Copyright statement.....	19
Chapter 1 Introduction to thesis	20
1.0 Medical error	20
1.1 The National Reporting and Learning System (NRLS).....	23
1.2 Thesis aims structure	24
Chapter 2 Literature review	29
2.1 Introduction.....	29
2.2 Methods for review	29
2.2.1 Search strategy	30
2.2.2 Screening process.....	31
2.2.3 Assessment of articles/study quality.....	32
2.2.4 Data extraction	33
2.3 Results	33
2.3.1 General results.....	33
2.3.3 Other topics	60
2.4 Conclusions.....	62
2.4.1 Strengths and limitations.....	64
2.5 Lessons learned for development of models	65
Chapter 3 Data descriptions, handling and summary	67
3.1 Introduction.....	67
3.2 Description of data processing and data set	67

3.2.1 Data receipt, processing and characterisation	67
3.2.2 Requested data extract and format received.....	68
3.3 Summary statistics	72
3.4 Selected relationships between data items.....	83
3.5 Conclusions	83
Chapter 4 Methodological considerations for the analysis of count data.....	85
4.1 The Poisson distribution and Poisson regression	85
4.2 Error structure and overdispersion.....	87
4.2.1 Bootstrapping & likelihood profiles	89
4.2.2 Scaled deviance models	90
4.2.3 Mixture/compound distribution models	90
4.2.4 Multilevel models.....	93
4.3 Scaling/standardizing of covariates	97
4.4 Predictive versus explanatory models	98
4.4.1 Presentation of coefficients	99
4.5 Assessment of model fit and performance.....	99
4.5.1 Model diagnostics and global fit	100
4.5.2 Predictive performance	101
4.6 Summary	103
Chapter 5 Count models of NRLS.....	104
5.1 Theoretical model	104
5.2 Dataset construction.....	106
5.2.1 Hospital Episode Statistics (HES) data.....	106
5.2.2 Aggregate dataset for modelling	108
5.2.3 Organisations included	111
5.3 Parameterisations of incident and exposure data.....	113
5.3.1 Describing exposure data in the aggregated dataset	113
5.3.2 Time period/seasonality	116
5.3.3 Excluded predictors.....	117

5.3.4 Summary of constructed dataset	118
5.4 Single-level model fitting and output	122
5.4.1 Methods	122
5.4.2 Results	123
5.4.3 Discussion	129
5.5 Mixed/random-effects models.....	130
5.5.1 Methods	130
5.5.2 Results	132
5.5.3 Discussion	135
5.6 Model selection	135
5.6.1 Methods	136
5.6.2 Results	136
5.6.3 Discussion	139
5.7 Extending models to longer time periods	140
5.8 Discussion and conclusions	143
Chapter 6 Non-parametric modelling techniques: GAMs, Trees & Neural Networks	147
6.1 Introductions	147
6.2 Generalized additive models (GAMs).....	147
6.2.1 Structure of GAMs	148
6.2.2 Smoothers	149
6.2.3 Estimation of GAMs.....	152
6.2.4 Model selection and degrees of freedom	153
6.2.5 Random-effects models in GAMs	154
6.2.6 Fitting GAM models to incident data	154
6.2.7 GAM model conclusions.....	159
6.3 Algorithmic methods	160
6.3.1 Tree-based methods.....	160
6.3.2 Applying regression trees and extensions to NRLS data	164
6.4 Artificial Neural Networks	167

6.4.1 Artificial neural network structure and estimation	167
6.4.2 Fitting neural networks to NRLS data	172
6.4.3 Conclusions for neural networks	174
6.5 Conclusions and comparisons with GLMM.....	175
Chapter 7 Development of death or severe harm models.....	178
7.1 Introduction	178
7.2 Development of death/severe harm incidents model.....	178
7.3 Conclusions	182
Chapter 8 Developing a risk-adjusted indicator for NHS regulatory use.....	184
8.1 Methods for UK health regulators	184
8.1.1 Regulators	184
8.1.2 Current users of NRLS incident reporting data	186
8.2 Creating a standardised incident reporting ratio (SIRR)	188
8.3 Monitoring techniques used by CQC	188
8.3.1 Comparison using funnel plots	189
8.3.2 Transformation to z-scores	191
8.3.3 Estimation of overdispersion using an additive model.....	193
8.3.4 Time series monitoring using ‘cusum’ charts.	196
8.4 CQC-style techniques applied to NRLS data models.....	198
8.4.1 Marginal vs. conditional.....	198
8.4.2 NRLS results.....	200
8.5 Conclusions	213
Chapter 9 Text mining models.....	216
9.1 Introduction	216
9.2 Text mining techniques.....	216
9.3 Previous work with NRLS text	216
9.4 Preparing text for modelling	218
9.5 Word frequency and document frequency	223
9.6 Topic models	226

9.7 Latent Dirichlet Allocation (LDA)	226
9.7.1 Using LDA to predict incident harm-level.....	227
9.7.2 Model tuning, fitting and results	229
9.8 Conclusions.....	235
Chapter 10 Development of a reporting tool.....	238
10.1 Introduction.....	238
10.2 Designing and building reporting processes	238
10.2.1 Software architecture for designing a report.....	238
10.2.2 Building and coding SQL Server stored procedures.	240
10.2.3 Modelling procedures and creating an R-package.....	241
10.2.4 Final processing	243
10.2.5 Construction of analysis module	243
10.3 Module release review and update.....	246
10.4 Summary.....	246
Chapter 11 Discussion	248
11.1 Incident reporting in the context of patient safety	248
11.2 NRLS data set structure	250
11.3 Statistical models build, and overdispersion.....	252
11.4 Applications in NHS organisations and barriers	257
11.5 Wider applications.....	258
11.6 Future NRLS analyses.....	259
11.7 Recommendations.....	260
11.8 Final comments	261
References	263
Appendices	283
Appendix A: Full Literature review search strategy	283
Appendix B: Screen captures.....	286
Appendix C: Supplementary tables	294
Appendix D: Interactive module development process flow chart.....	322

List of figures

Figure 1.1 ‘Swiss cheese’ model of accident causation	21
Figure 2.1: Search and screening processes for literature review.....	32
Figure 3.1 Example of CSV file splitting.....	68
Figure 3.2 Time of day of NRLS incident reports.....	79
Figure 3.3 Histogram of subject ages in incident reports	81
Figure 4.1 Illustration of single-level and random-intercepts on simulated data	95
Figure 5.1 Relationship between total incidents and main IP, OP & A&E counts, reported per month, per trust.....	118
Figure 5.2 Distributions of predictor variables from NRLS-HES combined dataset.....	121
Figure 5.3 Comparison of estimated model coefficients and 95CIs for NRLS-HES Poisson regression models.....	125
Figure 5.4 Model diagnostic plots for NRLS-HES Poisson regression models.....	125
Figure 5.5 Model diagnostic plots for NRLS-HES scaled Poisson models: Quasipoisson, Negative Binomial and Gen. Poisson	128
Figure 5.6 Model diagnostic plots for NRLS-HES random-intercept models	134
Figure 5.7 : Example of variation in slopes of random-intercepts for six outlier organisations	142
Figure 6.1 Approximation examples for non-linear relationships	148
Figure 6.2 Examples of controlling spline smoothers	151
Figure 6.3 Marginal smooth plots for NRLS-HES Poisson GAMs.....	156
Figure 6.4 Marginal smooth plots for NRLS-HES Negative Binomial (NB2) GAMs.....	157
Figure 6.5 Example structure for NRLS regression tree models	161
Figure 6.6 Schematic of a single hidden layer, feed-forward artificial neural network.....	169
Figure 6.7 Training and validation error for NRLS-HES Neural Network	174
Figure 7.1 Histogram of counts of death or severe incident reports per trust, per month	178
Figure 7.2 Relationships between death or severe, and total, NRLS incident reports	182
Figure 8.1 Example funnel plot using the ‘medpar’ dataset.....	190
Figure 8.2 Theoretical normal distribution and z-scores	192
Figure 8.3 Example of Winsorisation of a distribution of z-scores	194
Figure 8.4 Distribution of z-score transformation and adjustment methods.....	202
Figure 8.5 Comparison of overdispersion adjusted funnel plot methods	203
Figure 8.6 Standardised Incident Reporting Ratio funnel plots for total incident report models	205

Figure 8.7 Standardised Incident Reporting Ratio funnel plots for total incident report models (2).....	206
Figure 8.8 Standardised Incident Reporting Ratio funnel plots for death or severe harm NRLS incident report models	207
Figure 8.9 Illustrative CUSUM charts for Standardised Incident Reporting Ratios (SIRRs)	211
Figure 9.1 Wordclouds of top 100 words at different stages of cleaning	221
Figure 9.2 Wordclouds of top 100 words, after cleaning/ stemming, by harm level.....	222
Figure 9.3 Wordclouds of top 100 bigrams, from ski-gram models, by harm level	224
Figure 9.4 LDA metrics for word token models.....	230
Figure 9.5 LDA metrics for skip-gram(bigram) token models.....	233

List of tables

Table 2.1 Primary subjects assigned to NRLS-related articles, identified by systematic review.	
.....	36
Table 2.2 Most common subject tags applied to NRLS-related articles identified by systematic review.....	37
Table 2.3 Sources of potential bias identified in NRLS-related articles identified by systematic review.....	38
Table 2.4 Evidence Summary tables from NRLS literature review	51
Table 3.1 Data field names and metadata in NRLS extracts	71
Table 3.2 NRLS incidents reports per year reports received by NRLS	72
Table 3.3 NRLS incidents reports per year of incident occurrence.....	73
Table 3.4 Incident reports, per year and harm classification	74
Table 3.5 Care settings of incident reports.....	75
Table 3.6 Incident reports by location levels 1 and 2	77
Table 3.7 Incident reports by level 1 specialty groups.....	78
Table 3.8 Time of day of NRLS incident reports.....	79
Table 3.9 Incident reports by level 1 incident types	80
Table 3.10 Incident reports by subject age groups.....	80
Table 3.11 Incident reports by Patient Sex groups	81
Table 3.12 Proportions of Acute hospital incidents per year	82
Table 4.1 Summary of model variance functions for Poisson-based, variance scaled, and mixture models.	93
Table 5.1 Summary statistics for combined NRLS-HES modelling dataset	119
Table 5.2 Model coefficients for NRLS-HES Poisson GLM with Wald and bootstrapped confidence intervals.....	124
Table 5.3 Model coefficients for NRLS-HES Quasipoisson, Negative Binomial and Generalized Poisson GLMs.....	127
Table 5.4 Model coefficients for NRLS-HES random-intercept Poisson and mixture GLMMs.	133
Table 5.5 Prediction error and AIC summary NRLS-HES models	137
Table 5.6 Comparison of predictors significance for NRLS-HES models.....	138
Table 6.1 GAM model outputs for NRLS-HES models.....	158
Table 6.2 Mean Absolute Error (MAE) for regression tree-based NRLS-HES models.....	167
Table 6.3 Comparison of GLM, GLMM, GAM, Tree and Neural Network prediction errors....	176
Table 7.1 Mean absolute prediction error for death or severe harm NRLS-HES model.....	181
Table 8.1 Funnel plot outlier organisations for 'All incidents' and death or sever harm (DS) incident models.....	208

Table 9.1 Example of ‘tidy’ tokenization and database representation.....219

Table 9.2: Results of multi-class prediction models, using word tokens, for level of harm.....231

Table 9.3 Distribution of incident reports in harm classes.....232

Table 9.4 Results of multi-class prediction models, using skip-gram tokens, for level of harm
.....233

Table 9.5 Results of multi-class prediction models, including NRLS categorical data, for level of
harm235

Table 10.1 Summary of stages to build HED NRLS reporting module240

Copyright statement

The data used in this thesis are sourced from the National Reporting and Learning System (NRLS), and the Hospitals Episode Statistics (HES). The following notice applies to all uses of HES data within the project, this document, and in any discussions of outputs:

Copyright © 2019, the Health and Social Care Information Centre (NHS Digital). Re-used with the permission of the Health and Social Care Information Centre (NHS Digital). All rights reserved.

Chapter 1 Introduction to thesis

1.0 Medical error

The capacity for error is ever-present in healthcare, and a major cause of morbidity and mortality globally (Jha et al., 2013). This risk has been described as being '*...on an entirely different scale from error tolerated elsewhere*' and having '*different consequences from error in other service sectors*' (Elwyn and Corrigan, 2005). It is reasonable for patients to expect their care to be safe and free from errors, but this is an impossible task given the nature of error. Patient safety as a research area and discipline has seen significant growth over the last two decades, with the broad aim of seeking '*type(s) of process or structure whose application reduces the probability of adverse events resulting from exposure to the health care system across a range of diseases and procedures.*' (Shojania et al., 2002). Learning from error is a major component of patient safety, and this thesis is concerned with analysis methods for a specific learning system, described in this chapter.

Study of errors in fields other than healthcare has suggested complex interactions between most failures and elements including social, behavioural, cultural and technological factors (Donaldson, 2002). There are various conceptual bases for patient safety terms, ontologies and frameworks for understanding causation (Carson-Stevens, 2017, Runciman et al., 2010), with a general consensus that errors are not simply the fault of practitioners, but also the systems and environments in which they occur.

The UK's Chief Medical Officer, Prof. Sir Liam Donaldson, used work by Reason (1990) to describe the risk of accident and adverse events as if they were holes in a series of slices of Swiss cheese (Figure 1.1) (Donaldson, 2002). In this image, the solid pieces of cheese represent system defences and processes, with the holes representing vulnerabilities in these systems. Dangers arise in instances where the holes 'line up', allowing an incident to occur and slip through successive defensive layers, representing a chain of vulnerabilities in these systems.

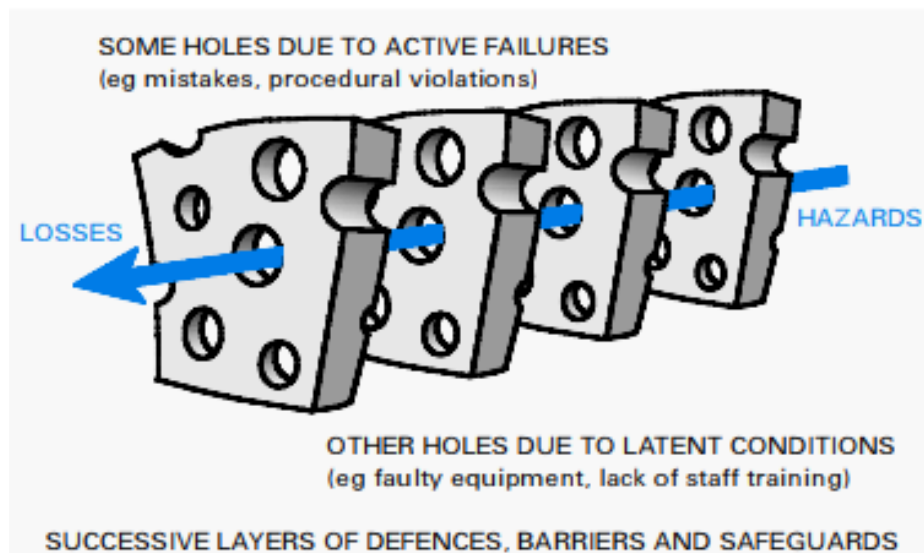


Figure 1.1 'Swiss cheese' model of accident causation

Figure taken from Donaldson (2002), based on the work of Reason (1990), illustrating the defensive systems as the solid parts of each slice, with holes representing vulnerabilities. Adverse events are often the result of an alignment of several system weaknesses, represented by the blue arrow.

Following the publication of the US Institute of Medicine's report '*To err is human*' (Kohn et al., 2000), Prof. Donaldson commissioned a review of the English system, estimating that between "6,000 and 25,500 NHS patients each year suffer serious disability or death as a result of healthcare interventions" (Hogan et al., 2012, Donaldson, 2000). These estimates were based on retrospective case-note review techniques, developed as part of the Harvard Medical Practice study (Brennan et al., 1991), where clinicians reviewed patient records. Although such estimates have been questioned (McDonald et al., 2000), these reports focussed attention on medical error, and acknowledged that although impossible to eradicate, errors may be used for learning and building defensive safety systems to prevent recurrence.

Case-note review is a well-established method of identifying problems with care, but it is resource-intensive, requiring training of reviewers, and consistent reviewing criteria (Hogan et al., 2012, Brennan et al., 1991, Williams et al., 2015, Wilson et al., 1995, Vincent et al., 2001). These methods can be well-designed and tested for agreement between reviewers, but may be hindered by inherent subjectivity or reliability issues, and may be biased by medical training and prevailing cultural attitudes (Hogan et al., 2012). Electronic incident reporting has been proposed as an additional approach for learning from error (Sheikh and Hurwitz, 1999) (Pronovost et al., 2006), based on the experience of reporting systems in other industries.

Incident reporting cultures are commonly found in high risk industries (Barach and Small, 2000, Francis, 2013) such as aviation (NASA, 2019, The CHIRP Charitable Trust, 2020), nuclear power (2015, IAEA, 2010) and oil production(Christou and Konstantinidou, 2012). The purpose of these systems is to promote learning by allowing the identification of risks that require further examination, and to highlight broad areas for targeting improvement activities (Macrae, 2016).

The example of other industries has pointed toward incident reporting providing useful signals for further investigation, but often the incident report itself is not considered to contain much information (Macrae, 2016). The message that signals should trigger in-depth investigation (Macrae and Vincent, 2014) has been heeded in recent years in the NHS, with the establishment of the Health and Safety Investigation Branch (HSIB) (Department of Health & Social Care, 2017), but incident reporting in healthcare has been established in several countries for many years, including the USA (Nuckols et al., 2007, Santell et al., 2003) and Australia (Runciman and Moller, 2001). Such systems are thought to improve reliability and safety by closing the ‘information loop,’ where experience of errors provides information for root cause analysis, with expert review building new evidence and driving system improvements (Carter et al., 2015). This NHS took its cue from earlier systems, including those in the US, where the collation of established regional reporting systems was recommended, with both voluntary and compulsory elements discussed in *‘To err is human’* (Kohn et al., 2000).

To further examine incident reporting in healthcare, we must define what incidents are reported. Another common term used somewhat interchangeably with “incidents” is “adverse events.” In clinical trials settings, adverse events commonly relate to any form of harm to patients, whether it is caused by the treatment under trial or not (Wittes et al., 2015). This definition does not fit well with healthcare delivery settings, as we do not always know the outcome. In this thesis, the term ‘adverse events’ will refer to: *“Any unintended event caused by the health care that either did or could have led to patient harm”* (Sari et al., 2007). The agency tasked with the creation of the UK’s reporting database (National Patient Safety Agency, 2004) defined ‘incidents’ as: *“any unintended or unexpected incident that could have or did lead to harm for one or more patients receiving NHS-funded healthcare,”* but this definition has grown to include events related to staff and organisational factors as well. It has also grown to include potential events, referred to as ‘near misses.’

The UK has a variety of data collection systems for incidents in healthcare. This confusing landscape demonstrates a degree of overlap and potential duplication or unnecessary collection (Mayer et al., 2017):

- **Strategic Executive Information System (StEIS):** used by managers to notify regulators about serious incidents, especially but not exclusively, the ‘never events’ list (NHS England, 2015).
- **Medicines and Healthcare Regulatory Agency (MHRA) yellow card scheme:** reporting system for regulator of new medications and devices (Medicines and Healthcare Regulatory Agency (MHRA), 2016).
- **Care Quality Commission (CQC):** organisations must notify CQC of adverse events, often through NRLS, but also directly in some cases (Care Quality Commission (CQC), 2016).
- **NHS Safety Thermometer:** a point of care survey that includes items related to harms (NHS Digital, 2012).
- **Serious Adverse Blood Reactions & Events (SABRE):** MHRA mechanism for reporting blood related event (Medicines and Healthcare Regulatory Agency (MHRA), 2010).
- **Serious Hazards of Transfusion Scheme (SHOT):** blood related voluntary incident reporting scheme that can also be completed via SABRE (Serious Hazards of Transfusion, 2016).

1.1 The National Reporting and Learning System (NRLS)

The National Reporting and Learning System (NRLS) is the English and Welsh NHS’ repository of incident reports from healthcare, that can be used for analysis and learning (Carter et al., 2015). Data are available to NHS organisations and researchers, at the discretion of NHS Improvement (NHSI), and quantitative analyses of these data is the subject of this thesis.

The size of this database is both a strength and an impediment to analysis. In evidence to the Mid-Staffordshire Hospitals enquiry, Prof. Donaldson who was responsible for implementing the UK response to the patient safety agenda, described the situation thus: *“The number of reports received is ... huge, so that raises the question of how can we analyse them all properly. Decisions therefore need to be made as to whether we need tighter rules on incident reporting, and the distinction between local and national level reporting and follow-through”* (Francis, 2013).

The NRLS was originally implemented and ‘owned’ by the National Patient Safety Agency (NPSA), who were merged, first into NHS England (NHSE), and then NHSI. Much has been learnt from it during its existence, including identifying risks in airway management between critical care and other settings (McGrath and Thomas, 2011), that drug-related errors are commonly related to wrong administration (Cousins et al., 2012) (with vinca-alkaloids a prominent example (Franklin et al., 2014)), identifying the risks of shock and death associated with the use of bone cement in hemiarthroplasty for fractured neck of femur surgery (Rutter et al., 2014) etc. Many of these messages are discussed in greater depth in the literature review in the next chapter, but have also formed the bases for ‘Rapid Reports’ or ‘Patient Safety Alerts’ from NPSA/NRLS/NHSI (Panesar et al., 2009). Despite many patient safety alerts, academics studies, and major investments in patient safety systems, there have been few evaluations of their effectiveness (Carson-Stevens et al., 2018).

NHSI’s patient safety team spend considerable effort reviewing severe harm and death incidents, where free-text descriptions of incidents are read by clinically trained staff. This is both expensive and time-consuming, with nearly 80,000 such reports for fiscal years 2010/11 – 2016/17 (calculated from NRLS data supplied for this project). These reports could be considered ‘the tip of the iceberg,’ representing just 0.72% of reports during this period. This leaves the majority of incident reports, that may provide valuable information, unused at national level. It also leads to questions about the purpose of collecting such a vast resource that cannot practically be analysed. Statistical and machine learning methods have seen some limited implementations, particularly around text mining (Bentham and Hand, 2012, Bentham, 2010, Bentham and Hand, 2009, Altuncu et al., 2018), and investigating associations with harm level (Cuong Pham and Colantuoni, 2010, Howell et al., 2015, Wahr et al., 2014, Hutchinson et al., 2009), but regular analytical methods, other than simple reporting rates, are not in common use.

1.2 Thesis aims structure

This applied statistics thesis attempts to address this gap by investigating previous analytical approaches, examining the NRLS, borrowing strength from other datasets for casemix-adjustment, and examining how outputs can fit into current regulatory structures. It also addresses the practical application of these techniques to build a tool to interrogate data, as well as accessible text analysis approaches that might be applied by regulators or investigators. The aim is to aid regulators, NHS organisations and researchers to identify organisations of

interest that might require investigation or may be examples of best practice from whom the whole system could learn. The intended outputs are the statistical methods for application by others, preliminary model data that can be investigated with the regulators, and the construction of an interactive online tool delivered through the Healthcare Evaluation Data (HED) benchmarking tool.

The objective of patient safety research, and this thesis is to prevent future harm by advancing methods to learn from error in the NHS. The current methods of review used by NHS Improvement, CQC and others are effective for learning but are limited by the resources required for case-note review. NHS Improvement's patient safety team cannot feasibly review the huge number of reports they receive each year. This is a barrier to learning and improvement, but also calls the purpose of national data collection into question. Valuable information is contained in some of these reports, reviewers may be looking for a 'needle in a haystack.' Therefore, the main question addressed by this thesis is: **'can we make better use of the NRLS data we are already collecting?'**

To answer this question, the aims of the project are:

1. To identify what types of learning have been derived from NRLS data so far, their strengths and their limitations.
2. To investigate and identify relevant data structures and necessary preparation steps for quantitative analysis of NRLS data.
3. To examine what statistical modelling processes and techniques are appropriate for NRLS data, develop and apply relevant methods.
4. To examine presentation methods for outputs that allow regulators and NHS organisations to interpret and use model output.

This thesis describes the current situation where the majority of incident reports (that do not lead to severe harm or death) cannot be reviewed at national level due to the scale of the task. My thesis addresses this unmet need by developing statistical, machine learning and text mining methods to identify differences in incident reporting culture between organisations. It provides monitoring methods and tools with better adjustment for differences between hospitals, and it explores whether quantitative analysis of free-text may be routinely applied to these data. The next chapter begins this by examining how NRLS data has been analysed in published literature at the start of this project.

Two analytical strategies were pursued in this thesis: analysis of incident reports as count data, and analysis methods for free text. These two methods are independent, but are reported in order, addressing the count methods first, and the free-text methods second. The thesis is structured in four main components, aiming to identify current analysis techniques, their limitations and propose new methods for greater insight. These sections are broadly: literature review, development of statistical models for count-based analyses, methods for presenting such count-based models in current regulatory frameworks, and methods of free-text based analysis. The chapter structure and content are summarised below.

Chapter 2 addresses aim 1, through identifying and examining a literature base around NRLS analysis. The difficulties in identifying this literature through common search practices are described, using an iterative screening process to reduce the volume of data returned. This chapter highlights the clinical areas where NRLS data have provided information, the problems with the incident reporting data and the NRLS specifically. It examines reporting practices and reliability debates, and describes the limited quantitative statistical analyses performed to date.

Chapter 3 addresses aim 2 by describing the NRLS data and the methods required to process it at scale. It details the methods for receiving, cleaning and processing the raw data. Summary data are presented, and data fields are described. Limited statistical testing is also performed, demonstrating some association between NRLS categorical data, but with few helpful conclusions using NRLS alone. A method of aggregating data to form a count data set, with predictors from other sources, is proposed.

Chapter 4 addresses aim 3 by examining how count data are commonly modelled. Chapters 2 and 3 suggest that NRLS data can only be feasibly modelled as count data, and this chapter discusses Poisson regression as a basis for the development of further models. Overdispersion and various methods for dealing with it are discussed, including mixture models, quasi-likelihood and multilevel modelling approaches. Issues around model checking and comparison are examined and used to inform the modelling strategy in the following chapters.

Chapter 5 addresses aims 2 and 3 by introducing a conceptual model for analysing incident reports focussing on 'exposure' and 'culture' factors. Exposure is then examined by

constructing the count data set suggested in chapter 3, using casemix factors identified from the Hospital Episode Statistics (HES) (NHS Digital, 2017d). The methods from Chapter 4 are used to develop a generalised linear modelling approach, with overdispersion identified as a major issue. Alternative models are examined to deal with overdispersion, with multilevel models including random-intercepts for clusters of repeated measure proving the most successful.

Chapter 6 continues to develop aim 3 by applying methods better suited to 'noisy', non-linear, or non-parametric data, relevant to NRLS. Smoothing approaches can aid modelling when working with 'noisy' datasets, and Generalised Additive Models are used to this end. Use of appropriate smoothers and estimation of smoothing parameters are discussed, and random-intercept structures developed in Chapter 5 are also applied. Regression trees, 'boosting', 'bagging' and 'Random Forests' are introduced as alternative algorithmic approaches from machine learning methods. Random Forests showed the most promise of these techniques and were taken forward to subsequent chapters. Artificial Neural Networks were also investigated to examine if their non-linear methods would better predict data but did not outperform the regression-based models already examined.

Chapter 7 takes the model architectures in previous chapters and refits them to the subset of severe harm or death incidents, the group that are manually audited by NHSI at present. These further pursues aim 3, by targeting the subgroups that represent the most urgent signal and are at the highest risk. These events are rare, and models were affected by this sparsity, with simpler models proposed and evaluated.

Chapter 8 addresses aim 4 by introducing the common statistical tools used by NHS regulators, including CQC, who have the most detailed methodology for examining organisational variation. Conditional and marginal model predictions are discussed, and marginal predictions are used to apply current regulatory methods. A Standardised Incident Reporting Ratio (SIRR) is proposed as a ratio of observed to expected counts from the casemix models developed in previous chapters. An additive overdispersion adjustment, based on meta-analysis techniques is described, and applied to calculate adjusted z-scores and funnel plots. Methods for monitoring are investigated through the use of Cusum control charting techniques.

Chapter 9 does not follow sequentially from previous analyses, but instead reports a parallel analysis approach targeting free text. This chapter answers aim 3 by examining the other major source of information in each report, the incident description, that has been ignored in earlier chapters. The challenges and standard practices for text analyses are explained and applied to NRLS data. Information on term frequency and how to use these metrics are discussed. Text are then used to build topic models based on latent topics identified by Latent Dirichlet Allocation (LDA). LDA models are then used to predict levels of harm and compared to the recorded values, showing promise for analytical techniques and with potential to aid data quality/completeness in future systems.

Chapter 10 takes the methods developed in chapters 5-8 and shows how they have been converted to an interactive online tool, for use by organisations, within the Healthcare Evaluation Data (HED) benchmarking tool. This chapters described the practical processes to achieve this using SQL, R and web technologies. This chapter contributes to aim 4, and the wider dissemination and practical use of the methods developed in previous chapters.

Chapter 11 is a final discussion chapter, drawing the conclusions of the various chapters together. This chapter addresses all four aims of the thesis and considers them in the wider context of the NHS, statistical methodology and implications for practise. The chapter included suggested wider applications and recommendations for stakeholders.

A reference list and various appendices are included, with relevant screenshots, supplementary tables and figures, at the end of this thesis.

Chapter 2 Literature review

2.1 Introduction

To determine how best to examine the NRLS data, a survey of previous analyses, research articles and the types of information they could access was required. Initial investigations of the NPSA website revealed only a limited number of resources, related mostly to their regular reporting functions, although some special reports and analyses by subject experts was published. NRLS analysis and use in academia was less clear, as there were no specific links to these publications. A review of academic and published literature was required to provide context and background information to answer aim 1 of this project, by identifying the what has been learned from NRLS to date.

This chapter describes the methods, difficulties and information gleaned from this review. It was conducted at the beginning of the project and has not been subsequently refreshed, but later publications are discussed where relevant in later sections of this thesis.

2.2 Methods for review

The aims for this review were to identify the areas where NRLS data has been analysed, what insights have been gained from these analyses and what methods have been used to gain such insights. Given these aims, it was necessary to perform a wide-ranging appraisal of the published literature. Various review methods may be used for literature reviews, including:

- **Narrative Review** – A summary of the state of knowledge in a field, generally collated by a topic expert that (Gregory and Denniss, 2018). Reviews of this type are typically not systematic in their approach, relying on researcher's experience, interpretation and knowledge of their field. They may contain biases and interpretations but can be used to take a wide view of a field and are particularly useful when a study question cannot be addressed by systematic review methods.
- **Systematic review** – Rigorous protocols to identify and collate all available research, fitting pre-specified eligibility criteria, to answer a specific research question (Higgins et al., 2019). Such reviews are necessarily narrow in focus, usually targeted at specific study types according to a hierarchy of evidence. They commonly involve careful assessment of quality and rigour and may involve meta-analysis of treatment effects where relevant.

- **Scoping review** – A type of review that may be used to identify and map a literature base, identify evidence in a particular field, or highlight particular research methods (Munn et al., 2018). Their methods are not as strictly defined as systemic reviews, but methods may overlap with systematic or narrative reviews standard but may involve similar processes. Scoping review may be precursors to systematic reviews where a narrowly focussed question has not yet been developed.

The review methods used for this chapter resemble a scoping review in many ways, with methods for systematic searches of published literature adopted from systematic reviews. This was necessary due to the diversity of the literature identified, and a desire to include as many relevant articles as possible. This has, in some cases, included letters and opinion articles where evidence has been drawn from the NRLS in some fashion. These articles would not have featured in a narrower systematic review, but this diversity presented difficulties in assessing the quality of articles. It also prevented the use of systematic checklists, commonly used in critical appraisal of evidence, or formal tools for assessment of quality. Studies were, instead, assessed for their method susceptibility to bias (section 2.2.3).

Systematic searches of major medical research databases were performed, targeting published research articles, conference presentations, theses and reports where NRLS data has been used directly. Patient safety alerts or reports issued by NPSA / NHSE or NHS Improvement have not been included in searches. Editorials or summaries of these alerts appearing in professional journals as practice development articles, rather than primary research, were also excluded.

2.2.1 Search strategy

Electronic database searches were used to identify articles published from 2001 (the year the NPSA was established) onwards. Text-based searches for variants of “NRLS” or “NPSA” in conjunction with variants of “incident” and “reporting,” as well as variants of “incident” and “report” with “NHS” or “database” from Great Britain were used as main search terms. Medical Subject Headings (MESH) were considered and searched, but MESH and other index terms were not used in final searches. Index terms were inconsistent across articles, with few conserved terms, in pilot searches. Articles were also inconsistently indexed under a variety of different terms across search engines, as many appeared in more than one source. See Appendix A for full search criteria.

Database sources used included: the Cochrane Library, DARE, CRD, Medline, Health Management Information Consortium (HMIC), Embase, CINAL, Proquest Dissertations and Theses, The British Library Theses, ProQuest Risk management and Web of Science.

2.2.2 Screening process

References were imported into Endnote X7 and limited de-duplication performed using Endnotes tools before manual screening. Screening was conducted in rounds using title, title & abstract and full-text (Figure 2.1). Inclusion and exclusion criteria were interpreted conservatively with uncertain articles included for the following round.

The first screening round was based on article title and used exclusion criteria:

- Articles related to subjects other than incident reporting (e.g “prostate specific antigen” (PSA))
- Articles focussed on countries or health systems other than England or Wales
- Articles with no obvious link to NRLS

Articles described as ‘audits’ in their titles were not informative enough to be excluded at this stage.

Second round screening focussed on title and abstract, using inclusion criteria:

- Articles directly mention or use NRLS data

Where abstracts were not indexed in search engines, full-text article were retrieved and where abstracts were not available due to article style, executive summaries/overviews or the first page of body text were screened in the same manner.

Full-text screening was based on inclusion criteria:

- NRLS data must be analysed in some form, including (but not limited to):
 - Summary statistics (e.g. frequencies)
 - Qualitative analysis (e.g. text entries reviewed, themes identified and data extracted)
 - Statistical analysis (e.g. regression modelling)

Summaries of NPSA safety reports/alerts were excluded from the review, as these formed part of the NRLS’ own work programme, rather than primary research. One protocol article was included, as it directly referenced incident counts from its primary data.

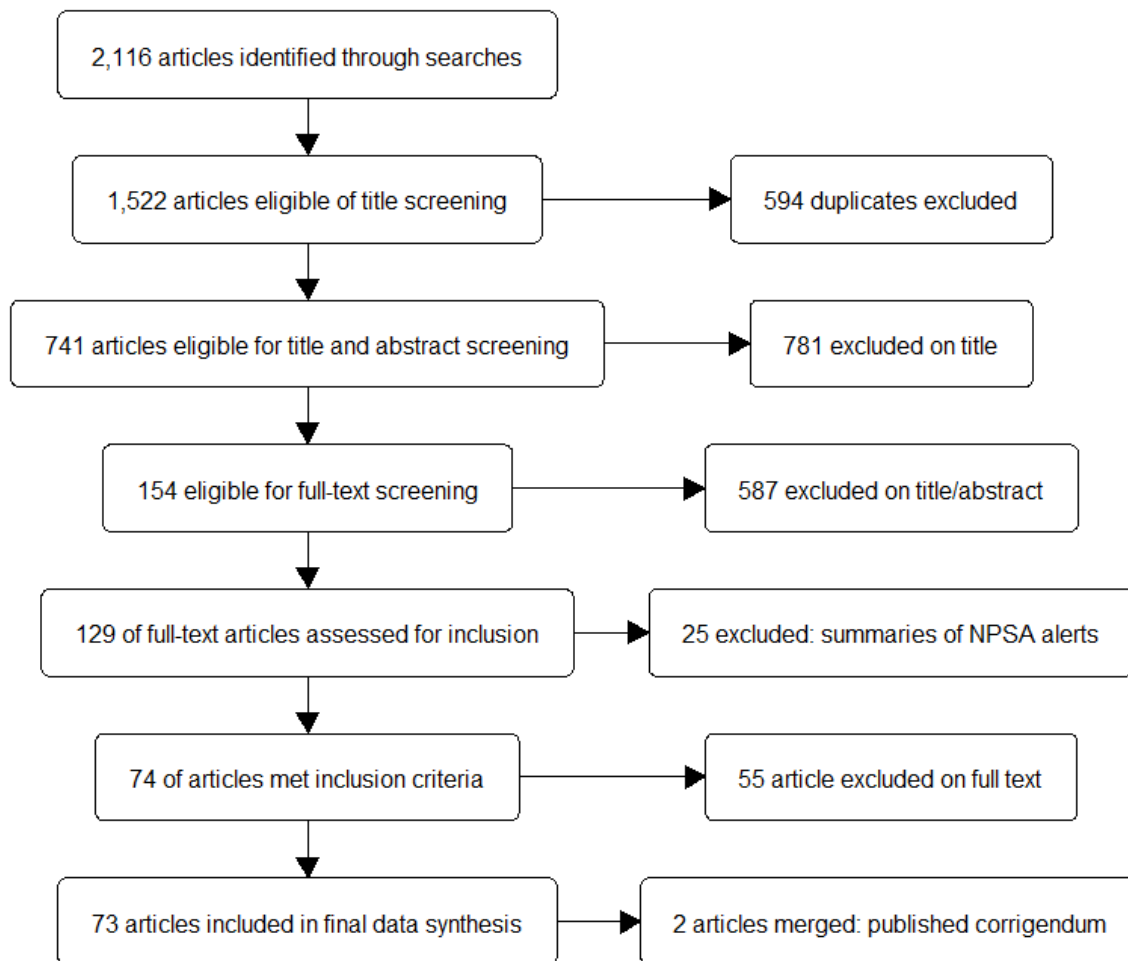


Figure 2.1: Search and screening processes for literature review.

Search and screening processes for literature review. Numbers of articles retrieved and excluded, are shown with arrows showing the process flow.

2.2.3 Assessment of articles/study quality

Standardised assessment tools commonly used in systematic reviews of clinical trials (e.g. the CONSORT statement (Schulz et al., 2010)) or specific study types (such as the CASP checklists (Critical Appraisal Skills Programme, 2016)) were not applicable to this review due to the range of articles included. The structures of these articles were not consistent, and not amenable to assessing with a structure-based checklist. Many articles were not peer-reviewed or not constructed in a manner resembling a traditional cross-sectional study or standard trial design. Quality was therefore assessed in terms of the strengths and weaknesses of each article, focussing on methodological rigour, NPLS search protocol/method of identifying relevant incident reports, and susceptibility to (or acknowledgement of) bias (table 2.3). Common biases were coded during data extraction and reviewed. Where biases undermined studies, they are explained in the results section.

Factual errors were identified in a number of papers, e.g. mis-matching percentage calculations (Cousins et al., 2012), unexplained gaps in before and after studies (Flood et al., 2014, Flood et al., 2015), biased comparisons (Flood et al., 2015) and lack of denominators of total incidents reviewed (Flood et al., 2015, Hutchinson et al., 2009, Fisher et al., 2015). Individual errors did not necessarily undermine the arguments of a given study, but in cases where authors or studies showed multiple errors, the reliability of those studies is questionable.

2.2.4 Data extraction

A custom data extraction and recording method was developed for this review due to the differing nature of articles. The full-text articles were read and extracted information (defined as: any data, data-based hypotheses or conclusions expressed uniquely in that article) were recorded. A primary subject was selected and list of secondary subjects (“tags”) was created that allowed multiple tags to be assigned to articles.

A relational database was designed, using Microsoft SQL Server 2014, for data capture and analysis (structure detailed in appendix B.1). Data input forms, connected to the database, were created in Microsoft Access 2013. Primary subjects were used to structure results in table 2.1.

The database tool was designed prior to extraction, but retained flexibility to be altered. Referential integrity was enforced on all ‘lookup’ fields so that if a tag was altered or removed, all records were subsequently updated. Information recorded as data or learning derived from studies were extracted into their own fields. Strengths and limitations were grouped separately and additional notes, such as structural issues, were recorded in their own fields. Appendix B.2 is a screenshot of the initial data collection tool.

A validation process was conducted at the end of screening and analysis ensuring all studies had an inclusion status, reconciled with Endnote, and compulsory fields were populated.

2.3 Results

2.3.1 General results

Searches identified 74 relevant articles after screening, with two articles subsequently merged, as they were revisions of the same manuscript in different publications. See table 2.4 for full evidence summary table.

Studies ranged in size from 13 reports (Rocos and Donaldson, 2012) to 5.9 million reports (Howell et al., 2015). The differing scales of studies necessitated different designs, with small studies allowing detailed analyses of individual reports and larger studies restricted to categorical and numerical analyses.

The majority of articles used mixed methods or qualitative methods, due to problems with NRLS categorisations (Table 2.1) (Barai et al., 2014). The depth of methodological detail varied widely, even amongst literature published in the same journal. Some articles included detailed NRLS search terms (Milligan, 2012) explaining how their sample of reports was identified, some included potential biases and reasons for missing data whilst others simply reported numbers of incidents (Panesar et al., 2012b) with no further comments.

Many studies were conducted in a 'clinical audit' style, with clinically trained reviewers reading the free-text descriptions of incidents and making clinical experience-based judgements. A small number of studies applied specific qualitative techniques such as the 'recursive model of incident analysis' (Rees et al., 2015b, Williams et al., 2015) or the 'constant comparative method' (Panesar et al., 2012a, Rees et al., 2015a), with one article developing a framework for analysis of incidents, referred to as the "Primary Care Patient Safety (PISA) Classification System" (Carson-Stevens et al., 2015). Well-developed studies explained the training process for reviewers (Carson-Stevens et al., 2015), how conflicts in coding were resolved and validation checks (Rees et al., 2015b, Carson-Stevens et al., 2015, Williams et al., 2015, Panesar et al., 2014).

Three studies used 'before and after' designs (Flood et al., 2014, Flood et al., 2015, Laker, 2009) to evaluate the effects of 'patient safety alerts' that were influenced by NRLS data. Two articles (Flood et al., 2014, Flood et al., 2015) used long 'before' periods and short 'after' periods (27 months vs. 16 months & 72 months vs. 15 months) and could be viewed as under-powered for detecting the desired change. The 'after' periods were too short to observe the relevant incidents at rates similar to the 'before' period, and therefore lacked the precision to assess their hypotheses. Both studies contained unexplained gaps between before and after periods in which incidents may have been reported.

Five articles applied at least one statistical analysis/modelling technique to describe NRLS. No consistent method was observed, but techniques included:

- Bayesian hierarchical model to examine variance in harm rate between and within hospitals (Pham et al., 2010).
- Rao-Scott chi-squared tests to compare proportions of medication errors between and within US and UK hospital intensive care units (ICUs) (Wahr et al., 2014).

- Generalised estimating equations (GEE) logistic regression to compare reports related to drug classes in between US and UK ICUs, accounting for hospital-level effects (Wahr et al., 2014).
- Log-linear models for contingency tables and linear mixed models to examine effects of location, incident and outcome on reviewer's severity score using (Templeton et al., 2011).
- Development of an error index as a function of error severity and propensity (Panesar et al., 2013b).
- Correlation coefficients and regression models to examine relationships between harm levels and other hospital outcomes measures (Howell et al., 2015).

The majority of articles focussed on secondary care settings, with small numbers in mental health, learning disabilities, or primary care. Low numbers of reports from some secondary care organisations, and other settings, combined with high harm rates (Williams et al., 2015) suggested either limited awareness of incidents or limited reporting in these contexts. A lack of 'no harm' or 'low harm' incidents suggests under-reporting, deflates denominators for total incidents, and prevents robust comparison of harm levels in these organisations.

Levels of harm reported across articles suggested most incidents were classified as 'no harm', 'low harm' or 'moderate harm', with few leading to 'severe harm' or death. E.g. Howell et al. (2015) reported 70.3% as no harm and 0.9% as severe harm or death.

A range of primary subjects, assigned during data extraction, was identified and summarised in Table 2.1. Many assigned subjects aligned with specific NRLS database categories or work-streams from specific research groups (e.g. anaesthesia, critical care or primary care). Subjects assigned to studies overlapped with other categories. In these cases, a primary subject was chosen from the apparent target of the authors, e.g. anaesthetics and surgery in (Arnot-Smith and Smith, 2010) was assigned as an anaesthetics primary subject.

Subject	Conference Abstract	Editorial	Mixed	Qualitative	Quantitative	Total
Acupuncture				1		1
Airway	3		3			6
Ambulance			1			1
Anaesthesia	2		4		1	7
Anaphylaxis	1					1
Cardiac arrest			1			1
Chest Drain			1			1
Critical Care			3			3
Dentistry			1			1
Dermatology			1			1
Diagnostics			1			1
Dialysis	1					1
Falls			3			3
Incident Reporting			3		3	6
Medication	1		10			11
Mental Health					1	1
Nutrition/Hydration			1	1		2
Obesity					1	1
Obstetrics	1					1
Ophthalmology			3			3
Orthopaedics			3		4	7
Paediatrics	2	1	1			4
Resuscitation	1					1
Self-Harm			1			1
Staffing			1			1
Suicide			1			1
Surgery	1		1		1	3
Transfers			1			1
Vaccination			1			1
Total	13	1	46	2	11	73

Table 2.1 Primary subjects assigned to NRLS-related articles, identified by systematic review.

Most articles touched on multiple subjects or overlapped with other classifications, e.g. ‘obesity’ may overlap with ‘secondary care,’ ‘ward settings,’ ‘airway management,’ ‘medication/dosage’ etc. Articles were also ‘tagged’ against the apparent primary subject and all other apparent subjects using a secondary classification system that allowed multiple tags.

Table 2.2 summarises the top 25 categories when multiple subject coding was considered. Medication was by far the most common subject, with medication themes in 70% of articles. A majority of article were secondary care based (55%) with surgery (36%), Administration (of drugs of treatment) (33%) and equipment (32%) the next most common categories.

Subject	Studies tagged	Percentage of Studies tagged (n=73)
Medication	51	70%
Secondary care	40	55%
Surgery	26	36%
Administration (of drugs or treatment)	24	33%
Equipment	23	32%
Communication	20	27%
Documentation	20	27%
Skills/Training	20	27%
Procedure	19	26%
Staff mistakes	18	25%
Incident Reporting	18	25%
Monitoring	18	25%
Dosage	16	22%
Primary Care	16	22%
Transfers	16	22%
Precribing	15	21%
Airway	15	21%
Anaesthesia	15	21%
Mental Health	14	19%
Infrastructure	14	19%
Delays in treatment	13	18%
Critical Care	13	18%
Falls	12	16%
Staffing	12	16%
Patient accident	10	14%

Table 2.2 Most common subject tags applied to NRLS-related articles identified by systematic review.

Table 2.3 summarise the common sources of bias/errors identified in the articles, due to the nature of NRLS data collection and the research methods used:

- **Allocation of specialty** – The recording of treatment specialty is important in classifying the type of incident, but multiple professionals/specialties may be involved in a report. Barai et al.(2014) suggested 69% of surgical incidents had at least one possible alternative classification and that current classifications were too rigid. Baird et al.(2009) estimated 40% of anaesthetic incidents were miscoded as surgery, whilst Cassidy et al. (2011) suggested the misclassification to be as high as 60% in the sample they reviewed.

Description	Articles affected (n) (Total n=73)	Article affected (%)
Allocation of specialty	32	43%
Alternative reporting route	16	22%
Anonymisation	18	24%
Ascertainment of reports	61	82%
Classification of harm	60	81%
Duplication	10	14%
Lack of detail	58	78%
Missing data	54	73%
Poor/lack of search terms	17	23%
Potential vs. actual harm	8	11%
Re-classification by authors	31	42%
Search terms not specified	18	24%
Under reporting	71	96%
Unlikely dates	4	5%

Table 2.3 Sources of potential bias identified in NRLS-related articles identified by systematic review.

- Anonymisation** - Data submitted to the NRLS are anonymised to protect patient confidentiality, and direct identification of patients was deemed unnecessary for the NRLS to function as a learning tool. This, however, prevents analysis of repeated incidents with the same patient or staff member, the tracking of incidents across care settings, and linkage of local data to NRLS to retrieve further details (Panesar et al., 2012a). This disproportionately affected certain studies where repeated behaviours/incidents were involved e.g. airway/choking incidents (Guthrie et al., 2015), self-harm (James et al., 2012) or attempted suicide (Bowers and James, 2011). A high number of choking incidents across many patients could be considered a different signal to repeated choking incidents for few patients.
- Classification of harm** – Only severe harm or death incidents are mandatory to report from 2010 onwards. Low harm incidents may not be perceived as incidents at all, and their reporting is still not mandated. One study commented on this, stating “...either there were no clinical complications ..., or that our data is not representative of all ... complications occurring during the time period examined. The answer is most certainly the latter” (Innes and Curtis, 2013). Although the NPSA guidance suggested using established grades of harm for analysis, many incidents appear to be mis-graded with some studies re-grading a proportion of their data. A common criticism by authors was that reports classified potential harm rather than actual harm. It has been suggested that a quarter of anaesthetic incidents were mis-graded (Baird and Smith,

2009) and severity of incidents differed when a specialist reporting system was compared to the NRLS (Guthrie et al., 2015).

- **Classification of incidents** – As with specialty, incident type is also inconsistent. 57% of wrong-site surgery reports in orthopaedics were suggested to be misclassified, with no wrong-site surgery occurring (Panesar et al., 2011). Some authors even reclassified their dataset during or prior to analysis to account for this (Thomas et al., 2009). NRLS classifications were also suggested to be insufficient to capture chronological elements in reports (Carson-Stevens et al., 2015).
- **Lack of detail** - Word count in the free-text incident descriptions varied considerably. A median word count of 20, ranging from a single word to hundreds of words was noted in a text-mining study (Bentham and Hand, 2012). Some entries represented error messages, single full stop characters, or series of 'x' characters. Some articles used word count as a proxy of reporting 'quality' (Sevdalis et al., 2010, Scott-Warren et al., 2012). Although a clear description may contain only a few words, it is unlikely that very short descriptions convey incidents adequately. Conversely it is not necessarily true that a long report is of high 'quality' and may contain spurious or unnecessary information such as software error messages. A lack of detail has prevented many reports being used effectively by researchers (Panesar et al., 2012a).
- **Missing data** – Many NRLS fields are non-mandatory and their completeness varies (Hignett, 2013). The age field was found to be populated in only 62% of eligible cases in one study (Martinez et al., 2011), and degree of harm populated 66% in another (Thomas et al., 2009), despite this being mandatory. If data are missing 'completely at random', imputation methods may be used. In the case of NRLS, we cannot determine if data are unreported, deliberately omitted, missing in error, or represent data mapping problems. Missing data may severely undermine the validity of extracted data and analyses, and imputation methods were not suitable for the studies. MacLennan and Smith (2011) suggested that missing data also affects case ascertainment, potentially excluding/including the wrong cases in studies.
- **Search terms** – free-text entries and clinical judgement appear to be the most appropriate way to identify incidents and avoid mis-classification. There is no standard validation of the free-text entered. It contains acronyms, medical terms, inconsistent names, spelling mistakes and corrupted entries (Bentham, 2010). Stronger articles considered multiple terms and anticipated different classifications, e.g. consulting SNOWMED terms (Arnot-Smith and Smith, 2010), using word variants (Catchpole et al., 2008, Milligan, 2012, Rutter et al., 2014, Booth et al., 2011), anticipating mis-spellings

(Kelly and Jalil, 2011, Kelly and Barua, 2011) or using specific mappings of terms (Cousins et al., 2012).

- **Under-reporting** - all incident reporting systems are prone to bias from under-reporting (Panesar et al., 2009). Many articles directly acknowledge this, but almost all the studies are affected, making estimations of harm rates or prevalence inconsistent and non-generalisable.
- **Volume of data** – The volume of incident reports in the system is overwhelming for researchers reviewing the text. This high volume often led authors to choose pragmatic samples. For example, a study of hydration incidents identified 7,856 incidents, and chose to review all 142 deaths and 257 severe harm incidents and 50 of each other harm category (Lecko and Best, 2013).

Estimated Harm rates									
Authors	Year	Records examined	Period	Death	Severe	Severe or Death	No Harm	Not Estimated	Summary
Acupuncture									
Wheway J, Agbabiaka TB, Ernst E(2012)	2012	325 of 468	Jan-09 - Dec-11	0.31%			63.38%	FALSE	Major themes included retained needles and extended treatment time. Dizziness, loss of consciousness and referral to ambulance crew or A&E in some cases.
Airway									
Thomas AN, McGrath BA(2009)	2009	1085	Oct-05 - Sept-07			10.10%	39.00%	FALSE	28.8% airway device incidents related to neonates or babies, 71% in adults. 82% were post-procedural problems. Partial displacement resulted in more harm than total displacement. Although less frequent than medication or equipment incidents, associated with higher degrees of harm.
Robertson JA, Smith AF(2010)	2010	1885 of 195810	13/01/2006 - 03/04/2009					FALSE	Airway incidents with devices and intubation often related to poor dentition, difficult intubation, difficult airway, operator inexperience, and equipment failures. Deficiencies in pre-operative assessment, equipment provision, information and skills implicated.
McGrath BA, Thomas AN(2010)	2010	453 of 968	Oct-05 - Sept-07		18.00%			FALSE	Management of tracheostomies on wards can lead to more incidents of harm, primarily due to lack of infrastructure, appropriate airway equipment and skills for staff dealing with patients.
McGrath BA, Thomas AN(2011)	2011	494	Oct-05 - Sep-07			18.00%		FALSE	Risk of harm from post-placement tracheostomy incidents higher in ward settings compared with others. Availability of appropriate equipment and skilled staff in managing patients on ward was highlighted.
Templeton R, Webster K, McGrath BA(2011)	2011	494	Oct-05 - Sept-07					FALSE	Tracheostomy incidents in ward settings have significantly higher rate of harm. Differences in practice between settings confound result as ICU and ward tracheostomies and protocols differ.
Guthrie S, Lecko C, Roddam H(2015)	2015	436	Jan-10 - Dec-10			10.60%		FALSE	Choking hazards in mental health and learning disabilities do not correlate well between local and national systems. Major choking hazard information such as food type or behavioural factors not routinely provided to NRLS.

Estimated Harm rates									
Authors	Year	Records examined	Period	Death	Severe	Severe or Death	No Harm	Not Estimated	Summary
Ambulance									
Fisher JD, Freeman K, Clarke A(2015)	2015	Un-specified	Apr-10 - Sept-10, Oct-10 - Mar-10, Apr-11 - Sept -11					TRUE	Improved reporting by ambulance trusts over time, linked to their size, but highly variable. Major incident categories were 'access, admission, transfer and discharge', 'patient accident,' 'medical device/equipment.' Authors suggest low reporting rates may reflect poor safety culture rather than few incidents.
Anaesthesia									
Catchpole K, Bell MDD, Johnson S (2008)	2008	12,606	Jan-04 - Feb-06			2.10%	75.30%	FALSE	Reasonably high harm rates in anaesthesia, with largest groups related to inappropriate or delayed treatment or 'other' suggesting poor classification fit.
Catchpole K, McCulloch P(2009)	2009	12,649	Jan-04 - Feb-06			2.00%	75.00%	FALSE	Most incidents low harm but higher harm rate in epidural anaesthesia. Highest groups treatment/procedure and infrastructure /equipment.
Baird M, Smith A(2009)	2009	4,900	Jan-06 - Mar-06	0.30%	1.10%	1.30%	77.40%	FALSE	40% of incidents misclassified by specialty and 25% misclassified harm, with 20% over-estimation and 5% under-estimation 5%.
Arnot-Smith J, Smith AF(2010)	2010	231	2006 - 2008	0.40%	5.00%	6.00%	31.00%	FALSE	Anaesthetics incident reports well identified, often misclassified under surgery (40%). Non-availability of drugs, unintentional awareness and allergic reaction major incident types.
Cassidy CJ, Smith A, Arnot-Smith J(2011)	2011	1,029	2006 - 2008	0.00%	0.50%	0.50%	89.00%	FALSE	Majority of anaesthetic equipment errors due to failure of equipment, but user error and unfamiliarity implicated. Checklists recommended.
MacLennan AI, Smith AF(2011)	2011	606	Jan-06 - Dec-08	1.00%	7.90%	8.90%		FALSE	In-depth report classifying causal issues. Categorised medication incidents, primarily due to administration and dosage, airway incidents, particularly induction of anaesthesia, artificial airway, haemorrhage, disconnection of equipment, failure of monitoring equipment and duplication of reference data.
Scott-Warren J, McPherson D, Mahajan R et al(2012)	2012	318 local vs. 318 national.	Oct-09 - Sept-10					FALSE	Comparison of specialty-specific reporting portal (SSP) with NRLS data for same organisation. Degree of harm lower in SSP, reviewers judged harm, specialty and incident category allocation superior in SSP, but still deficient in many cases.

Estimated Harm rates										
Authors	Year	Records examined	Period	Death	Severe	Severe or Death	No Harm	Not Estimated	Summary	
Anaphylaxis										
Worth A, Panesar S, Healy L et al.(2012)	2012	1858	2005 - 2010					TRUE	Most incidents involved exposure in patients who were known to be allergic. Often involved antibiotics, anaesthetics analgesics and contrast media. Lack of knowledge, route of administration, equipment failure, failure to recognise deterioration and documentation were main groups.	
Cardiac arrest										
Panesar SS, Ignatowicz AM, Donaldson LJ(2014)	2014	30	Jun-10 - Oct-12	100%				FALSE	Cardiac arrest death incidents suggest miscommunication involving crash number, shortfall in staff attending arrests, equipment deficit, and poor application of knowledge/skills.	
Chest Drain										
Akram AR, Hartung TK(2009)	2009	2152	Jan-05 - Mar-08	0.50%	0.70%	1.30%	63.80%	FALSE	Incorrect placement of chest drains major source of harm, reporting 0.5% incidents resulting in death, primarily from puncture of solid organs.	
Critical Care										
Thomas AN, Galvin I(2008)	2008	1021 of 12084	Aug-06 - Feb -07			3.00%	69.00%	FALSE	18.1% involved pumps/infusion devices, 16.1% ventilators, 10.5% haemofilters and 7% monitoring equipment. Failure or faulty equipment most common but incorrect setting or use also common.	
Thomas AN, Panchagnula U, Taylor RJ(2009)	2009	5615 of 6649	Jan-08 - Mar-08			0.94%		FALSE	Most common incident groups were medication (25.8%), infrastructure/staffing(23.0%) and implementation of care (18.6%). Communication between professional teams, documentation and processes of transfer in and out of ICU implicated.	
Wahr JA, Shore AD, Harris et al(2014)	2014	2837	2003 - 2008	0.07% UK 0.03% US	0.81% UK 0.16% US		80.71% UK 84.43 US	FALSE	Difference between UK and US: wrong does (44% vs. 29%), omitted doses (8.6% vs, 27%). Gentamicin cited more frequently in UK. Heparin, insulin, potassium, and opioids frequently cited in moderate, severe harm or death in both countries. Incident rate higher in prescribing in US and administration in UK, but 'correcting' role of UK pharmacists without reporting as incident speculated. Similar harm rates suggested.	

Estimated Harm rates									
Authors	Year	Records examined	Period	Death	Severe	Severe or Death	No Harm	Not Estimated	Summary
Dentistry									
Thusu S, Panesar S, Bedi R(2012)	2012	2012	2009					TRUE	Low reporting rates in dentistry. Commonly clerical error, injury to lip (often using bur) main incidents. Small number of medical emergencies noted, often due to underlying condition or ingestion of hypochloride, bur, crown or bridge. Surgical checklists suggested.
Dermatology									
Gawkrödger DJ(2011)	2011	394	Jan-05 - Sept-09					FALSE	Dermatology incidents included drug prescribing, monitoring and follow-up, particularly isotretinoin. Phototherapy comparatively highly reported with excessive treatment duration wrong device and dose error noted.
Diagnostics									
Sevdalis N, Jacklin R, Arora S et al(2010)	2010	1674	Nov-03 - Oct-05				55.00%	FALSE	Harm rate in diagnostic incidents is higher than in non-diagnostic incidents. Predominantly occurring on hospital wards, but significantly fewer in wards than non-diagnostic, with diagnostic incidents also more likely in emergency department.
Dialysis									
Rylance P, Fielding C, Hutchison A et al(2015)	2015	94	12-month period between 2007 - 2015	70.21%	29.79%			TRUE	12-month study in nephrology units revealed 94 severe harm or death incidents with 40% related to management of patient including delay in medical or nursing care. Major theme was related to haemorrhage and infection related to fistulae and dialysis catheters. Dislodgement of venous needles highlighted, and under-reporting suggested.
Falls									
Healey F, Scobie S, Oliver D et al(2008)	2008	206,323 of 206,350	Sept-05 - Aug-06		0.60%	0.60%	64.70%	FALSE	Falls are major category of error at 32% reports, 82.6% of which in over 65s, with peak times between 10:00am and 11:50am. Falls rate per 100 beddays varies substantially between settings, and estimated to cause 11,265 - 12181 lacerations, 447 -626 fractured neck of femurs and 281-512 other fractures.
Hignett S, Sands G, Griffiths P(2011)	2011	6,577 of 44,202	Sept-06 - Aug-07					FALSE	70% falls unwitnessed and location information sparse. Most falls at bed/chair with different patterns observed in 'frail' or 'confused' patients.
Hignett S, Sands G, Griffiths P(2013)	2013	19,890 of 20,036	Sept-05 - Aug-08				67.00%	FALSE	78% falls unwitnessed and location available in only 47% of incidents. Falls in bed space less likely to result in harm than other locations. 'Frail' or 'confused' display different patterns of harm.

Estimated Harm rates									
Authors	Year	Records examined	Period	Death	Severe	Severe or Death	No Harm	Not Estimated	Summary
Incident Reporting					Incident Reporting				
Shaw R, Drever F, Hughes H et al(2005)	2005	28,998	Sept 01 - Jun-02			0.45%		FALSE	NRLS pilot/feasibility study. Main groups 41% slips trips and falls, 9% medication, 8% resource issues, 7% treatment issues. Different grading of harm used. Feasibility demonstrated, but interoperability of systems poor.
Hutchinson A, Young TA, Cooper KL et al(2009)	2009	Not stated	Apr-04 - Nov-05					TRUE	Highest reporters showed lower proportions in slips trips and falls. Correlations with higher reporting rates and positive data on safety from staff survey and better risk-management rating from NHS Litigation Authority. No significant correlations with other incident types, severity, staff survey questions (except observing a recent error), MRSA bacteraemia, HSMR, death in low-mortality HRGs, pressure ulcers or sepsis or within casemix factors.
Pham JC, Colantuoni E, Dominici F et al(2010)	2010	104,674	2006					FALSE	Calculated Harm Susceptibility Ratio (HSR) for 20 trusts and 12 work areas, with 55% of harm attributed to variation between trusts. Variation in work areas notable within trusts. HSR suggested for capturing within trust variation, with A&E, radiology and therapy consistently showing highest probability of harm.
Donaldson LJ, Panesar SS, Darzi A(2014)	2014	2010	Jun-10 - Oct-12					TRUE	Deaths suggest mismanagement of deterioration, failure to prevent falls, infections etc., medication error, delayed test results, dysfunctional patient flow and equipment errors such as unavailability or misuse.
Carson-Stevens A, Hibbert P, Avery A et al(2015)	2015	13,332	Apr-05 - Sep-13			1.90%		FALSE	Protocol report for analysis of general practice incidents. Summary incident figures reported with 42, 729 reports from general practice. Article presents a comprehensive and rigorous analysis plan and process for developing the qualitative framework 'Primary Care Patient Safety (PISA) Classification System,' to analyse these types of reports.
Howell AM, Burns EM, Bouras G et al(2015)	2015	5,879,954	Jan-03 - May-13			0.90%	70.30%	FALSE	Elderly medical patients most vulnerable to harm, but reporting rate and harm rate vary by specialty. No significant correlations with hospital size, mortality, or patient satisfaction. Significant correlations with harm rate and clinician staffing, as well as litigation ratios. Clinicians more likely to report death, but lower overall reporting rates than other staff. Suggests reporting rate should not be used to assess quality.
Medication									
Stubbs J, Haw C, Dickens G(2008)	2008	17 of 767,716	Nov-05 - Nov-06			0.00%	64.00%	FALSE	Tablet crushing incidents reviewed with none reported from mental health. Modified release opiate and cytotoxic drugs main groups. Lack of staff knowledge of what should not be crushed and poor communication to patients highlighted. Under-reporting suggested.

Authors	Year	Records examined	Period	Estimated Harm rates					Summary
				Death	Severe	Severe or Death	No Harm	Not Estimated	
Thomas AN, Panchagnula(2008) U	2008	2,428 of 12,084	Aug-06 - Feb-07			0.35%		FALSE	Medication incidents in critical care setting commonly involved morphine, gentamicin & noradrenalin. Noradrenalin and Insulin most commonly associated with harm. 61% incidents associated with administration. Incorrect prescriptions identified after patient contact in 50% incidents.
Mahajan R, Mathews L, Russell J et al(2009)	2009	157	Jan-05 - Mar-08					TRUE	80.3% of incidents 'wrong drug' related to administration, with 7.6% prescribing, 8.9% preparation. Little harm observed, but most preventable with adequate checking process.
Laker MF(2009)	2009	147	Nov -07 - Jan-08 Sept-07 - Nov-07	0.00%			88.00%	FALSE	NRLS reporting of anticoagulant incidents used as proxy for increased safety culture. Lag time to report did not significantly reduce. Communication, administration, testing and prescription errors were leading causes.
Milligan FJ, Krentz AJ, Sinclair AJ(2011)	2011	768	Jan-05 - Dec-09	0.13%				FALSE	Incidents concerning insulin and oral glucose-lowering agents in care home setting were usually low harm and related to incorrect dosing, frequency or omitted doses. No compulsion for care homes to report NRLS, so likely under-estimate.
Cousins D, Rosario C, Scarpello J(2011)	2011	16,600	Nov-03 - Nov -09			0.07%	76.00%	FALSE	High proportion of insulin incidents report harm, occurring mainly during administration, prescribing and dispensing. Delayed/missed administration, dosage errors or wrong insulin product accounted for 60% of incidents.
Cousins DH, Gerrett D, Warner B(2012)	2012	526,186 of 5,437,99	Jan 05 - Dec-10	0.05%	0.10%	0.90%	83.00%	FALSE	Medication incidents represent ~10% of incidents. Low degree of harm in general, but large number of incidents suggest many administration, prescribing and dosage incidents occur.
Franklin BD, Panesar SS, Vincent C et al(2014)	2014	38 of ~9 million	Nov-03 - May-13				100%	FALSE	Vinca alkaloid incidents did not lead to harm but identified incidents in storage, timing, location and potential for confusion between IV and intrathecal medication. Low harm reports can provide information on risks of rare but serious events before they happen.
Innes J, Curtis D(2013)	2015	28	Aug-10 - Jul-11				100%	TRUE	Small number of rapid tranquilisation incidents reported, mainly within mental health units related to administration, prescribing, drug availability and decision to administer. NRLS insufficient to monitor these incidents properly.
Flood C, Matthew L, Marsh R et al(2015)	2015	Unclear	Oct-02 - Nov-08 Jun-09 - Aug-10					TRUE	Reduction in midazolam incidents reported after NPSA RRR, but compared 74 months of all incidents to 15 with no further severe harm of death incidents reported (with no denominator). Gap between reporting periods not adequately justified.

Authors	Year	Records examined	Period	Death	Estimated Harm rates				Not Estimated	Summary
					Severe	Severe or Death	No Harm			
O'Grady I, Gerrett D(2015)	2015	1,882	Jan-05 - Dec-13	0.32%					TRUE	Missed doses are common in patients who are nil by mouth or have swallowing problems. Major reasons include prolonged fasting due to overrunning theatre lists, lack of awareness or availability of alternative administration routes, delays/lack of assessment of swallowing and communication of risk.
Nutrition/Hydration										
Lecko C(2010)	2010	1,433 and 897	2006 -2007 and 2008						TRUE	Nutrition related incidents represented 20% and 23% of incidents respectively in searches. Incidents often related to artificial feeding, patients being 'nil by mouth' and oral feeding. Themes included poor communication and documentation particularly in transfer/ admission/discharge and fasting for theatre, inadequate staff training or awareness, lack of nutrition service or assessment, and failure to follow protocol or implement changes of feeding/fluid.
Lecko C, Best C(2013)	2013	368 of 7,856	2003 - 2012	1.80%	3.30%	5.10%			FALSE	Hydration incidents often reported. Poorly designed systems, lack of local guidance and failures of recognition, implementation, poor awareness, and excessive demands on staff.
Obesity										
Booth CM, Moore CE, Eddleston J et al(2011)	2011	388 of 555 identified.	Jan-05 - Aug-08	1.00%	1.00%	2.00%	86.00%		FALSE	Anaesthetic incidents had higher rates of harm due to difficulty intubating or maintaining airway. Unavailability or failure of bariatric equipment highlighted. Haemorrhage, DVT/PE, unintended damage, recognising complications and wound breakdown major groups. Medication incidents related to dosage common for heparin/warfarin.
Obstetrics										
Sandall J, Watson K, Wiseman O(2012)	2012	9,121	2009	3.70%					FALSE	264 potentially avoidable factors grouped into 10 themes. Major themes were failure or delay in monitoring, diagnoses or assessment, failure to recognise deterioration and concerns with resourcing, staffing and equipment.
Ophthalmology										
Fetherston T(2007)	2007	144 of 3,127 incidents	Mar-01 - Mar-06	0.70%	8.30%	9.00%	37.00%		FALSE	Treatment procedure incidents including complications and equipment, documentation incidents and patient accidents major causes. Complications of cataract surgery leading type.
Kelly SP, Barua A(2011)	2011	166	2003 - 16 Jun 2010.		6.00%		81.00%		FALSE	6% incidents severe harm, but majority low. No near misses reported. Major themes were treatment delay, missing records, prescription and severe inflammation with ranibizumab.

Authors	Year	Records examined	Period	Death	Estimated Harm rates			Not Estimated	Summary
					Severe	Severe or Death	No Harm		
Kelly SP, Jalil A(2011)	2011	164	2003 - Jan-10					TRUE	No near-misses reported in intraocular lens transplant suggesting lack of awareness. Inaccurate biometry, wrong lens selection were leading causes but not specified in 37.8% cases. Many cases required further surgery.
Orthopaedics									
Robinson PM, Muir LT(2009)	2009	79	Mar-05 - Jun-07	0.00%				FALSE	After correct site surgery guidance introduced, 79 wrong-site surgery incidents in orthopaedics reported. Anaesthetic blocks were most common procedure, unclear what stage most were noticed.
Panesar SS, Noble DJ, Mirza SB et al(2011)	2011	116 of 316	2008				91.00%	FALSE	Wrong-site surgery reports read and classified. 42% met inclusion criteria. 58% miss-classified. 9% harm, 91% near-miss. Smaller proportion of near-misses being prevented by checklist than those that result in harm. Estimated that checklist could have prevented incidents in 21.1% of cases.
Panesar SS, Simunovic N, Bhandari M(2012b)	2012	4,521	Since 2003...' paper published in 2011				4.00%	FALSE	96% of patients suffered some form of hip-fracture from surgery being delayed.
Panesar SS, Carson-Stevens A, Mann BS et al(2012a)	2012	257	2005 - 2009	100%				FALSE	Many deaths could not be properly assessed due to lack of detail, but of those that could 32% died in relation to infection, 44% had failure in non-technical skills mainly related to situational awareness. Recommended improved checklists, protocols and trigger tools.
Panesar SS, Carson-Stevens A, Salvilla SA et al(2013a)	2013	48,095 of 163,595	Jan-09 - Dec-09	0.15%			69.90%	FALSE	Largest specialty contributing to surgical events (29.4% incidents). 30.1% resulted in harm. Major categories included implementation of care, monitoring, self-harming behaviours and infection control. High proportions of harm in several categories including patient accidents.

Authors	Year	Records examined	Period	Estimated Harm rates					Summary
				Death	Severe	Severe or Death	No Harm	Not Estimated	
Panesar SS, Netuveli G, Carson-Stevens A et al(2013b)	2013	48,971	2009 - 2010	0.10%	0.40%	0.50%	70.50%	FALSE	Orthopaedic error index developed, with mean value of 7.09/year. 5 of 155 hospitals identified as outliers, 3 tertiary centres carrying out complex surgery and two others unusually high.
Rutter PD, Panesar SS, Darzi A et al(2014)	2014	360	2005 - 2012	11.30%	6.00%	17.20%		FALSE	High proportion of deaths with incidents related to bone cement during hip hemiarthroplasty. Cardiac arrests and periarrest leading cause. Most reports describe acute deterioration during or immediately after cement insertion.
Paediatrics									
Rees P, Carson-Stevens A, Williams H et al(2014)	2014	2,347	None specified.					FALSE	Paediatric vaccination incidents quoted in response to another article. Wrong number of doses 38.3%, wrong vaccination 28.5%, and wrong timing 17.2%.
Rees P, Edwards A, Panesar S et al(2015a)	2015	1,788 of 46,902	Apr-03 - Jun-12	0.40%	0.50%	0.90%	57.30%	FALSE	Paediatric family practice reports, 42.7% described harm to children. Vaccination incidents made up high proportion. Priority areas identified, due to high harm rates included: diagnosis/assessment, treatment procedure, referral issues and medication provision.
Rees P, Edwards A, Powell C et al(2015c)l	2015	2,191	2003 - 2013					TRUE	Priority areas for improvement include staff mistakes, knowledge, failure to follow protocols and organisational factors (e.g. inadequate protocols, service availability). Provision in community pharmacy, diagnosis, assessment and timely referral of acutely unwell patients during out-of-hours and communication with and about a child highlighted.
Rees P, Edwards A, Powell C et al(2015b)l	2015	1,745	2002 - 2013	0.17%			38.28%	FALSE	Vaccination incidents predominantly nursing related, often leading to harm as child had extra vaccinations. Major themes were incorrect dosage, vaccine or number of vaccinations, often due to poor documentation.
Omar A, Rees P, Evans HP et al(2015)	2015	1,242	Not specified					TRUE	Fragmentation of care services major factor in incidents concerning vulnerable children including poor transfer or information. Breakdown in consent communication and care in 52.3% of cases.

Estimated Harm rates									
Authors	Year	Records examined	Period	Death	Severe	Severe or Death	No Harm	Not Estimated	Summary
Resuscitation									
Flood C, Gull N, Thomas B et al(2014)	2014	39 of 1,166	Jan-06 - Mar-08 Nov-08 - Apr-10	2.00%	1.40%	3.40%		FALSE	Borderline significant reduction in resuscitation incidents in mental health and learning disabilities settings after NPSA Rapid response alert compared to before alert, but biased reporting period and underpowered analysis with multiple comparisons. Gap between reporting periods not adequately justified.
Hawkes C, Chambers S, Satherley P et al(2015)	2015	4,538	Not specified	3.10%	13.00%			FALSE	DNACPR incidents were predominantly in secondary care, but also in ambulance and community hospitals. 16% resulted in severe harm or death. Communication with patients/relatives, record keeping, clinical review after change in patient status and processes around DNACPR decisions (including requesting, implementing and communicating this information between services) were implicated.
Self-Harm									
James K, Stewart D, Wright S et al(2012)	2012	448 of 14,271	Jan-09 - Dec-09		2.00%		30.00%	FALSE	3 times as many self-harm incident in women as men but anonymisation confound repeated events. Significantly higher rates in forensic services when weighted on beds. Men's methods were more outwardly aggressive than women's. Conflict with staff frequently mentioned as a trigger.
Staffing									
Francis R(2013)	2013	940	2005 - 2010					FALSE	NRLS data used as a proxy to understand Mid-staffs safety culture. Staff shortages in and inadequate skills highlighted. Significant under-reporting with A&E analysis suggesting lack of awareness or learning culture in A&E, with increased reporting in later periods indicating trust's attempts to improve.
Suicide									
Bowers L, Dack C, Gul N et al(2011)	2011	602 of 711	Jan-09 - Dec-09					TRUE	Higher rates in women but confounded by excluding successful suicides, with higher rate in men, and anonymous data does not show repeat attempts. Mainly strangulation attempts, higher at night and in acute psychiatric wards, but may reflect under-reporting of other settings.

Authors	Year	Records examined	Period	Death	Estimated Harm rates				Summary
					Severe	Severe or Death	No Harm	Not Estimated	
Surgery									
Martinez EA, Shore A, Colantuoni E et al(2011)	2011	4828	Jan-03 - Feb-07				72.00%	FALSE	Cardiac surgical incidents outside the operating room have a higher rate of harm compared to those reported in the operating room. Distribution of incident types varied between settings, with medical equipment incidents 6 times more likely in operating room and treatment procedure incidents 3.7 times more likely.
Rocos B, Donaldson LJ(2012)	2012	13	Mar-04 - Mar-11		0.80%			FALSE	Small number of surgical fire incidents with 84% judged to be misused of equipment causing ignition. Most cases due to not allowing skin preparation to dry or drapes/swabs soaked with fluid too close to surgical field.
Barai I, Howell AM, Burns E et al(2014)	2014	703	Not stated					TRUE	69% of surgical incidents may have been classified in a different manner. NRLS classifications inflexible and blunt learning.
Transfers									
Williams H, Edwards A, Hibbert P et al(2015)	2015	598	Apr-03 - Jun-12	0.16%	0.50%	0.67%	15.22%	FALSE	Deficiencies in the discharge processes led to significant harm. Communication between secondary and primary care, particularly around referral processes & medication, with poorly designed or poorly executed protocols and missing information.

Table 2.4 Evidence Summary tables from NRLS literature review

‘Not Estimated’ refers to studies where proportion of incident resulting in harm was not estimated, with TRUE meaning harm was not estimated. Where ‘Not Estimated’ is FALSE, but a value is missing, proportions of at least one, but not all, harm levels were reported.

2.3.2 Selected topic-specific themes:

A high volume of information was extracted from the literature related to specific clinical or organisational settings. Multiple studies were identified under common themes, with the twelve most common themes summarised below. The remaining review data are summarised in section 2.3.3, due to the length of this review.

2.3.2.1 General Incident reporting

Five articles used incident reporting figures to assess reporting, deaths, and correlations with other measures.

In the NRLS pilot study (Shaw et al., 2005), staff from 18 NHS trusts including acute, mental health and ambulance trusts, as well as primary care, took part in a prospective voluntary scheme during 2001-2002. The main incident group identified was 'slips, trips and falls' at 41% of reports, with other major groups including 'medication', 'resource issues' and 'treatment issues.' The grading of harm differed from the final version of the NRLS and interoperability with other systems was deemed to be an issue, but the article demonstrated the feasibility of data collection.

Two studies assessed correlations of incident reports with other measures. Data between April 2004 and November 2005 revealed that the highest reporting organisations showed lower proportions of incidents in the 'slips, trips and falls' group (Hutchinson et al., 2009). Correlations were observed between higher reporting rates and positive responses on safety culture questions from the NHS Staff Survey and better risk-management ratings from the NHS Litigation Authority. No significant correlations with other incident types, severities, MRSA bacteraemia, the Dr Foster Hospital Standardised Mortality ratio (HSMR), death in low-mortality HRGs, incidence pressure ulcers or sepsis, were observed.

Rates of harm and correlations to other measures were examined between January 2003 and May 2013 (Howell et al., 2015). Elderly medical patients were most vulnerable to harm but reporting rates and harm rates varied by clinical specialty. No significant correlations with hospital size, the Summary Hospital Mortality Index (SHMI), or patient satisfaction measures were observed. Significant correlations with harm rate and clinical staffing, as well as litigation ratios were identified. Authors suggested that clinicians were more likely to report death incidents but showed lower overall reporting rates than other staff groups.

Variance in harm rates between trusts and across specialties within trusts were examined in 20 hospitals between 2002 and 2004 by Pham et al.(2010). Authors suggested 55% of variation in harm rate was attributed to variation between trusts, but highlighted work areas where the

probability of harm was highest within (rather than between) trusts, including A&E, radiology and therapy.

Incidents leading to death between 2010 and 2012 were specifically examined and suggested mismanagement of deterioration, failure to prevent falls, infections, medication error, delayed test results, dysfunctional patient flow and equipment errors including unavailability and misuse as major causes of incident reports (Donaldson et al., 2014).

2.3.2.2 Medication

Twelve articles examined medication, the largest subject area in this review. Reports were predominantly low harm ($\approx 75\%$), but their high frequency ($\sim 10\%$) represents a significant number of incidents (Cousins et al., 2012). Incidents were often described during drug administration, prescribing or in calculating drug dosage. A study related to 'wrong drug' errors suggested $\approx 80\%$ related to administration and others related to missed or wrong doses (Mahajan et al., 2009). In critical care, administration errors accounted for 61% of medication incidents, including incorrect recording, incorrect rates of infusion and missed doses (Thomas and Panchagnula, 2008).

Laker (2009) described a pilot study using NRLS reporting rates for selected medication errors (specifically mentioning anticoagulants), with missing blood results and communications failure common problems. The study assessed feasibility of using NRLS data to monitor a patient safety programme but made no clear assessment of its success. A similarly vague picture was described in relation to rapid tranquilization (RT) incidents that were a priority for the NHS Litigation Authority (Innes and Curtis, 2013). The study identified very few incidents, none of which led to harm or adverse effects, suggesting that there are either no complications from RT or, more likely, under-reporting meant their data was not representative of all RT complications (as previously quoted in section 2.1).

Medication errors in critical care commonly involved morphine, gentamicin or noradrenalin (Thomas and Panchagnula, 2008). The highest proportion of harm was observed in noradrenalin and insulin reports. Communication between staff when transferring patients in or out of critical care was a common cause of errors.

Vulnerability to error was observed in prescribing, communicating, preparing and administering drug treatments. Missed doses were a major area of concern in patients who were nil by mouth, have swallowing problems, are on prolonged fasts due to over-running theatre lists, or for those who experience delays/lack of assessment of their swallowing (O'Grady and Gerrett, 2015). Staff were not always aware of alternative administration routes for drugs. Confusion over routes of administration was a common theme noted in

administration of vinca alkaloids where inadvertent intrathecal administration is potentially fatal (Franklin et al., 2014), in relation to tablet crushing leading to inaccurate dosing in mental health settings (Stubbs et al., 2008), and in relation to insulin (Cousins et al., 2011).

Insulin was also a theme for two other studies. With no compulsion for care-homes to report to NRLS, reports in these settings were primarily made by visiting NHS practitioners, commonly relating to wrong doses, frequency of administration or missed doses (Milligan et al., 2011). Cousins et al. (2011) discussed a similar pattern, highlighting the potential for wrong insulin products and staff unaware of the differences in hospital. Major causes included missing prescriptions on admission, poor use of abbreviations, decimal errors, incorrect monitoring and dose adjustment, duplicate doses, errors in intravenous dose calculations, and poor documentation. 'Look-alike' and 'sound-alike' drugs were also highlighted as problems.

2.3.2.3 Anaesthesia

Seven articles focussed on incidents in anaesthesia. Approximately 2% of anaesthetic reports led to severe harm or death, which was consistent across studies (Catchpole et al., 2008, Catchpole and McCulloch, 2009). These articles suggested that the harm rate, of approximately one quarter of incidents, was relatively high and reflected comparatively high the opportunity for error and potential for harm in anaesthetic practice. Specialty classification was estimated to be inaccurate in 40% of anaesthetic incidents, and likely due to the variety of specialties involved in patient care, e.g. surgeons, anaesthetists, nurses or support staff (Baird and Smith, 2009). Incident type was estimated to be mis-classified in 40% of reports (Arnot-Smith and Smith, 2010), with major incident types being unavailable drugs, unintentional awareness in patients and allergic reactions. MacLennan & Smith (2011) and Robertson & Smith (2010) designed their subsequent studies with these misclassifications in mind, identifying both anaesthetic and surgical incidents containing relevant free-text, noting: *“data were often not adequate to determine whether the anesthetist made the error.”* Level of harm was classified incorrectly in approximately 25% of cases, with authors suggesting that many staff reported potential harm rather than actual harm. This poses the wider questions about the NRLS, namely: are the categorisations commonly understood and applied, and, can such classifications adequately describe incidents in their current form?

Major causes of anaesthetic incidents, common to most of the articles, were medication error (including duplicate administrations), equipment failure (primarily of monitoring equipment), non-availability of drugs or equipment, airway maintenance, and communication (including duplicated or missing records and non-communication of existing allergies). Cassidy et

al.(2011) suggested the main causes of incidents in anaesthesia were equipment failure, but user error and unfamiliarity with equipment were also common, leading them to recommend using checklists to standardised processes and guard against errors of omission.

The well-developed reporting culture, and known submissions to NRLS in anaesthetics, were used to validate an improved reporting procedure using a specialty-specific local reporting form (Scott-Warren et al., 2012). Authors suggested that degree of harm, specialty and incident category allocation was improved in the local specialty-specific system, but was still deficient in many cases.

2.3.2.4 Airway management

Six articles focused on airway management incidents. Major themes included a lack of staff availability, competency or equipment (Robertson and Smith, 2010). Lack of available airway equipment led McGrath (2010) to recommend the universal stocking of sterile 'airway kits' to accompany all admitted patients, suggesting that infrequent use would allow them to be passed on between patients, keeping the cost burden acceptable, but providing timely access to appropriate equipment.

Tracheostomy incidents indicated that partial displacement was more likely to lead to harm than total displacement (Thomas et al., 2009). Two studies suggested that apparent differences between critical care units and wards were insensitive to difference in tracheostomy protocol (McGrath, 2010, Templeton et al., 2011). Whilst higher harm was observed in ward settings, replacement of tracheostomy on a ward was more likely to be long-term maintenance, whereas replacement in critical care was more likely to become a trial removal (Templeton et al., 2011).

Choking reports in the NRLS were compared to an enhanced local reporting system, suggesting that a lack of information on choking incidents in the NRLS and significant under-reporting (Guthrie et al., 2015). Authors suggested choking incidents were being missed in patients with intellectual disabilities and that key information such as food type or behavioural factors were not present in the NRLS and limit the learning from it.

2.3.2.5 Orthopaedics

Seven articles focussed on orthopaedic reports, but a number of others showed overlapping themes. Orthopaedics was the most common specialty contributing to surgical incidents ($\approx 29\%$) (Panesar et al., 2013a) and was acknowledged to have a high harm rate ($\approx 30\%$) with

implementation of care, patient monitoring and infection control leading incident types. A study of incidents leading to death called for standardisation of practices and use of improved surgical checklists, protocols and trigger tools (Panesar et al., 2012a). Many deaths could not be properly assessed due to a lack of details, but data suggested 32% died in relation to infection and 44% showed failure in a non-technical skill (mainly situational awareness).

Reports related to delays in hip fracture surgery estimated harm in 96% of cases, although authors acknowledged the 'self-reporting' bias where a delay is less likely to be reported if patients experience good outcomes (Panesar et al., 2012b).

'Wrong site surgery' was a major theme, with one study branding orthopaedics as the "worst" specialty for wrong-site surgery between 1998 (Robinson and Muir, 2009). This study used NRLS reports to examine whether a patient safety alert for wrong-site surgery had been effective at reducing its occurrence. They identified 79 cases of wrong-site surgery with anaesthetic block being the most common procedure reported, which is not an orthopaedic surgical procedure but an essential part of pain management during surgery. These may be better be described as anaesthetics incidents and are subject to the misclassification issues described in 2.3.2.3. Panesar et al.(2011) further examined wrong-site surgery reports, identifying 58% of reports as mis-classified and were not, upon inspection, wrong-site surgery. Within the remaining incident reports, 9% led to harm, of which 21% were viewed as preventable if standard surgical checklists were used.

Rutter et al. (2014) used the size of the NRLS to their advantage to review rare events in relation to bone cement implantation syndrome. They reported a high proportion of deaths, especially involving cardiac arrests and peri-arrests, immediately or shortly after cement insertion during hip hemi-arthroplasty.

An orthopaedic error index was created to quantify error severity and propensity (Panesar et al., 2013b). The article presented an average of 7.09 incidents per hospital per year and identified 5 of 155 hospitals as outliers, including 3 tertiary centres carrying out complex surgery and two others with unusually high values.

2.3.2.6 Paediatrics

Five articles focussed on paediatric incident reports. Reporting to NRLS from general practice is generally lower than secondary care, but Rees (2015c) identified just 2191 reports between 2003 and 2013 featuring 'unwell children' in primary care. Paediatric vaccination was highlighted as a major source of incidents (Rees et al., 2014, Rees et al., 2015a). Incidents

were often high harm, and areas for further work and intervention were identified as staff mistakes, staff knowledge, failure to follow protocols, inadequate protocols or service provision. Provision of community pharmacy, diagnosis, assessment, timely referral of acutely unwell patients in 'out-of-hours,' and communication with (and about) children were also suggested as priorities (Rees et al., 2015c). Incidents in vulnerable children were particularly linked to fragmentation of care services and sharing of information (Omar et al., 2015).

A major mixed methods study on paediatric vaccination events in primary care, using structured clinical reviews and standardised coding to identify themes, suggested many errors were administration-based and showed significant correlation with vaccination type and severity (Rees et al., 2015b). Delays were predominantly due to incorrect dosage, incorrect vaccine or incorrect/missed timing that were implicated in three child deaths from meningitis and pneumonia. Documentation error linked to vaccination status was a major contributor, particularly for socially or medically vulnerable patients. Parental challenge and knowledge of vaccination records were sometimes cited as preventing incidents. Conversely some reporters described their expectations of parental knowledge to aid them when they encountered missing information.

2.3.2.7 Surgery

Three studies focussed on surgical practice not previously described as Orthopaedics in section 2.3.2.5. These studies overlapped with other subjects including orthopaedics and anaesthetics, paediatrics, or obesity such as surgical damage in obese patients (Booth et al., 2011). This uncertainty around classification was mentioned in several articles (Barai et al., 2014, Baird and Smith, 2009), suggesting 69% of surgical incidents had the potential to be classified in a different manner.

Surgical fires were identified as rare but harmful events, again using the NRLS' size to identify rare events. Alcohol-based skin preparations were the primary cause, either drenching drapes or not drying on patient's skin before being ignited by surgical instruments (Rocos and Donaldson, 2012).

Cardiac surgery incident rates were analysed by Martinez et al.(2011) suggesting higher harm rates when procedures were performed outside the 'operating room.' Distribution of incident types differed, causing authors to suggest a focus on reducing medical device/equipment errors in the operating room and on medication and patient accidents outside of the operating room.

2.3.2.8 Patient Falls

Three articles specifically examined incident reports related to falls. Falls represent one of the largest incident groups in the NRLS, representing 32.3% of incidents between September 2005 and August 2006 (Healey et al., 2008). Falls happen in a variety of settings and most frequent occur in the elderly, often in relation to activities like mobilising, transferring, washing/showering or visiting the bathroom. Most falls happened from beds or chairs (38%). 70% - 78% of falls were un-witnessed (Hignett et al., 2011, Hignett et al., 2013) and degrees of harm varied significantly between care settings with the highest rate of harm in mental health settings, although this may be an artefact of reporting culture (Healey et al., 2008). Falls occurred throughout the day but demonstrated a peak between 10:00 and 11:59, leading the authors to suggest elements of staff and patient routines as a contributing factor. Falls were estimated to have caused 21,124 lacerations, 528 fractured neck of femurs, and 442 other fractures, between September 2005 and August 2006.

Variation between general patients and those described as 'frail' or 'confused' was noted by Hignett et al.(2011), with contributory factors differing between groups. Bed rails were significantly more likely to be described in incidents where the patient was described as 'confused'. Patients who were described as 'frail' were less likely to experience 'no harm' incidents and more likely to be witnessed, whilst patients with 'confusion' were more likely to have unwitnessed falls (Hignett et al., 2013).

2.3.2.9 Ophthalmology

Three articles examined incidents in Ophthalmology. Vitreoretinal surgery was examined by Fetherston (2007), suggesting equipment, documentation, patient accidents and complications were the main incident types, with cataract surgery the most common procedure.

Kelly and Barua (2011) investigated incidents related to anti-vascular endothelial growth factor (VEGF) intraocular injections. The authors suggest a majority were low harm and often related to treatment delay, missing records, unavailability of drugs but also describe instances of severe inflammation.

Kelly and Jalil (2011) examined intraocular lens transplants suggesting that a lack of 'near-miss' reports in NRLS represented a lack of awareness of incidents and/or a poor safety culture. Inaccurate biometry and wrong lens selections were the leading causes, with many cases requiring further surgery.

2.3.2.10 Nutrition and Hydration

Two articles examined incidents related to nutrition or hydration. An article by Lecko (2010) was unclear in whether it represented primary research or summarised prior work, but was included on the basis it contained NRLS analysis. Incidents concerned both artificial and oral feeding as well as patients who were nil-by mouth. Communication problems and inadequate documentation were major causes contributing to failures to follow or make changes to protocols. Lack of nutritional assessment or nutritional services was highlighted as well as poor staff awareness of choking risk including thickening of drinks or appropriate dietary support and monitoring.

Lecko and Best noted many hydration incidents were avoidable and suggested that, as with nutrition incidents, a lack of staff awareness, excessive demands on staff, and poorly designed or poorly implemented local guidance was a major causes of incidents (Lecko and Best, 2013).

2.3.2.11 Self-harm and attempted suicide

Two studies that examined self-harm or attempted suicide both focussed on mental health settings. Both studies suggested notably higher rates in women than men, but both were confounded to a degree by their design and the NRLS' anonymisation. Examining attempted suicides, Bowers et al. (2011) excluded completed suicides to examine only failed attempts and noted higher rates in women, but also referenced a higher suicide 'completion' rate in men from another source. This was further confounded by the anonymisation of NRLS with repeated suicide attempts from the same patient recorded as new incidents, losing the valuable information of whether the higher rates in women represents repeated, unsuccessful, attempts. Anonymisation also affected the self-harm, study in mental health organisations by James et al. (2012), who observed higher rates in women, despite the prevalence of self-harm being similar. Authors suggest that higher rates of repeated self-harm in women may drive this, but NRLS was unable to provide evidence to support this. Self-harm incidents were significantly higher in forensic units when weighted on bed days, with men more likely to use outwardly aggressive methods of harm, and women more likely to restrict breathing. Suicide attempts were mainly strangulation attempts and were more common at night and in acute psychiatric wards. Conflicts with staff were often cited as antecedents.

2.3.2.12 Critical care

Two articles focussed on critical care settings, but a number of other articles referenced critical care but with different primary themes, e.g. medication (Thomas and Panchagnula, 2008), or airway incidents (Thomas and McGrath, 2009).

Medications errors and documentation errors were common, with the processes around transfer in and out of critical care implicated. Authors suggested that communication between critical care and other teams had the highest potential for incidents (Thomas et al., 2009).

Wahr et al.(2014) compared occurrence of medication incidents in critical care settings between UK and US hospitals. This study suggested errors occurring at different stages in each health system, with the UK reporting more wrong doses and the US reporting more missed doses. The UK reported higher rates of error with gentamycin and noradrenalin whilst US reported higher rates of error with salbutamol and insulin. Heparin, insulin, potassium, and opioids were frequently cited in moderate, severe harm or death incidents in both countries. Authors suggested cultural differences around the role of pharmacists were confounding results. US pharmacists are encouraged, and in some cases mandated, to report prescribing errors and UK pharmacists more likely to correct prescribing errors, seeing it as part of their job, without perceiving them as incidents.

Equipment incidents in critical care were examined by Thomas and Galvin (2008). A variety of equipment contributed to incidents: 18.1% pumps/infusion devices, 16.1% ventilators, 10.5% haemofilters and 7% monitoring equipment. Failure of the equipment and intermittent faults were common but practitioners using incorrect setting or using equipment incorrectly was also common.

2.3.3 Other topics

A variety of other topics were examined in single articles and are summarised below. Articles explored particular medical specialties or other cross-cutting themes.

Allergic reaction was noted in a number of settings related to medication and anaesthesia. Worth et al.(2012) suggest that allergies were often known and documented prior to incidents, with many related to antibiotics, analgesics and contrast media.

Communication was a major factor in several studies. Deaths from cardiac arrest suggested that confusion over 'crash call' phone numbers, lack of staff responders, and deficiencies in equipment or skills were the main themes (Panesar et al., 2014). Communication across services was a major component of incidents related to 'Do not attempt cardiopulmonary

resuscitation' (DNACPR) orders. Incidents in community hospitals, ambulances and secondary care showed high harm rates and suggested that staff were unsure of when and how to assess, implement and communicate about these orders with patients and each other (Hawkes et al., 2015). For communication incidents that related to discharge from secondary care, many cited the transfer of information between services (Williams et al., 2015). Continuity of medication and referral to community practitioners were major themes, with poorly designed or poorly executed protocols, and missing information identified as major themes.

Unintended damaged was identified as a risk to all patients during the placement of chest drains, causing high levels of harm, primarily from punctured solid organs (Akram and Hartung, 2009). Unintended damage was also a factor in the care of obese patients. Incidents overlapped other primary subjects, but often related to processes where the care of obese patients deviated from that received by other patients, e.g. dosage adjustment, availability of bariatric equipment, difficulties in maintaining airways, DVT/PE, or surgical damage (Booth et al., 2011).

Diagnostics errors were more likely to occur in emergency departments and less likely to occur on wards, with incidents of diagnostic error having a comparatively high harm rate (Sevdalis et al., 2010).

Incidents in obstetrics detailed failures or delays in monitoring patients, diagnoses, or assessments. Failure to recognise deterioration, and the availability of staff and equipment were primary concerns (Sandall et al., 2012). Staffing issues were common to many of the reports, with this interpretation support by the NRLS's contribution to the Mid-Staffordshire enquiry (Francis, 2013).

Dialysis incidents suggest that much of the harm is to do with patient management and often relates to haemorrhage or infection from fistulae and dislodgment of needles or catheters (Rylance et al., 2015).

Dermatology incidents included drug prescribing, monitoring and follow-up, particularly in relation to isotretinoin. Phototherapy was comparatively highly reported with excessive treatment duration, wrong devices and dose errors leading themes (Gawkrödger, 2011).

Incidents reports in dentistry suggested low reporting rates, likely due to NRLS' voluntary nature and practitioners' reluctance to disclose incidents (Thusu et al., 2012). Main incident types were clerical error, and injuries to lips, commonly related to use of a bur. A small number of medical emergencies were described, usually related to underlying condition or to

the ingestion of 'hypochloride' [sic], bur, crown or bridge. Authors suggests much of the harm occurring in dentistry could be controlled using adequate checklists.

2.4 Conclusions

This chapter has contributed to aim 1 of this thesis, identifying what types of learning have been derived from the NRLS data. Research articles published using the NRLS have a wide remit, and a variety of methods, that have mainly focussed on review of the free-text descriptions of incidents. Researchers have derived learning in many settings, but the lack of standardised methods reduces the validity and generalisability of results. There are signs of a systematic, standardised method gaining traction, with several qualitative analyses and a protocol published for primary care incidents (Carson-Stevens et al., 2015, Rees et al., 2015a, Rees et al., 2015b, Omar et al., 2015, Williams et al., 2015). Clinical audit style reviews were most common in analysis of secondary care incidents, with the limited attempts at high-level statistical modelling of NRLS hampered by inconsistent and poorly understood categorisation and missing data.

Categorisations within the NRLS remain problematic. Fowler (2013) described this, noting that *"...generic, web-based reporting systems (such as NRLS) are limited by inherent taxonomic limitations and the multi-factorial aetiology of harms...."* Many reports do not fit squarely within a single category used in NRLS, with the 'real signal' more adequately conveyed in free-text descriptions that may reference several categories. Reports often require manual reading by a reviewer to verify or reclassify. Several studies suggested 'potential harm' was being reported rather than 'actual harm,' with misclassification of incident type and clinical specialty surprisingly common. Despite a logical requirement for standardised severity (harm level), specialty, incident type and location, it must be acknowledged that incidents can fall into 'grey areas,' even for these clearer categories. For example: incidents in theatre may involve anaesthetics, surgery, nursing and issues of equipment failure, medication etc., or problems with syringe drivers may be classified as medical device or medication incidents but are, in reality, both. The contrast of primary subject and multiple subject coding used in this review served to highlight this problem. Incidents may genuinely represent several categories simultaneously, but the current classification scheme does not allow this, limiting its effectiveness for case ascertainment and learning. Although some elements of the NRLS' mappings have been designed to provide multiple routes to a particular classification, incident classifications would be better represented as a 'one-to-many' design, relational database parlance for allowing multiple categories for a single incident. This may improve accuracy of

review, and aid case ascertainment, but it will inevitably lead to an increase in complexity, and may increase the reporting burden on staff and organisations.

Analyses have mainly focussed on severe harm and death incidents, or 'harm' incidents with comparatively little use made of 'no harm' / near-miss reports. Most articles analysing 'no harm' incidents did so within a narrow range of a clinical specialty, or used a sample of no-harm reports to identify additional themes after the primary analysis of severe harm or death incidents. Some authors have suggested that high non-mandatory reporting rates represent a mature safety culture (Kelly and Jalil, 2011, Laker, 2009), but this cannot be substantiated using the NRLS alone. It is unclear if an organisation with a low non-mandatory reporting rate genuinely represents low incidence. Cultural and occupational factors may mean different members of staff may not perceive a given event to be an incident, and so may not report it, whereas another member of staff would. The same is true at organisational level. A low non-mandatory reporting rate may relate to missing data, as these grades of are not mandatory to report. This dynamic undermines the validity of analyses using non-mandatory reports, with the Francis enquiry into Mid-Staffordshire hospital (2013) suggesting it has no justification for the system *'...continuing to be voluntary, particularly as part of the system is now mandatory'*. Overall, the system remains agnostic as to whether a high reporting rate is a positive or negative indicator, and analyses that make either assumption will lead to unjustified conclusions.

As detailed in the introduction of this thesis, a major practical problem of the NRLS is the sheer volume of incidents reported, particularly at lower harm levels. As these reports are comparatively under-used, it poses a question about the purpose of collecting them.

A degree of naivety about the causes of incidents remains, as many reports (and articles in this review) were not sensitive to the underlying causes of incidents and future preventative measures. Clinical training and experience appear to be essential to most interpretations of incident reports, e.g. high harm from tracheostomy removal on wards compared to critical care was a potentially alarming finding, but given a clinical context to explain differences in removals and recordings, this difference was to be expected (Templeton et al., 2011).

Many of the themes raised in articles will be familiar to staff and patients. Poor communication between teams and with other services, poor recording or transmission of relevant information (e.g. current medication, allergies or DNACPR status), the availability and reliability of equipment, the implementation of care protocols and monitoring of patients, and recognition of deterioration all featured prominently. Suggestions within articles highlighted staffing, a lack of, or deficient, equipment and information systems, and staff

training/knowledge. Authors made a variety of suggestions including the use of surgical and anaesthetic checklists, greater involvement of other professionals in care processes, e.g. dietetics, and the stocking of 'airway kits' to avoid equipment shortage. Cousins et al.(2012) suggested the mapping of drug names in medication incidents requires changes to make it effective, but few other articles make practical suggestions about changes to NRLS after commenting on its deficiencies.

The NRLS is a vast and valuable resource for learning but cannot be treated as a representative central dataset of adverse events. Its part-mandatory nature, under-reporting, conceptual problems with identifying 'incidents,' poor classification, varying depth and quality of reports, and its large scale render it particularly challenging to analyse effectively at scale. Any techniques used to analyse NRLS should consider these points as a baseline.

2.4.1 Strengths and limitations

This review used a wide range of sources to target NRLS-based research, identifying appropriate methods and summarising learning. It has not examined output directly from the NRLS/NPSA and successor organisations. The high volume of output from these organisation would require a substantial review of its own, but could be considered the main use of NRLS (Panesar et al., 2009). The clinical review and count methods from these contexts were considered, but academic publications were favoured as they were, perhaps falsely, assumed to be more rigorous investigations. This assumption was due to my own bias in constructing this review, and this protocol is unable to assess the rigour of studies. Further systematic reviews with narrower questions and stronger inclusion/exclusion criteria would allow better assessment of quality. Rigour would come from clear descriptions of methods, clear reporting, acknowledgement of bias and limitations.

A wide variety of article types and styles were encountered, including journal papers, conference abstracts and academic reviews published on websites, not all of which were peer-reviewed. Although a broad framework for analysis was created, and subject-based tagging developed, the nuance of individual articles may not be adequately conveyed in this review format. Restricting this review to a particular style of article would have excluded large portions of relevant literature but enabled the use of common systematic review and 'quality assessment' tools.

Identifying relevant articles via search engines was challenging, as NRLS-based studies are inconsistently described and indexed. In some cases, articles did not adequately describe the use of NRLS in their title or abstract, but all manual screening stages were performed conservatively, retaining articles when uncertain. Despite a broad approach, it is possible the search strategy did not identify some relevant articles.

2.5 Lessons learned for development of models

The outputs of the literature review will frame the development of statistical models and treatment of the data prior to modelling in the coming chapters. Key lessons for the next stage include:

- **Classification of harm:** This is subjective in many regards and although it can be examined it cannot be assumed to be universally agreed or applied.
- **Non-mandatory incidents:** Death and severe harm incidents should be well represented in the data and may potentially act as population-level data. Non-mandatory incidents, whilst contributing to the largest proportions of data, should not be considered comprehensive or directly comparable between organisations.
- **Specialty or care setting:**
 - Specialty or care setting should be considered as overlapping and may not accurately reflect a given incidents. This may mean that modelling by speciality, even if relationships are identified, may not represent the incident in the 'real world.'
 - Better populated/studied specialties may provide benchmarks for model evaluation. For example, modelling critical care, anaesthetic or medication incidents would allow comparison of models and interpretations to be compared with published work.
- **Supporting data items:** Non-mandatory (and mandatory in some cases) data items may be poorly recorded or incorrectly submitted. Missing data should be considered when modelling and appropriate techniques, such as imputation assessed if appropriate.
- **Under-reporting:** The nature of incident reporting relies on incidents being observed, the incident reported, and that report being accurate. All of these stages are variable, and likely to depend on organisational culture and wider NHS policy. They should be considered in modelling as a known source of variance (e.g. as random-effect terms).

- **Free-text reporting:** Text descriptions offer greater depth of information but are challenging to analysis with 'traditional' quantitative statics and summaries. Modelling should either be restricted to categorical data, or apply text mining principles, such as those pioneered by Bentham (Bentham and Hand, 2009, Bentham and Hand, 2012, Bentham, 2010).
- **Types of organisations:** Many organisations submit data to NRLS, but submissions differ between setting and organisations. Models would be best targeted at specific settings with secondary care representing the richest source.
- **Organisational size:** An organisation seeing twice as many patients as another should be expected to have more incidents and must be scaled appropriately.

When combined, the lessons from this chapter suggest that directly modelling the NRLS data as if it was a representative national dataset, would be unwise. There may be much to be compared across organisations, but the data in its raw form lack a credible denominator for the size or risk of the organisation the data represent. Chapter 3 further examines the data empirically to understand the data completeness, distributions, and data quality features.

Chapter 3 Data descriptions, handling and summary

3.1 Introduction

National Reporting and Learning System (NRLS) data are the primary focus for analysis in this thesis. With over a million data rows per year, data handling and management require industrial-scale tools and systems, and cannot be contained or analysed in spreadsheets or text files. This chapter follows from the lessons learned in chapter 2 and is focussed on aims 1 and 2 of the thesis: examining the strengths and weaknesses of the NRLS data, and the data structures and preparation steps. It describes the process for accessing, receiving and handling NRLS data, summarises the data received, data completeness, distributions, and tests plausible associations between variables. The chapter concludes by assessing the extent to which statistical models can be built, the likely limitations, and a proposal for constructing a modelling dataset.

3.2 Description of data processing and data set

NRLS data can be submitted by two routes: patients or staff directly submitting reports to NRLS using webforms, or as extracts from NHS organisations' risk management software systems, but is almost exclusively the later. Risk management software, such as the commercially available Datix system, are designed for local reporting, examination, root cause analysis and learning. They also have many functions beyond the remit of NRLS to facilitate local investigation etc.

3.2.1 Data receipt, processing and characterisation

A data sharing agreement between NHSE/NHSI and University Hospital Birmingham NHS Foundation Trust (UHB), my employer and PhD sponsor, allowed UHB to hold and analyse NRLS data. Data are provided monthly by secure file transfer protocol (SFTP). All NRLS data are anonymised at source, prior to submission, with the NRLS team removing any identifiers accidentally included in free-text fields. Data were extracted, cleaned and loaded into a secured database environment at UHB using the process described below, with extracts received monthly

3.2.2 Requested data extract and format received.

An extraction, upload and cleaning process was developed to load data into MS SQL Server, for both my PhD work and for UHB's reuse (see Chapter 10). Data were received as 'Comma-delimited' (CSV) format, a text file format where commas are used to denote the breaks between data columns, with rows indicated by line breaks. In CSV format, text-columns are commonly surrounded by limiters, e.g. "my text", so that commas within the limiters do not split the data into additional columns.

The load process took several weeks to build and refine, using a 'dynamic SQL' process (Pollack, 2018). This process is controlled by a list of file names for the received CSV files and includes a mechanism to capture duplicated incident IDs (as incident data maybe updated and resubmitted to NRLS), isolate them in a duplicate table, and use only the most recent record for each unique incident identifier.

Issues identified and resolved during the build and load process were:

- Not all free-text fields were properly quoted (surrounded by " " characters), causing fields containing commas to split at inappropriate positions: e.g. a phrase such as "this is the first, and second" may split as (Figure 3.1):

Source Data field	Split 1	Split 2
"this is the first, and second"	this is the first, and second"	
this is the first, and second	this is the first	and second

Figure 3.1 Example of CSV file splitting

Text values that are properly quoted (surrounded by " " characters) will not cause extra splits (row 1) and entries without quotes will split at each comma regardless of whether the field is complete (row 2)

- Discussion with the NRLS team identified their use of SAS software for creating extracts, that did not always conform to the CSV standard described above. To resolve this, a SAS script was written that uploaded the CSVs with maximum string length settings, reformatted them to explicitly include quotes, and created output files as tab-delimited. This output format was chosen as 'tab' characters, that are illegal in NRLS free-text fields, would not cause incorrect delimiting.
- Additional blank data lines were attached to some extracts and not others. This can lead to null data rows or import errors if not properly inspected and corrected.
- 'Holes' in the data were identified and the NRLS confirmed that some of their extracts were based on incident date and some on the date they were received by NRLS. This

was subsequently adjusted so all records used the 'Received by NRLS date' and replacement extracts sent.

- Reports can be overwritten in the system, and it is common practice to do so once local organisations complete their investigations of incidents. Where this is observed, the most recent records were used. NRLS agreed in principle to issue an annual 'update' to reflect any changes, but these files have not been received.
- Four incorrectly split data lines were identified in November/December 2012. No logical reason for this could be identified and no data was missing. These records were manually corrected in UHB's data extracts.
- A number of columns contain either 'NULL' (empty) or the data field heading. These are likely to be interpreted as Boolean fields (yes or no, 1 or 0), akin to 'dummy coding' categorical variables for statistical modelling.

Data fields included in the extract are summarised in table 3.1. Most fields in NRLS are categorical and many of them are mapped to values by submitting organisations, with NRLS generating several processed fields. Many data fields are mapped to the NRLS's nomenclature/classifications, indicated in the 'Mapping Required' field of 3.1.

.

Field Name	Data Type	Manadatory Y/N	Requires Mapping Y/N	Minimum Length	Maximum Length	% Null or blank	Notes
RP01 Unique Incident ID	Integer	Y		7	8	0.00%	
RP05 Local Trust incident ID	Character	Y		1	43	0.45%	
RP02 Care Setting of Occurrence	Character	Y	Y	16	73	0.00%	
RP07 NHS Organisation Code	Character	Y		3	8	0.18%	Using National ODS codes
Date record exported to NRLS_Cleansed	Date			9	9	0.00%	NRLS added field after their 'cleaning'
IN01 Date of Incident	Date	N		9	9	0.00%	
IN03 Location (Ivl1)	Character	Y	Y	5	43	0.00%	
IN03 Location (Ivl2)	Character	Y	Y	5	59	2.11%	
IN03 Location (Ivl3)	Character	Y	Y	4	42	27.82%	
IN03 Location - Free Text	Character	Y	Y	1	197	94.96%	
IN05 Incident Category - Lvl1	Character	Y	Y	5	84	0.00%	
IN05 Incident Category - Lvl2	Character	Y	Y	5	81	11.23%	
IN05 Incident Category - Free Text	Character	Y	Y	1	483	75.82%	
IN06 Number of contributing factors	Categorical/Boolean	N	Y	1	1	92.36%	Multiple choice options
• Communication factors (includes verbal, written and non-verbal between individuals, teams, and or Organisations)				112	112	99.30%	
• Education and training factors (e g availability of training)				62	62	99.65%	
• Equipment and resources factors (e g clear machine displays, poor working order, size, placement, ease of use)				111	111	99.81%	
• Medication factors (where one or more drugs directly contributed to the incident)				81	81	99.65%	
• Organisation and strategic factors (e g abelingional structure, contractor agency use, culture)				100	100	99.65%	
• Other				5	5	99.47%	
• Patient factors (e g clinical condition, social physical psychological factors, relationships)				99	99	97.14%	
• Task factors (includes workguidelines procedures policies, availability of decision making aids)				101	101	99.27%	
• Team and social factors (includes role definitions, leadership, support, and cultural factors)				94	94	99.73%	
• Unknown				7	7	98.70%	
• Work and environment factors				22	151	97.16%	

Field Name	Data Type	Mandatory Y/N	Requires Mapping Y/N	Minimum Length	Maximum Length	% Null or blank	Notes
IN06 Contributing factors - Free text	Character	N		2	422	99.69%	Free text
IN07 Description of what happened	Character	Y		3	5291	0.00%	Free text description (supposed to be minimum 5 characters)
IN10 Actions Preventing Reoccurrence	Character	N		1	5435	46.61%	Free text
IN11 Apparent Causes	Character	N		1	7854	64.99%	Free text
PD01_A Age at time of Incident	Int	N		1	7	28.43%	Date of birth submitted but converted to age at incident
PD01_B Patient Age Range	Character	N		12	17	28.40%	Date of birth submitted but converted to age at incident
PD02 Patient Sex	Character	N		4	20	9.31%	
PD04 Adult Paediatrics' Specialty	Character	Y	Y	18	23	18.43%	Acute Only
PD05 Specialty - Lvl 1	Character	Y	Y	5	45	0.73%	
PD05 Specialty - Lvl 2	Character	Y	Y	3	69	14.85%	
PD05 Specialty - Free Text	Character	Y	Y	1	137	86.08%	
PD09 Degree of harm (severity) - display	Character	Y	Y	3	13	0.00%	Only Mandatory if patient was harmed
MD01 Med Process	Character	Y	Y	5	68	88.04%	Mandatory if IN05 coded Medicine
MD02 Med Error Category	Character	Y	Y	5	79	88.07%	Mandatory if IN05 coded Medicine
MD05 Approved Name (Drug 1)	Character	Y	Y	1	326	90.98%	Mandatory if IN05 coded Medicine
DE01 Type of Device	Character	Y	Y	5	56	96.98%	Mandatory if IN05 mapped to devices or IN06 to equipment
Date incident received by NPSA	Date			9	9	0.00%	
DV01 Patient Age at Time of Incident	Character			2	8	28.35%	Categorical/Coded
IN02_A_01 Hours				9	21	10.76%	

Table 3.1 Data field names and metadata in NRLS extracts

Each row represents a data field, with columns showing data types, whether items are mandatory, if data are 'mapped' to NRLS categories, variable lengths and the proportion of missing values are displayed, along with relevant notes.

3.3 Summary statistics

The NRLS extracts provided span the fiscal years 2010/11 – 2016/17, defined by the ‘Received by NRLS date’. A total of 55,502 duplicate records were removed from the data set with the most recent record retained for each unique incident ID, for a total of 10,964,514 records. A total of 426 distinct organisations (grouped into 10 organisational types) submitted reports (see Appendix C.1), mainly in secondary care settings. Tables 3.2 and 3.3 summarise incidents reported per fiscal year, grouped by the date reports were received by NRLS (3.2), and per year of incident occurrence (3.3). Both tables demonstrate the scale of reports received by NRLS, in the order of 1.5 million records per year. Both tables show a large increase from 2010 to 2011, stabilising to a year-on-year increase of $\approx 9\%$.

Fiscal Year Incident received by NRLS	Incidents	% of total	% change vs. previous year	
2010/11	1,063,162	9.70%		n/a
2011/12	1,349,119	12.30%	↑	27%
2012/13	1,452,719	13.25%	↑	8%
2013/14	1,605,556	14.64%	↑	11%
2014/15	1,722,134	15.71%	↑	7%
2015/16	1,841,165	16.79%	↑	7%
2016/17	1,930,656	17.61%	↑	5%
NULL ^a	3	0.00%		n/a
Total	10,964,514	100.00%		

^a Incident reports with missing 'Date incident received by NPSA' data field

Table 3.2 NRLS incidents reports per year reports received by NRLS

Incidents were selected from all data extracts provided by NHSI, between 2010/11 and 2016/17. Green arrows indicate an increase in incident numbers compared with the previous year.

Table 3.2 shows an increase in the total number of incident reports received per year for each year of the measuring period, with the latest percentage change between 2010/11 and 2011/12. Mandatory reporting of severe harm and death was instituted in 2011 and this may have been the catalyst for a general increase in incident reports (of all harm levels) being submitted to NRLS. Many organisations report data to NRLS via third-party software, such as Datix, that they use for local incident management. It may also represent organisations that had not previously used such a system, starting submission as these systems came online.

Fiscal Year of Incident occurrence	Incidents	% of total	% change vs. previous year	
<2010/11 ^b	939	0.01%		n/a
2010/11	1,250,249	11.40%		n/a
2011/12	1,366,042	12.46%	↑	9%
2012/13	1,469,558	13.40%	↑	8%
2013/14	1,602,708	14.62%	↑	9%
2014/15	1,766,163	16.11%	↑	10%
2015/16	1,844,282	16.82%	↑	4%
2016/17	1,664,554	15.18%	↓	-10%
>2016/17 ^c	19	0.00%		n/a
Total	10,964,514	100.00%		
^b Incident dates ranging from 1910 - 1989, likely to represent data entry error.				
^c Incident dates ranging from 2020 - 2049, likely to represent data entry error.				

Table 3.3 NRLS incidents reports per year of incident occurrence

Incidents were selected from all data extracts provided by NHSI, between 2010/11 and 2016/17. Green arrows indicate an increase in incident numbers compared with the previous year, and red arrows indicate a decrease.

Table 3.3 shows more stability than table 3.2 as it is does dependent on the information flow and timing with NRLS. It also highlights data quality issues where some reports received in the period 2010-11 – 2016/17 had invalid dates of incidents (0.01%). This questions data validation rules used by NRLS, as these reports do not appear to have been rejected or corrected. The drop observed for 2016/17 suggests there may be missing reports, that were likely to be reported in 2017/18, with incident dates in 2016/17. This is a facet of using a cut-off, and the consequences of the practicalities of receiving incidents in batches from many trusts.

Severity of harm is an important classification within NRLS, detailing the consequences for patients, but also whether incidents are taken for national review. Proportions in the different harm categories have remained fairly constant from year to year, with death and severe harm incidents accounting for 0.25% and 0.47% respectively. This can be seen in the context of the increasing number of reports, in that severe harm and death reports have remained in similar proportions but absolute numbers have doubled when comparing 2010/11 to 2016/17 (table 3.4). This may suggest that the incidence of severe harm or death incidents is increasing, but may also be due to increasing healthcare activity, increased awareness of the incidents, or better reporting practices.

Fiscal Year incident received by NRLS	Severity of Harm						Patient Group*	Grand Total
	No Harm	Low	Moderate	Severe	Death	NULL		
2010/11								
Incidents	734,497	253,778	65,826	6,380	2,648	21	12	1,063,162
% (row)	69.09%	23.87%	6.19%	0.60%	0.25%	0.00%	0.00%	100.00%
2011/12								
Incidents	926,619	323,792	87,545	8,074	3,073	4	12	1,349,119
% (row)	68.68%	24.00%	6.49%	0.60%	0.23%	0.00%	0.00%	100.00%
2012/13								
Incidents	985,118	362,122	94,045	7,724	3,693	2	15	1,452,719
% (row)	67.81%	24.93%	6.47%	0.53%	0.25%	0.00%	0.00%	100.00%
2013/14								
Incidents	1,099,052	400,574	94,681	7,170	4,070		9	1,605,556
% (row)	68.45%	24.95%	5.90%	0.45%	0.25%	0.00%	0.00%	100.00%
2014/15								
Incidents	1,212,129	411,306	87,204	7,550	3,915	2	28	1,722,134
% (row)	70.39%	23.88%	5.06%	0.44%	0.23%	0.00%	0.00%	100.00%
2015/16								
Incidents	1,318,977	435,789	74,122	7,401	4,854	-	22	1,841,165
% (row)	71.64%	23.67%	4.03%	0.40%	0.26%	0.00%	0.00%	100.00%
2016/17								
Incidents	1,401,598	449,588	66,689	7,247	5,500	-	34	1,930,656
% (row)	72.60%	23.29%	3.45%	0.38%	0.28%	0.00%	0.00%	100.00%
NULL								
Incidents	-	-	-	-	-	3	-	3
% (row)	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
Total Incidents	7,677,990	2,636,949	570,112	51,546	27,753	32	132	10,964,514
Total % (row)	70.03%	24.05%	5.20%	0.47%	0.25%	0.00%	0.00%	100.00%

Table 3.4 Incident reports, per year and harm classification

Numbers of report per harm classification level, by fiscal year reports were received by NRLS. Columns represent harm levels, and rows represent incident counts and percentages, grouped in fiscal years (grey). *The free-text phrase 'Patient Group' was included in this field in a number of reports, presumably as a data submission error.

'Care setting' detailed the location based on the type of service an organisation provides (table 3.5). The majority of incidents were reported in Acute/General Hospital settings at 72.6%. Other major care settings were in mental health services at 12.7% and community nursing, medical and therapy service (including community hospitals) at 11.4%. General Practice (GPs) represent a particularly small number of reports in relation to their high levels of activity. This may be confounded by the nature of general practice, with short consultations that do not necessarily require interventions, and the capacity to observed and report incidents in this setting. The awareness of NRLS in general practice is also an unknown quantity.

RP02 Care Setting of Occurance	Incidents	% of Total
Acute / general hospital	7,960,518	72.603%
Ambulance service	60,290	0.550%
Community and general dental service	5,543	0.051%
Community nursing, medical and therapy service (incl. community hospital)	1,246,537	11.369%
Community optometry / optician service	234	0.002%
Community pharmacy	92,161	0.841%
General practice	46,436	0.424%
Learning disabilities service	164,585	1.501%
Mental health service	1,388,210	12.661%
Total	10,964,514	100.000%

Table 3.5 Care settings of incident reports

NRLS reported incidents between 2010/11 – 2016/17, showing the number of incident reports received by NRLS and the percentage of total incident reports they represent.

‘Location’ is a mapped field describing which area of an organisation an incident occurred in. Location is hierarchical, with three levels of detail nested within each other, such that each level 2 category is nested within a more broadly defined level 1 category. Levels 1 and 2 are summarised in table 3.6 and levels 1 – 3 in Appendix C.2. Inpatient areas of acute hospitals represented over half the reported locations at 51.53% and 9.87% in inpatient areas of Mental Health units. Other acute hospital areas representing the largest proportions of incident reports included accident/minor injuries assessment units, support areas and outpatient departments. Surprising areas included 4.47% of incidents with a location of ‘Private house/flat, 0.19% submitted as ‘Unknown’ rather than null, 1.38% describe as ‘Other’ at Level 1.

Location Levels (1/2)	Total number incidents reported	Incidents as a percentage of total
Ambulance (including call / control centre)	27,706	0.25%
Call / control centre	8,550	0.08%
In vehicle / in transit	13,061	0.12%
NHS Direct	13	0.00%
NULL	15	0.00%
Other	6,067	0.06%
Community hospital	572,122	5.22%
Day care services	9,497	0.09%
General areas	72,975	0.67%
Inpatient areas	391,753	3.57%
NULL	507	0.00%
Other	43,416	0.40%
Outpatient department	26,054	0.24%
Support Services	27,920	0.25%
General / acute hospital	7,671,154	69.96%
Accident (A) / minor injury unit / medical assessment unit	610,631	5.57%
Ambulatory care treatment centre	14,548	0.13%
Day care pre-assessment clinic	145	0.00%
Day care services	122,102	1.11%
General areas	286,145	2.61%
Inpatient areas	5,650,166	51.53%
NULL	496	0.00%
Other	83,928	0.77%
Outpatient department	446,565	4.07%
Outpatient pre-assessment clinic	121	0.00%
Support Services	456,307	4.16%
Mental health unit / facility	1,389,480	12.67%
Community mental health facility	136,662	1.25%
Day care services	17,386	0.16%
General areas	105,854	0.97%
Inpatient areas	1,082,612	9.87%
NULL	1,269	0.01%
Other	22,871	0.21%
Outpatient department	10,263	0.09%
Support Services	12,563	0.11%
Not applicable	18,330	0.17%
NULL	4	0.00%
Other	151,474	1.38%
NULL	151,472	1.38%
Other	2	0.00%
<i>continued on next page</i>		

Location Levels (1/2)	Total number incidents reported	Incidents as a percentage of total
Primary care setting	339,863	3.10%
Ambulatory care treatment centre	2,164	0.02%
Community pharmacy	84,891	0.77%
Dental surgery	20,951	0.19%
GP Surgery	37,136	0.34%
Health centre / out-of-hours centre	125,370	1.14%
NHS Direct	22,736	0.21%
NULL	29	0.00%
Optician / optometrist	3,650	0.03%
Other	34,232	0.31%
Rehabilitation centre	8,704	0.08%
Public place (specify)	37,035	0.34%
Residence / home	639,884	5.84%
Hospice	7,962	0.07%
Intermediate care setting	27,570	0.25%
NULL	290	0.00%
Nursing home	62,813	0.57%
Other	15,481	0.14%
Prison / remand centre	35,914	0.33%
Private house / flat etc.	489,854	4.47%
Social care facility	96,141	0.88%
Day care services	1,907	0.02%
Local Authority (non-residential)	558	0.01%
NULL	126	0.00%
Other	7,152	0.07%
Residential care home	86,398	0.79%
Unknown	21,321	0.19%
Grand Total	10,964,514	100.00%

Table 3.6 Incident reports by location levels 1 and 2

NRLS incident reports received between 2010/11 – 2016/17. Columns represent counts of incidents and percentage of total incidents, with rows representing locations at level 2 (white), grouped by at level 1 (grey).

Specialty is also nested within levels and described in Table 3.7 at level 1, with full list of level 2 specialties in Appendix C.3. Medical specialties accounted for the highest number of incident reports (30.73%), followed by surgery (14.59%) and mental health (12.59%). Articles in the literature review (Panesar et al., 2013a) suggested Trauma and Orthopaedic services to be the highest submitter which is not evident in the data presented, with General Medicine, Obstetrics, Care of older people (all in acute settings), adult mental health and community nursing representing higher proportions of incidents, although trauma and orthopaedics is the highest surgical specialty, which is what their paper is likely referring to.

Specialty (Level 1)	Incidents	% of total
Accident and Emergency (A)	672,352	6.13%
Anaesthesia Pain Management and Critical Care	149,591	1.36%
Children's Specialties	12,026	0.11%
Dentistry - General and Community	12,695	0.12%
Diagnostic services	400,177	3.65%
Learning disabilities	226,858	2.07%
Medical specialties	3,369,003	30.73%
Mental health	1,380,566	12.59%
Not applicable	106,360	0.97%
NULL	79,621	0.73%
Obstetrics and gynaecology	998,749	9.11%
Other	647,546	5.91%
Other specialties	296,734	2.71%
Primary care / Community	891,989	8.14%
PTS (Patient Transport Service)	38,018	0.35%
Surgical specialties	1,599,356	14.59%
Unknown	82,873	0.76%
Grand Total	10,964,514	100.00%

Table 3.7 Incident reports by level 1 specialty groups

Level 1 specialty descriptions for incidents reported to NRLS in 2010/11 – 2016/17. Rows represent specialty categories and columns represent incidents and percentages of total.

Incident reports are also specified with a time of day data field. This is reported in the NRLS extract as an hour of the day (table 3.8 and figure 3.2). The largest category was NULL, with 10.76% of incidents. There is also a major spike around midnight. This spike is seen in isolation, and not continued in hours 23 or 1, and may suggest data quality/default values of midnight submitted to/from incident reporting systems. Higher proportions of incidents are recorded during the 'working day', between 9 and 5. This may reflect more staff (such as ward clerks) to report incidents, or it may reflect incidents occurring related to daily routines, as suggested in Chapter 2 (Healey et al., 2008).

Time of Incident in Hours (0 = 12 a.m.)	Incidents	% of total
0	594,714	5.42%
1	196,338	1.79%
2	187,056	1.71%
3	171,751	1.57%
4	167,200	1.52%
5	163,825	1.49%
6	184,194	1.68%
7	233,598	2.13%
8	415,336	3.79%
9	617,661	5.63%
10	716,245	6.53%
11	699,472	6.38%
12	608,595	5.55%
13	491,636	4.48%
14	604,812	5.52%
15	581,781	5.31%
16	564,238	5.15%
17	436,626	3.98%
18	447,513	4.08%
19	410,837	3.75%
20	403,609	3.68%
21	309,086	2.82%
22	315,356	2.88%
23	263,465	2.40%
NULL	1,179,570	10.76%
Grand Total	10,964,514	100.00%

Table 3.8 Time of day of NRLS incident reports

NRLS incidents reported 2010/11 – 2016/17, with zero representing midnight.

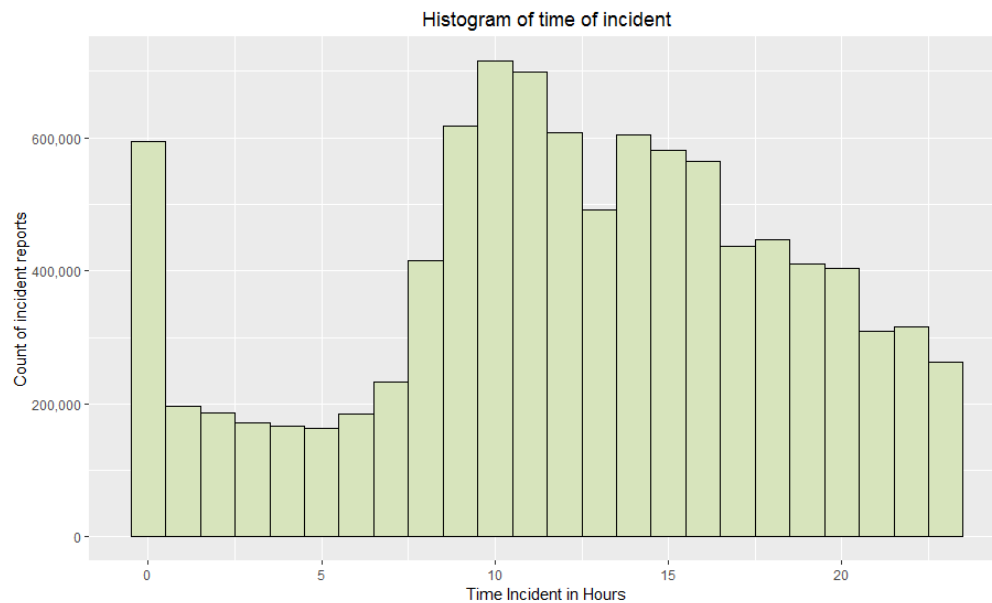


Figure 3.2 Time of day of NRLS incident reports

NRLS incidents reported 2010/11 – 2016/17, excluding NULL values (10.76%), bins represent single hours with zero representing midnight.

Incident Type is a 2-level category, presented at the highest level in table 3.9, with full table in Appendix C.4. Patient accidents represented the highest group of incidents (29.5%), twice the number of the nearest types, representing nearly a quarter of reports. Implementation of care and medication were also major groups, with pressure ulcers representing very few incidents (a point contradicted by findings in Chapter 9), likely due to other reporting routes.

Incident type (Level 1)	Incidents	% of total
Access, admission, transfer, discharge (including missing patient)	990,024	9.029%
Clinical assessment (including diagnosis, scans, tests, assessments)	543,290	4.955%
Consent, communication, confidentiality	400,101	3.649%
Disruptive, aggressive behaviour (includes patient-to-patient)	348,538	3.179%
Documentation (including electronic & paper records, identification and drug charts)	690,025	6.293%
Implementation of care and ongoing monitoring / review	1,257,205	11.466%
Infection Control Incident	194,714	1.776%
Infrastructure (including staffing, facilities, environment)	621,828	5.671%
Medical device / equipment	307,573	2.805%
Medication	1,230,715	11.225%
NULL	1	0.000%
Other	426,007	3.885%
Patient abuse (by staff / third party)	57,916	0.528%
Patient accident	2,357,183	21.498%
Pressure Ulcer	24	0.000%
Self-harming behaviour	406,280	3.705%
Treatment, procedure	1,133,090	10.334%
Grand Total	10,964,514	100.00%

Table 3.9 Incident reports by level 1 incident types

Level 1 incident type descriptions for incidents reported to NRLS in 2010/11 – 2016/17. Rows represent incident type categories and columns represent incidents and percentages of total.

Age range	Incidents	% of total
Under 28 days	78,287	0.71%
1 month to 1 year	126,678	1.16%
2 to 4 years	97,284	0.89%
5 to 11 years	108,895	0.99%
12 to 17 years	202,392	1.85%
18 to 25 years	450,340	4.11%
26 to 35 years	735,305	6.71%
36 to 45 years	555,031	5.06%
46 to 55 years	589,003	5.37%
56 to 65 years	710,797	6.48%
66 to 75 years	1,121,091	10.22%
76 to 85 years	1,687,194	15.39%
Over 85 years	1,388,455	12.66%
NULL	3,113,762	28.40%
Grand Total	10,964,514	100.00%

Table 3.10 Incident reports by subject age groups

Age groups for subject of incident reports submitted to NRLS in 2010/11 – 2016/17. Rows represent age categories and columns represent incidents and percentages of total.

Patient demographics are reported in tables 3.10 and 3.11. Large numbers of missing values are represented by 'NULL,' 28.4% for age and 9.31% for sex. Median age was 68 with an interquartile range of 44 (39-83), and the distribution shown in figure 3.3. Percentages of incidents were high in the under-4s and over-60s, suggesting age may be an important risk predictor. This should be considered in the context of higher healthcare use by the elderly (41.15% hospital episodes for patients 65 years and older in 2014/15 (The Health Social Care Information Centre, 2015)), allowing more opportunity for incidents. Another peak from mid-twenties to thirties may represent maternity episodes when considered in the context of hospital admissions.

Patient Sex	Incidents	% of total
Female	3,893,798	35.51%
Indeterminate	6,982	0.06%
Male	3,193,764	29.13%
Not stated / unknown	2,849,361	25.99%
NULL	1,020,471	9.31%
Patient Group*	138	0.00%
Grand Total	10,964,514	100.00%

Table 3.11 Incident reports by Patient Sex groups

Sex of subjects of incident reports submitted to NRLS in 2010/11 – 2016/17. Rows represent sex categories and columns represent incidents and percentages of total.

*The free-text phrase 'Patient Group' was included in this field in a number of reports, presumably as a data submission error



Figure 3.3 Histogram of subject ages in incident reports

Bins represent age groups of two years, and red line represents the median age (68), (28% missing data)

Sex included the option 'Not stated/unknown' which has been entered by reporters (table 3.11). This can be considered a 'known unknown,' as distinct from NULL that should be considered not just as missing, but also as an 'unknown unknown'. A higher proportion of incidents were reported for female patients, but 'not stated' and null values, when combined, were larger.

Literature review and descriptive study of NRLS suggested that the majority of reports stem from acute hospital settings. Comparisons within this setting may be more reasonable than comparing across all settings, as incidents in ambulance trusts, mental health trusts, general practice etc. are likely to differ. For the purposes of modelling in this thesis, analysis will be restricted to incidents where the field [RP02 Care Setting of Occurrence] was 'Acute / general hospital,' representing 72.6% of incidents in the dataset (table 3.12).

Harm Level	FiscalYear									Grand Total
	<2010/11	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	>2016/17	
No Harm										
Incidents	674	867,184	932,182	998,835	1,098,815	1,249,228	1,325,930	1,205,136	6	7,677,990
% acute	83.68%	73.42%	74.94%	76.76%	77.17%	78.09%	77.13%	76.38%	100.00%	76.44%
Low										
Incidents	184	293,532	333,836	364,566	398,695	419,348	433,588	393,187	13	2,636,949
% acute	67.39%	66.01%	64.94%	65.10%	64.52%	64.81%	64.54%	64.32%	69.23%	64.84%
Moderate										
Incidents	56	78,452	88,556	94,729	93,914	85,932	72,617	55,856		570,112
% acute	80.36%	66.12%	64.38%	59.13%	57.74%	54.67%	53.94%	52.67%		58.71%
Severe										
Incidents	17	7,782	8,234	7,629	7,233	7,598	7,249	5,804		51,546
% acute	76.47%	72.67%	70.20%	68.79%	69.82%	66.40%	65.65%	66.44%		68.69%
Death										
Incidents	8	3,265	3,216	3,778	4,043	4,029	4,869	4,545		27,753
% acute	12.50%	51.24%	46.86%	43.54%	38.24%	40.46%	37.34%	33.49%		40.87%
Patient Group**										
Incidents		12	12	19	8	26	29	26		132
% acute		33.33%	16.67%	5.26%	12.50%	7.69%	6.90%	0.00%		9.09%
NULL										
Incidents		22	6	2		2				32
% acute		81.82%	0.00%	0.00%		0.00%				56.25%
Total incidents	939	1,250,249	1,366,042	1,469,558	1,602,708	1,766,163	1,844,282	1,664,554	19	10,964,514
Total % acute	79.55%	71.16%	71.71%	72.60%	72.75%	73.66%	73.10%	72.59%	78.95%	72.60%

Table 3.12 Proportions of Acute hospital incidents per year

Proportions of incidents reported in the 'Acute hospital settings' group, by year of incident occurrence and degree of harm, for incident reported to NRLS between in 2011/12 – 2016/17. Columns at either end of the data range, (<2010/11 and >2016/17), are outside the scope of the dataset and may represent data submission errors.

**The free-text phrase 'Patient Group' was included in this field in a number of reports, presumably as a data submission error

Table 3.12 suggests that proportions of incidents attributable to acute trusts remained stable, but the number of reports increased dramatically between 2010/11 and 2015/16, by over 500 thousand reports per year. The reduction for 2016/17 is likely to represent incidents occurring in 2016/17 that were reported to NRLS 2017/18.

3.4 Selected relationships between data items

Several data items were investigated in this summary section to confirm relationships. Many of the variables in NRLS are not suitable for comparison with each other and represent nested levels within a hierarchy, identifiers or free-text descriptions. Level of harm is the main outcome indicator in the NRLS, with other variables describing the situations of incidents.

Tests were conducted using Chi-squared tests for categorical predictors, with p-values calculated using a Monte Carlo test (Adery, 1968). Significant associations with predicted harm level (with p-value < 0.001) were observed for:

- Incident Category (Levels 1-2)
- Weekday of Incident
- Hours of Incident
- Care Setting
- Location (Levels 1 – 3)
- Specialty (Levels 1 – 2)
- Medical Error Process
- Medical Error Category

These results, however, are unsurprising given the huge variety of incident types and how rare some of the higher harm levels are. E.g. it would be surprising for GP incidents to have the same proportions of harm as acute hospital settings. They do not appear to lend additional value to the data, beyond the descriptive statistics in section 3.3.

3.5 Conclusions

Rigorous data handling practices are required for processing NRLS extracts to adequately load, process and handle such large datasets. Manual inspection can only be used where specific records require checking, so a system of automated check and de-duplication is an appropriate way to manage this. Attention must be paid to the formatting of data, and any instances where formatting is 'broken,' should be identified and resolved before analysis.

Numbers of incident reports have increased year-on-year, with the proportion of death or severe harm incidents remaining stable. Reporting is mandatory for incidents of this type, and this leads to a question of whether death or severe harm incidents are increasing, or if another mechanism explains this. Incident reports commonly feature people aged over 60, neonates or young children, and appear to be more common in women of child-bearing age. Different types of organisation, with differing hierarchies of location and clinical specialty, report different kinds of incidents and differing levels of harm. Medical and surgical inpatients make up significant proportions of the data, with patient accidents the most common incident type.

The largest ‘target’ setting for statistical modelling is the acute hospital setting, and models in this thesis will focus here. Missing data and classification issues explained in Chapter 2 suggest that NRLS is not a good source of casemix data. NRLS contains information on reported incidents. Chapter 2 suggested that reporting systems only contain a fraction of the real incidence of incidents, but we also know nothing about the ‘exposure’ or the risk of incident. On a patient level, some patients will be at more risk of adverse events than others if, for example, they have trouble mobilising or receive complex healthcare interventions. This will also scale to organisational level, where risk will differ with the casemix. Chapter 5 puts this in a conceptual framework and introduces a secondary source of casemix data to counter this limitation in the NRLS. A dataset focussed on hospital activity is required to examine exposure in such a way. The Hospital Episode Statistics (HES) (NHS Digital, 2012) is one such source that will be examined in the coming chapters. HES and NRLS cannot be directly linked (see Chapter 5), but can be modelled at similar levels of aggregation, creating a count dataset. Chapter 4 details common methods and theory associated with modelling count data, and Chapter 5 applies these methods to NRLS-HES and explains the construction of the dataset. Chapters 6 – 8 extend these count modelling approaches and examine how they can fit with current NHS regulatory frameworks.

Chapter 4 Methodological considerations for the analysis of count data

Chapter 3 examined the NRLS at records level, suggesting that a secondary source of casemix data is used and models developed using identically aggregated count data, that will be examined in Chapter 5. This chapter contributes to aim 3 of the thesis by examining some of the theoretical considerations when using count data, appropriate modelling techniques and methods for examining model output.

Rabe-Hesketh & Skondral (2012) suggest that *“Counts can be thought of as aggregated versions or summaries of more detailed data on the occurrences of some event.”* Count data commonly arise from two possible sources, observation of point processes, or the discretization of continuous outcomes (Cameron and Trivedi, 2013e). Count data have several properties that distinguish them from continuous numeric data:

- They are natural, whole numbers (integers), sometimes referred to as ‘discrete.’ E.g. counts of 2 or 3 are possible, but 2.5 is not.
- They range from zero to infinity, with counts less than zero impossible
- Counts occur in a fixed time period, with a known average rate (μ) - counts that occur with a variable time period can be characterised as rates using a time denominator
- Since they are bounded at zero, and whole numbers, they are not normally distributed. Counts with low average rates are noticeable skewed but may appear asymptotically normally distributed when the average rate is high

Regression models, based on the normal distribution, may be a poor fit for count data although data are sometimes log-transformed and analysed in such a way. The use of a linear model on transformed data is a poor reflection of the distribution of the error term and may produce predictions outside the possible range $[0, +\infty]$.

4.1 The Poisson distribution and Poisson regression

Count data are often better modelled using Poisson regression, based on the Poisson distribution (Poisson, 1837, Cameron and Trivedi, 2013a), regarded as a limited case of the

binomial distribution. If random variable Y is Poisson distributed, the probability of observing a count of y is:

$$\Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!},$$

where μ is the expected average count or rate, e is Euler's number (the base of the natural logarithm: ~ 2.71828), and $y!$ is the factorial of y .

Poisson regression can be considered part of the wider family of Generalized Linear Models (GLM) propose by Nelder and Wedderburn (1972), extending linear modelling principles to other distributions from the exponential family. Data are estimated, using a 'link function' from the exponential family, to facilitate a linear model on the scale of this link function. This differs from the log-transformed linear model mentioned above in terms of estimation methods and assumptions about distributions of error terms. Error structures in GLMs are assumed to be distributed according to a distribution from the exponential family. This includes the Poisson distribution with the natural logarithm link function, but can be extended to other data types, such as binary data with the binomial distribution and a logistic link function, or waiting times data with the gamma distribution and a reciprocal link function (although a log link function may also be used).

A GLM has the basic structure:

$$g(\mu_i) = \mathbf{X}_i \boldsymbol{\beta},$$

where μ_i is the expectation of the random variable Y for the i^{th} row of a model matrix \mathbf{X} , g is a link function from the exponential family of distributions, and $\boldsymbol{\beta}$ is a vector of unknown parameters (model coefficients), estimated during the modelling process (Wood, 2017b).

In the case of Poisson regression, μ_i represents the expected count (or rate), with a Poisson response distribution/error structure, and a linear predictor of the form:

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} \dots + \beta_p X_{pi}$$

Where β_0 is an intercept term, β_1 the model coefficient of predictor variable X_1 for each line of model matrix i . This extends to p predictor variables. We then identify the most likely model coefficients during estimation.

To estimate models, we must define a loss function: a function that penalizes prediction error. There are various loss functions, but the most common is ‘squared error loss,’ where we seek to minimise the difference between Y and function of X : $L(Y, f(X)) = (Y - f(X))^2$ (Hastie et al., 2009a). In the context of regression, this loss is calculated from the sum of the pairwise differences between observed data points and predicted data points. These differences are referred to as the ‘residual’ error. We implement this in linear regression by calculating the sum of the squared residuals and solve our model by finding its minimum. This process commonly referred to ‘ordinary least-squares’ (OLS) and can be solved directly.

OLS is suitable for the estimations of GLMs only when we assume normality. Under other distributional assumptions, such as Poisson, it is usually performed by Maximum Likelihood estimation (MLE) (Wilks, 1938, Nelder and Wedderburn, 1972). MLE and OLS are equivalent for normally distributed data, but MLE cannot be solved in the same way as OLS, and must be iteratively estimated instead (Wood, 2017b). A probability density (or mass) function is defined as $f_l(y)$ for random variable, of which y is an observation, with l representing the unknown parameters of a model. Values of l that make $f_l(y)$ larger for a given y are more likely to be ‘correct’ than values that make it smaller. The natural log-transformed likelihood function, $\log(f_l(y))$ is used to judge the best fit, and the maximum log-likelihood value yields l : the ‘most likely’ parameter estimates.

In Poisson regression, where μ is a rate but occurs over a fixed interval for all observations, it can be regarded as a count. When intervals vary between observations, e.g. over different time intervals, or as proportions within different sized units, μ must be regarded as a rate. This can be specified in a model by using an ‘offset’ term. An offset can be regarded as a scale factor, achieved by fixing the model coefficient at unity (Cameron and Trivedi, 2013c), with the model estimated by constrained maximum likelihood. The model would then predict a rate of the outcome per unit of the offset. Offsets in Poisson models are usually transformed to correspond to the link function and avoid mismatches in scale.

4.2 Error structure and overdispersion

The Poisson distribution has a single parameter, μ , that is both the mean and the variance of the distribution. This is key assumption of a Poisson regression model, so the conditional

mean of a model is expected to equal the conditional variance, referred to as 'equidispersion.' Equidispersion is commonly violated with 'real-world' count data (Breslow, 1984).

Overdispersion, where the conditional variance is greater than the conditional mean, can occur for a variety of reasons (McCullagh and Nelder, 1983, Collet, 1952, Ver Hoef and Boveng, 2007, Cameron and Trivedi, 2013b), including:

- **Aggregation / Discretization:** Summarising or aggregating data loses some of the nuance of the variation within the data. In general, this reduces resolution and efficiency, and can result in overdispersion (Dean and Balshaw, 1997).
- **Mis-specified systematic component of a model:** Model parameters may not adequately describe the variation in the model. A better specification may include more, or different, predictors and relevant interactions that may reduce overdispersion. Parameterisation of the model (the way predictor variables are represented) may also affect this, as information may be lost if a parametrisation does not reflect the underlying relationships in the data e.g. using crude grouping of a continuous variable that may be better modelled as a continuous numeric predictor.
- **Presence of outliers:** Extreme values, not representative of the distribution, may distort models. Identification of these observations, and the reasons for extreme values, can inform decisions to omit observations or restrict the modelling range.
- **Variation between response probabilities (heterogeneity):** This may arise when measurements with identical predictors do not always produce the same response. It may relate to under-specification of the model, but may also be due to correlation within the data when each point is not truly independent. E.g. repeated measurements from an individual, or clustered structures within the data, such as centres within a clinical trial.

In the presence of overdispersion, the Poisson modelling process will underestimate the variance in the model, giving standard errors that are too small. With under-estimated errors, the significance of parameters (judged by t-tests, chi-squared tests, or used to construct confidence intervals) will be over-stated (Breslow, 1984).

Estimation of the error structure and significance of coefficients is based on the MLE, and can take a number of forms (Cameron and Trivedi, 2013d). The most common approach is to

calculate the standard errors of a Poisson model and use then to form 'Wald' confidence intervals (Wald, 1942). The Wald test is analogous to a t-test in linear regression, and its square is asymptotically chi-squared distributed on one degree of freedom (Hauck and Donner, 1977). The major flaws in this technique are the assumptions that the error distribution is entirely known, the standard error is correctly calculated, and is asymptotically normally distributed which may not be the case in GLMs. Wald confidence intervals are the default output of many statistical packages.

4.2.1 Bootstrapping & likelihood profiles

Common techniques to increase the accuracy of confidence intervals, particularly around random-effects (see section 4.2.4), tend to be more computationally intensive than calculating Wald intervals. These techniques include:

- **Profiled likelihood:** Profiling can be used for confidence intervals by maximising the likelihood of the joint distribution of the standard error and the mean expected count, for a fixed mean (Cameron and Trivedi, 2013d). Quantiles of the likelihood function can, after transformation, be applied to the quantiles of the normal distribution to derive confidence intervals (Bates et al., 2015, McGrath et al., 2013). This does not have to be normally distributed, and is accurate, but can take significant time to calculate on complicated models. They may also be affected by overdispersion.
- **Bootstrapped intervals.** The 'Bootstrap' (Efron, 1979) is a statistical inference approach based on resampling, with replacement, to build up a sampling distribution (Fox and Weisberg, 2012). Statistics can then be calculated on this sampling distribution. These calculations are repeated many times, once on each sub-sample. These estimates are normally distributed (if using a parametric bootstrap) and allow the distribution of these estimates to be compared with the original. The bootstrapped estimates of parameters, and sampling variance allow standard deviations to be calculated on the standard error estimates and substituted for the original terms (Cameron and Trivedi, 2013d). Bootstrapping also takes a substantial amount of time compared to other techniques, as the process is resampling and fitting the model repeatedly. It may also account for the effects of overdispersion (standard errors are too small), as the resampling gives a better estimate of the residual variance.

4.2.2 Scaled deviance models

Where overdispersion is suspected, a common next step is to formally test for it. A chi-squared test of the sum of the squared Pearson residuals, on the residual degrees of freedom (Bolker, 2018), can be used for this. The null hypothesis for the test is that the model is not overdispersed. A dispersion ratio, ϕ , can also be presented from this test where a value of 1 represents equidispersion, and a ratio higher than one signifies overdispersion.

A simple adjustment for overdispersion commonly uses a multiplicative scale factor to inflate the variance. Scaling in this manner does not alter parameter estimates, only the estimated error. The techniques presented below take this approach but an alternative, additive, overdispersion model used with overdispersed healthcare indicators (Spiegelhalter et al., 2012a, Spiegelhalter, 2005b) is discussed in Chapter 8. This approach is related to the random-intercept models described later in this chapter.

Multiplicative scaling options (summarised in Table 4.1) include::

- **‘Robust’ confidence intervals** - White-Huber (Huber, 1967, White, 1982) estimators (‘sandwich’ estimators) can be used to adjust for violations due to heteroscedasticity and estimate standard errors without assuming the full distribution is known. This method is often recommended over Wald-style tests and intervals.
- **Quasi-likelihood** (Wedderburn, 1974) models allow for the estimation of a scale factor, rather than fixing it at 1 in Poisson models. Quasi-likelihood models assume the variance is a multiple of the mean, i.e. $\text{variance} = \text{mean} * \text{scale}$, (Cameron and Trivedi, 2013d), allowing the scale parameter to be reported in output. They also relax the usual GLM distributional assumptions from a fully specified exponential family distribution to a simpler mean-variance relationship. This may lack a full distributional form and corresponding MLE, making comparisons between models challenging using common measure metrics such as the likelihood ratio test or AIC (require an MLE). Poisson quasi-likelihood models will be referred to as *‘quasipoisson’* in this thesis.

4.2.3 Mixture/compound distribution models

In some cases, a compound distribution can be used, combining assumptions about the distribution of counts and their generating processes. The term ‘mixture models’ refers to compound distributions such as Poisson-Gamma in this thesis. The term ‘mixture’ is also used to refer to distributions where sub-populations are contained within a larger population, but

this is not examined in this thesis. In the case of sub-populations, such as two normal distributions within a larger distribution (i.e. bimodal), data may be better described by finite mixture models.

Although these models could be described as scaled in terms of the error estimation, the parameter estimates are conditional on the variance and scale factors, so parameters estimates are not the same as those obtained from Poisson models.

Mixture models commonly used for count data analysis include (see Table 1):

- **Negative Binomial (NB):** Whilst Poisson models assume variance = means, and quasi-poisson assume variance = mean * scale parameter, NB models are mixtures of the Poisson distribution for inter-cluster variation and a gamma distribution for intra-cluster variation. Therefore, NB models represent Poisson means, following a gamma distribution (Sellers and Shmueli, 2010a). There are two standard parameterisations of NB models, often referred to as **NB1** (constant dispersion) and **NB2** (mean dispersion) (Cameron and Trivedi, 1986), with both referring to subject i within cluster j :

- **NB1** models assume a group-specific expected count μ_{ij} , with the variance:

$$Var(y_{ij}|x_{ij}) = \mu_{ij}(1 + \alpha),$$
where $1+\alpha$ is a multiplicative overdispersion/scale factor. NB1 therefore takes a similar form to the quasipoisson model (Rabe-Hesketh and Skrondal, 2012), but with a full distributional form and corresponding MLE.
- **NB2** models can be framed as random-intercept models (see below), but rather than assuming a normally distributed random-effect, we assume a gamma distributed frailty with mean of one and variance α . This then has scale parameter α and shape parameter $1/\alpha$ giving the quadratic form:

$$Var(y_{ij}|x_{ij}) = \mu_{ij} + \alpha\mu_{ij}^2$$
(Rabe-Hesketh and Skrondal, 2012)

NB1 models assume the variance is scaled in the same manner as quasipoisson models, but NB2 models estimate the variance as quadratic to the mean. When referring to negative binomial models, most articles and textbooks refer to the NB2 parameterisation. NB2 models give higher weights to values with low mean counts, levelling off at $1/\alpha$, when calculating the scale parameter. If the variance is assumed to be overdispersed in a manner proportional to the mean count, NB1 is preferable, with NB2 being useful if the effects of overdispersion are likely to be proportionally higher with low mean counts, or stable across larger means counts.

NB models are more challenging to estimate than Poisson GLMs, and often iterate between a fitting procedure, and a scoring procedure for the scale parameter until convergence (Venables and Ripley, 2013). Sellers and Shmueli (2010a) quote (McCullagh and Nelder, 1989) suggesting NB2 models to be *“an unpopular option with a problematic canonical link.”*

- **Generalized Poisson (GP)** models (Bae et al., 2005, Famoye, 1993) allow for a more complex scale factor, α , where the variance is modelled as: $\mu_i(1 + \alpha\mu_i)^2$. When $\alpha = 0$, the model reduces to a Poisson distribution and when $\alpha > 0$, the model indicates overdispersion. This parametrisation allows both under and overdispersion to be modelled, and belongs to the exponential family of distributions in the case of a constant dispersion parameter. The disadvantages of GP are they no longer belong to the exponential family of distributions if the dispersion is observation-specific (Sellers and Shmueli, 2010a), and they are not in common use when compared to NB or quasipoisson models.
- **Conway-Maxwell-Poisson (COM-Poisson)** (Conway and Maxwell, 1962, Shmueli et al., 2005) proposes a more general form of the Poisson model. This extends the principles of the GP models, using a different scale factor ν , the rate of decay of successive ratios of probability, such that the variance is then modelled as: $\frac{1}{\nu}\mu_i$. This allows a distribution of expected counts to vary with the dispersion parameter, allowing the ratio between two consecutive values to be non-linear (Sellers and Shmueli, 2010a). COM-Poisson models were tested for the coming chapters but have not been included in the results due to implementation issues in R (the statistical coding language used for this project). Issues included not supporting varying dispersion parameters, lack of convergence, and estimation issues.

Table 4.1 summarises the model types and variance functions discussed above.

Model	Variance	MLE	Parameter estimates conditional on variance?
Poisson GLM	μ	Y	N
Huber/White Robust Standard Error	Sandwich estimator ((White, 1980)	Y	N
Quasi-Poisson	$\mu\theta$	N	N
Negative Binomial 1	$\mu_i(1 + \alpha)$	Y	Y
Negative Binomial 2	$\mu_i + \alpha\mu_i^2$	Y	Y
Generalized Poisson	$\mu_i(1 + \alpha\mu_i)^2$	Y/N	Y
Conway-Maxwell Poisson	$\frac{1}{v}\mu$	Y	Y

Table 4.1 Summary of model variance functions for Poisson-based, variance scaled, and mixture models.

4.2.4 Multilevel models

Overdispersion in models may represent latent structures or correlations within data, where points are not independent (a key assumption in the GLM model). Goldstein (2010) described these structures as “*neither accidental nor ignorable*”, and explicitly modelling them is a logical step when clustering is suspected.

A class of models commonly used for these latent structures is the ‘multilevel’ model (also referred as ‘mixed model,’ ‘hierarchical models’ or ‘random-effects models’). These models are applicable in a variety of situations where data are not independent, but are related at different levels such as repeated measurements over time or at different levels of aggregation (Jackson et al., 2008). In these structures, the variance may be partitioned into distinct levels (‘variance components’). Multilevel models are particularly suited to clustered data, where the clustering introduces correlations (Breslow and Clayton, 1993), such as ‘between and within’ variation in a multi-centre clinical trial. We may expect a given measurements to be normally distributed across subjects, but the clusters within which measurements are made also exert effects (e.g. demographics associated with a centre, poorly calibrated equipment at some centres, effects related to clinical practice within centres etc.). Measurements from each centre are likely to be related, and data cannot be viewed as truly independent.

A common extension to the GLM framework, and sharing connection with quasi-likelihood models, is the use of Generalized Estimating Equations (GEE). GEEs allow for estimation of models with correlations due to longitudinal data. They can account for the effect of the specified correlations or covariance structures when estimating parameters. GEEs do not require a fully specified distribution and estimation proceeds from the first two moments of the distribution only (Liang and Zeger, 1986). GEEs are robust to a degree of misspecification of the correlation structure, but may be limited by their predictive scope. GEEs predict the ‘population-average’ or ‘marginal’ effects (further discussed in Chapter 8). This ‘removes’ the effects of clusters and does not allow cluster-specific (‘conditional’) predictions. The lack of a fully specified distribution, and therefore an MLE, also limits their value for model comparisons using likelihoods.

The Generalized Linear Mixed Model (GLMM) is a more complicated approach in comparison to GEE. It allows the modelling of different variance components (‘random-effects’) as well as the explanatory variables (‘fixed effects’) that have, so far, been referred to as ‘predictors’ in single-level GLMs. There are many definitions of fixed and random-effects (Gelman and Hill, 2006a), and predictors can be modelled as either fixed or random depending on the assumptions underlying a particular model. The choice of which random-effects to include should be considered as part of experimental design, considering where they are logically and theoretically possible. They can be related to hierarchies, e.g. patients recruited within treatment centres in a clinical trial, repeated measures e.g. from the same patients over time, or using data at different levels of aggregation.

A basic GLMM structure follows from a GLM (Wood, 2017b), as:

$$g(\mu_{ij}) = X_i\beta + u_j, \quad u_j \sim N(0, \theta)$$

With an additional random variable: u_j , the random-effect, for j th cluster. Random-effects are usually assumed to be normally distributed with mean zero and a standard deviation θ . In matrix form, the model now also depends on a dispersion matrix of unknown variance components (Breslow and Clayton, 1993). Only the variance of the random-effect is estimated, as the model coefficients remain fixed at zero. Fixed effects are now, also, conditional on the estimated random-effects.

In contrast to GEEs, GLMMs can be used to examine both the ‘conditional’ (cluster-specific) model predictions, but also make marginal predictions by integrating out the individual/cluster effect or predicting for a subject where the random-effect estimate = 0 (Lee and Nelder, 2004). They have fully specified distributions, MLEs and allow for comparison between models using likelihood ratio tests and AIC.

4.2.4.1 Relevant multilevel structures

A variety of multilevel structures can be fitted to data, including intercepts, slopes and random categorical coefficients, as well as cross-classified random-effects. The structures most relevant to repeated measures in count data include:

- **Random-intercept models:** where cluster-specific intercepts allow for deviation from the global intercept within cluster (see Figure 4.1). Models of this type address overdispersion due to clustering. The effects of x remain constant for each cluster, but the difference between the cluster-specific intercept and global intercept is captured in the random-effect.

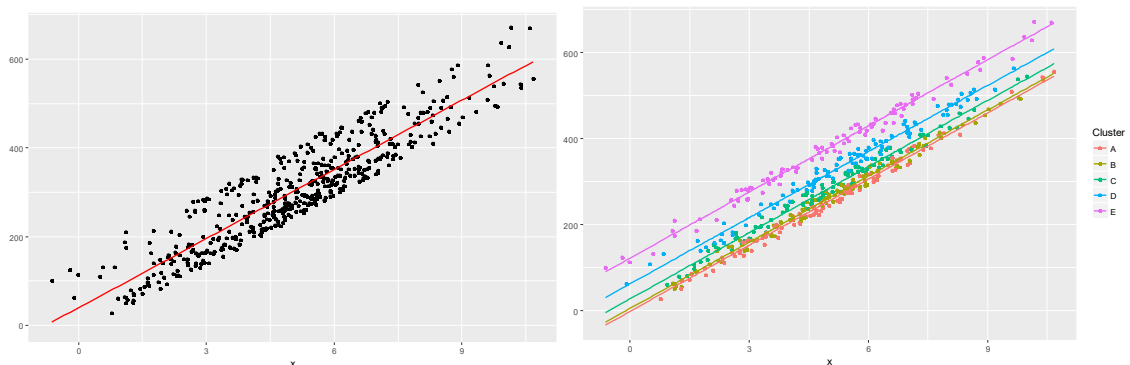


Figure 4.1 *Illustration of single-level and random-intercepts on simulated data*

Left panel shows single-level regression of x against y , with a single intercept. Right panel shows random-intercepts for coloured clusters within the data, with the effects of x constant.

- **Individual-level random-effects:** A special case of multilevel models is to allow the overdispersion itself to be modelled as a random-effect. This can be considered as a Poisson-lognormal model in comparison to the NB2 models, discussed earlier in this chapter, that are Poisson-gamma. An index term can be used for the random-effect ‘mopping up’ the residual overdispersion (Elston et al., 2001, Rabe-Hesketh and Skrondal, 2012). This is computationally intensive and, in some cases, requires Markov chain Monte Carlo (MCMC) methods to estimate. This method also creates problems for likelihood profiling, which involves systematically varying a parameter whilst holding others constant. The individual-level random-effect may alter

drastically in this process as it is always attempting to scale to the residual variance, causing errors in profiling procedures.

4.2.4.2 Accuracy of estimation

Models in the following chapters were primarily estimated using the statistical programming environment `R`, using the `lme4` package (Bates et al., 2015), and its `glmer` function, considered the standard function for this in `R`. The package estimates GLMM models by Laplace approximation (Laplace, 1986) (originally 1774), a technique applied to the maximization of penalized likelihood. Laplace approximation is considered an improvement on pseudo-likelihood methods (SAS Foundation, 2018), and works by sampling the marginal likelihood at one point. This single integration point might be considered inaccurate when more than one point can be evaluated.

Gauss-Hermite Quadrature (GHQ) (Pan and Thompson, 2003) is considered an improvement on Laplace approximation, particularly for the estimation of random-effects. Although GHQ samples the marginal likelihood at more points, it is more restricted in the models it can fit, restricted to a single random-effect in `lme4`. Despite these concerns, the accuracy of Laplace approximation has been demonstrated with the number of random-effects less than $n^{1/3}$ (Shun and McCullagh, 1995), where n is the sample size, giving the least biased estimation with lower prediction error (Handayani et al., 2017) than GHQ, using `lme4`. This may, however be a result of the implementation in this particular modelling procedure or a quirk of a particular data set. It is advisable for models fitted by Laplace approximation, for speed, also be checked by fitting by GHQ, potentially in different statistical packages to verify the accuracy of estimation. When developing models in subsequent chapters, models were fitted using both GHQ and Laplace methods in `R` and SAS, with no significant differences observed between methods or software.

For NRLS-based models, we have good reason to suspect cluster effects from repeated measurements due to monthly reporting at organisations. It is reasonable to assume that monthly reporting rates from one organisation may be correlated due to the exposure and culture variables at that organisation. Failure to model this structure will affect estimates of the error in any model.

4.3 Scaling/standardizing of covariates

Model predictors of differing units may be on differing natural scales, e.g. a proportion from 0 – 1, a given count variable may range from 1 – 10,000, or the natural logarithm of the same variable ranging from ~0.00 – 9.21. Transformations of data are primarily used to alter the distributional form to reflect better the relationship between variables, but they usually change the scale of the data as well.

Large differences in scale can make models harder to interpret and, particularly for GLMMs, can lead to convergence problems (Bolker, 2017). Scaling of variables can be performed, without altering the distribution, and may aid model convergence.

It is also common to scale data to assess the relative importance of predictors on similar scales and allow easier calculation for GLM and particularly GLMM estimation (Gelman, 2008).

Scaling can be performed in many ways, such as dividing by a constant, or “min-max” scaling where values are scaled between zero and one based on their minimum and maximum values.

A common recommendation in regression literature (Schieleth, 2010) and statistics internet forums (e.g. ‘Cross-validated’: <https://stats.stackexchange.com/>) is to centre variables on their mean and scale by the standard deviation, creating a z-score representation.

For each numeric x , the standardisation is:

$$\frac{x - \bar{x}}{\sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}}$$

This representation renders the mean as zero, with a change of one in x representing an increase of one standard deviation. This alters the interpretation of the model, allowing each model coefficient to be interpreted as the mean value for the parameter, and the intercept as the average value when all parameters are at their mean. Coefficients from models that are not centred and scaled should be interpreted as the change for one unit of x , holding all other parameters constant, and the intercept as the value of y when x is zero. Once centred and scaled in this manner, model coefficients can be directly compared to each other within models for their relative effects, without larger parameter values dominating. If models are applied to new data, a decision must be taken over whether to scale according to the mean and standard deviations in the data set, or whether to recalculate the model given the mean and standard deviations which will change with different inputs.

Gelman takes this argument further, suggesting that mean-centred variables are divided by two standard deviations rather than one (Gelman, 2008), to allow scale to be matched with untransformed binary variables. The rationale behind this is to use a consistent set of comparisons between low and high values, across different types of predictors, comparing units that differ in their input values by two standard deviations. The interpretation of this becomes, the change in y for a change of two standard deviations in x . This is approximately similar to the comparison of zero/one for a binary predictor, and can effectively be viewed as the change from 'low to high' for a variable.

4.4 Predictive versus explanatory models

Models are generally created with either an explanatory or a predictive goal. In an explanatory model, parameter estimates are of primary importance, but the goal in a predictive model is accurate prediction and generalisation to new data (Shmueli, 2010). These modelling paradigms necessitate different approaches to 'significance' of parameters and assessing model performance.

When assessing a model, parameter estimates and their standard errors should be inspected. Large standard errors may indicate poor parameterisation which, in turn, may suggest a lack of support in the data or correlations with other predictors. Parameter estimates will show which factors contribute most to prediction and indicate whether some effects are dominant. Although parameter estimates are of interest in all models, they are a major output for an explanatory model. The primary purpose of NLS models presented in the following chapter is prediction. Predictive models may require the inclusion of parameters that are not directly interpretable to the reader (e.g. projected features or neural network weights), or 'significantly' associated with the outcome. Models may also be "*inferior in terms of parameter bias, but superior in terms of predictive accuracy*" (Sellers and Shmueli, 2010b). Comparisons of predictive accuracy are the primary measure of model fit in this case. In predictive modelling, distinctions should be made between testing and training data (see section 4.5.3).

The significance of parameter estimates is sometimes used to decide which covariates to include/retain in a model, particularly in explanatory models. This approach is not ideal and can lead to incorrect assessment of significance. The conditional distribution of a reduced

model may differ from the full model, as parameter estimates are conditional on each other, and should all be reassessed if a parameter is removed. Confounding predictors, if removed from a full model when they do not appear 'significant', may lead to poorer overall performance of the model due to this conditional nature. An argument can be made for including all factors with a rational, theoretical mechanism/justification in a model regardless of their estimated significance.

There are also issues around degrees of freedom and multiple comparisons when removing predictors from a full model. A reduced model may be presented as if it is a full model, despite being chosen from a larger set of predictors, and its significance may then be overstated by assuming reduced degrees of freedom. Shrinkage methods such as LASSO regression (Tibshirani, 1996), where parameters are shrunk towards zero if they are not predictive, may provide better solutions. Harrell discusses problems related to this at length, particularly regarding step-wise regression (Harrell, 2001, Harrell et al., 1996).

Variables used in predictive models should, therefore, be chosen a-priori for theoretical reasons rather than being data driven where possible.

4.4.1 Presentation of coefficients

Poisson GLM parameter estimates are commonly presented as Incidence Rate Ratios (IRRs), calculated by exponentiating the model coefficients, due to the use of the log link function. The additive/multiplicative relationship between the original and log scale make these estimates multiplicative when transformed back. It is common to interpret them as multiplicative effects using IRRs, multiplying Y by the IRR for each increase in 1 in the corresponding X covariate, whilst holding all other predictors constant. Although model coefficients are presented in the following chapter, they have not been transformed to IRRs, as the focus is not on direct interpretation of model predictors. Rather, differences in parameter estimates across models, significance of groups of factors (e.g. age related), and relative sizes of standard error are worth noting when comparing different model classes. The emphasis is, however on predictive ability.

4.5 Assessment of model fit and performance

A well-fitted model is important, when trying to describe or predict, and model fit can be assessed in a variety of different ways. Examining measures of model fit, and predictive accuracy may both be used.

4.5.1 Model diagnostics and global fit

Plots of model outputs, including distribution of residuals and observed values against model predicted values, can be used to examine if there are any patterns or deviations from expected outputs. Standardised residuals are useful when comparing overdispersed Poisson models, as they are scaled by the standard deviation.

‘Goodness-of-fit’ in Poisson models is commonly assessed using chi-squared tests, comparing the Null deviance (one degree of freedom per data point) with the residual deviance of the fitted model. The null hypothesis of such a test is that the model is correctly specified, but in cases of overdispersion, the deviance is strongly affected by this, rendering this test unhelpful.

Nested models are commonly tested in terms of the reduction in residual variance, penalised by the number of additional parameters or degrees of freedom. In linear models, this can be performed using an f-test, or in Poisson and Logistic models, the likelihood ratio can also be tested (the ‘likelihood ratio test’ or LRT) using a chi-squared/Wald test. Where models are nested purely in fixed effects, or purely random-effect terms, these tests are valid. When testing a fixed effect model against a random-effects model, where the fixed effect model is considered nested within the random-effects model, this method is problematic for two reasons (Bolker et al., 2009, Greven, 2088):

1. The estimation of degrees of freedom for random-effects is debatable, with complex random-effects hard to define. They are commonly defined as a single degree of freedom per variance parameter, but ‘effective degrees of freedom’ may be more suitable when complex random-effect structures are present (Bolker, 2018, Wood, 2017b)
2. Wald tests are based on the normal distribution/z-test, but when comparing against a fixed effect model, the null hypothesis is that the residual deviance is zero. The test is therefore on the boundary of the parameter space, i.e. the variance is not normally distributed around zero, and can only be greater than or equal to zero (Molenberghs and Verbeke, 2005).

These concerns render the likelihood ratio test inaccurate for comparing the random-effects and fixed effects models. The boundary issue renders the LRT conservative with p-values approximately twice as large as they should be (Pinheiro and Bates, 2000).

A common generalisation of the LRT is the Akaike Information Criterion (AIC) (Akaike, 1998). AIC is an estimate of relative Kullback-Liebler divergence, and is based on the log-likelihood, but adding a parameter to penalize more complex models (Bolker et al., 2009):

$$AIC = -2 \ln(L) + 2k$$

Where L is the maximised likelihood and k denotes the number of parameters. It allows models to be compared but does not require them to be nested. Smaller values of AIC suggest that models (based on the same dataset) lose less information, and models with smaller AICs are generally preferred.

A small sample size correction is commonly applied to AIC as it can be prone to overfitting. Referred to as AICc, it is suggested when the ratio of the number of data points to model parameters < 40 (Burnham and Anderson, 2004). This is not the case for the all incident models, and traditional AIC is used in the following chapter but AICc used for the simplified death and severe harm models.

The use of AIC with multilevel models is controversial, with evidence that it is appropriate for the selection of fixed effects. The boundary constraint, mentioned above in relation to the LRT, renders it biased and it favours smaller models without random-effects (Greven and Kneib, 2010, Bolker, 2018). Greven and Kneib also note that the level of focus of a model affects which calculation is appropriate, e.g. population level prediction (using marginal 'mAIC', the usual AIC definition) or cluster-specific based on conditional modes (using a conditional 'cAIC' that they propose). In most statistical packages, mAIC is the common implementation, although cAIC is available in R. AIC can be used in comparison of models, but it should not be relied upon to distinguish between single-level and multilevel models.

4.5.2 Predictive performance

Competing models with different representations may need to be compared to identify the best approach. Model selection is traditionally performed to do this, but there are few 'perfect' ways to do this. Maximisation of predictive ability of a model, or reduction in residual variance are standard approaches. Common ways to select on predictive ability vary between different model classes and include measures such as sensitivity/specificity or the 'area under the Receiver Operator Characteristic' (ROC) curve (also known as the 'C-statistic') for binary classification models such as logistic regression (Harrell, 2001).

When applied generally to regression, the error rate of prediction can be calculated, commonly:

- Root Mean Squared error (RMSE): $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
- Mean Absolute Error (MAE): $MAE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$

Where y_i is the observed value and \hat{y}_i is the predicted value.

Whilst these measures are both commonly used, the square-root transformation on RMSE gives proportionally more weight to large error values. MAE might be considered the better choice if outlying values are to be treated equally. Although both approaches are presented for each model, it is uncommon to use them in isolation, as it is difficult to prove that minimum prediction error alone is a sufficient summary of both error and bias, so a measure of minimum variance is often used as well (Wood, 2015).

If model predictions are tested solely on the dataset used in fitting, overfitting may not be considered an issue as the model is not being generalized, but our models will reflect the idiosyncrasies of our data sample as well as the underlying relationships. In most predictive contexts, generalization is the aim. Prediction error should therefore be tested on data that were not used in the modelling process. In machine learning publications, and some predictive modelling settings, datasets used to build a model may be referred to as the ‘training set,’ and a separate set referred to as the ‘testing set.’ Some modelling paradigms (such as Boosting or Artificial Neural Networks, see Chapter 6) consider a further split with a ‘validation set’ that is used as an out-of-sample test during model training, but it is not used in the final assessment of models (where we would use the testing set).

K-fold cross-validation methods are similar in a sense, because a holdout sample (k) acts as a testing set whilst the rest of the data is used for training. Care should be taken when using k -fold cross-validation (and bootstrapping), or when splitting data for training and testing of multilevel models. The subsamples should be representative of the random-effects structure and models may become unstable if too few points are used. Decisions must also be made about the handling of new random-effects levels in a testing set. They may be excluded or fitted with zero estimates for random-effects, given that they have no precedent in the training set. Fitting them with an estimate of zero corresponds to the global average for random-intercept models, and is a rational choice, but becomes more complex if elaborate random-effect structures are used.

4.6 Summary

Count data are integer values with a boundary at zero, making them unsuitable for linear modelling via OLS techniques. Generalized Linear Models based on the Poisson distribution are the common approach for fitting such data but are often hampered by overdispersion, where the variance is underestimated, commonly due to misspecification. Scaling of model variance, mixture models and multilevel modelling techniques can be used to model overdispersion, but all have different assumptions and particular techniques will suit some datasets, and questions, but not others.

Explanatory models are often tested by verifying assumptions about the distributions of the data and may use likelihood ratio tests or AIC to compare models. These methods may be hampered by overdispersion or invalid for comparisons between GLMs and GLMMs.

Predictive models are best compared in terms of their ability to predict data, preferably on a new 'testing' dataset. Their performance on new data, or measures of cross-validation or bootstrapped error, reduce the chance of overfitting the training data. Mean Absolute Error can be considered as a good measure of a model's predictive accuracy, as it is less sensitive to extreme outliers when compared to root mean squared error and is preferable in cases of overdispersion.

The techniques described in this chapter will be fitted to NRLS incident reporting data in Chapters 5 -7, with some of the concepts relevant to text mining models in Chapter 9. This chapter contributing to aim 3 of the thesis, by developing the statistical framework for dealing with count data and overdispersion. These methods form the basis for subsequent modelling stages and how models will be compared and tested.

Chapter 5 Count models of NRLS

This chapter will contribute to aim 3 of the thesis by developing the first sets of statistical models. Published analyses on NRLS, examined in my literature review, and summary data (chapters 2 & 3) suggested it is poorly categorised and unsuitable for direct use in statistical models. Chapters 2 and 3 proposed models using an alternative source of casemix data, to be treated as a count dataset at the same level of aggregation as the outcome data from NRLS. This chapter explains a theoretical approach to NRLS modelling, the construction and examination of a modelling dataset using the secondary data source, and presents the results of initial count models.

This chapter is structured as a flow of model development, with each section building on the last, rather than forming distinct experiments. This flow introduces the secondary data sources, selects the appropriate data structures for modelling, applies models based on Chapter 4, tests for overdispersion and identifies and interprets the best performing model. The sections proceed as:

- 5.1 Theoretical model of incident reporting
- 5.2 Introduction to, and preparation of the Hospital episode Statistics (HES) as an alternative source of casemix data
- 5.3 Examining and choosing parameterisation of predictors for models
- 5.4 Fitting of Poisson models and single-level overdispersion-adjusted models
- 5.5 Fitting of multilevel, overdispersion-adjusted models
- 5.6 Model selection
- 5.7 Extending models to longer time periods

5.1 Theoretical model

In order to model NRLS data, and interpret it in practice, a conceptual model has been developed to explain factors that may affect reporting. This model is based on the literature review in Chapter 2, discussion with clinical and academic supervisors, and NHS Improvement (NHSI).

As an overarching theoretical model, we can consider incident reporting in NHS hospitals to be the combination of two functions:

$$\text{Incident reports} = f_1(\text{Exposure}) + f_2(\text{Culture}) + \text{natural fluctuations/error}$$

Where:

- **Exposure** is the opportunity for incidents to occur, e.g. a larger hospital could be expected to have more incidents than a smaller one due to more patient contacts, members of staff, equipment and facilities in use etc.
- **Culture** is, itself, a combination of 4 further functions:

$$\text{Culture} = f_3(\text{Safety Behaviour}) + f_4(\text{Awareness}) + f_5(\text{Priority}) + f_6(\text{Process})$$

Where:

- **Safety Behaviours** are a system's defensive actions to prevent, and learn from, errors. E.g. regular examination of incident reports, implementation of findings/changes, feedback to users of reporting systems etc.
- **Awareness** is the extent to which staff notice and categorise events as 'incidents.'
- **Priority** is the organisational interest in incident reporting, usually promoted by senior staff, and made visible throughout an organisation.
- **Process** is the systems, barriers and enablers for incident reporting. E.g. Well-designed or confusing incident forms, availability of incident forms/computers for access to electronic reporting, computer failures that may impact the reporting etc.

Most cultural factors are likely to be local (operating at hospital-level, or even ward/team level), but there may also be cultural factors driven by regional or national priorities, such as changes in Never Event policy or 'pathways' across several local providers.

The functions above can be viewed as 'latent variables' that cannot be directly measured. We must, instead, represent them using proxies that convey similar information, or reflect them in the structure of models (such as stratification or using correlation structures like random-effects or GEE, discussed in Chapter 4). Some of these effects are likely to be confounded, and latent variables may be represented by more than one proxy, leading to additional correlations.

The effects of exposure will be examined by the initial models in this chapter. Exposure has the most accessible proxies from casemix variables available from other secondary datasets. Culture will also be considered in later models by using random-effects. These models make allowances for hospital-specific variation, whilst allowing the average effects of exposure to be estimated globally for the model. Models can then be used to make predictions at the model average (marginal) values for a given set of casemix variables.

The rest of chapter 5 examines whether hospital-level exposure variables are associated with rates of incident reporting and whether they can be used to predict the number of incidents reported. Model output is intended for use in tools for reporting organisations and regulators NHSI or the Care Quality Commission (CQC), and these are further developed in Chapter 8.

5.2 Dataset construction

A hybrid dataset was constructed to examine the effects of casemix/exposure using a secondary source of casemix data. The dataset was constructed from incident reports submitted to NRLS by July 2017 (as reports can be retrospective), for incidents in fiscal year 2015/16, and Hospital Episode Statistics (HES) data for the same period (NHS Digital, 2016). Single-year models are the focus of this chapter, but models were also extended to five-year periods to test for stability and changes over time (see section 5.8).

5.2.1 Hospital Episode Statistics (HES) data

The HES is an England-wide data warehouse of hospital activity (NHS Digital, 2017d), containing 19.3 million inpatient (IP) episodes, 113.3 million outpatient (OP) appointments, and 20.3 million accident and emergency department (A&E) attendances in 2015/16.

HES records are anonymised, patient-level data entries that include details such as admission & discharge dates, types of hospital stay (e.g. maternity/birth episode, emergency admission, surgical episode etc.), hospital treatment specialties, treating organisation, referring organisations etc. Patient demographics including age on admission, sex, and deprivation group, as well as clinical diagnoses and procedures/interventions performed during the patient's visit. Although HES are collected for inpatients (IP), outpatients (OP), and accident and emergency (A&E) attenders, with specific extra sets for critical care and maternity, only

the IP dataset is mandatory. Despite monthly fluctuations, and errors at submitting organisations, HES data completeness is considered good (Herbert et al., 2017), and preferable to small clinical databases due to its coverage and completeness in some settings (Royal College of Obstetricians and Gynaecologists, 2012).

Hospitals in England are obliged to submit activity data through the 'Payment by Results' (Department of Health, 2013a) process where provider organisations essentially 'invoice' commissioners for the work they perform. The precursor to HES, described below, is akin to a payment receipt for the activity of a provider organisation for each patient they see. The system has a number of stages:

1. At the end of each month, patient medical notes and entries in other data systems are manually read and coded ("abstracted") by professional clinical coding staff. Coders use common national and international standard definitions, including ICD-10 (World Health Organization, 2017) for diagnoses and OPCS-4 (NHS Digital, 2017a) for procedures, that define both the codes and the rules for their use.
2. The assigned codes and admissions data from hospital systems are extracted, by each organisation's IT or informatics teams, to meet a common data standard known as the 'Commissioning Dataset' (CDS) (NHS Digital, 2017b).
3. Monthly extracts are uploaded to the Secondary Uses Service (SUS), a data warehouse administered by NHS Digital (NHS Digital, 2017e). Healthcare Resource Groups (HRGs) (NHS Digital, 2017c) are assigned to spells. These codes group activity into casemix-adjusted, chargeable units of provider resource, primarily for payment purposes. HRGs can be considered broad groups of similar activity, such as the HRG4 code = 'HB23C'. This is described as 'Intermediate Knee Procedure category 1 for trauma, without complications.' The codes often have national tariff values assigned and expected lengths of stay (LOS) associated with them (Department of Health, 2013b). Incentives, top-ups for long-staying patients and penalties are also included and may change from year to year.
4. Commissioning organisations use SUS to determine payments to providers, or dispute and resolve claims in some cases.
5. Once completed (approximately 3 months), the national data set is cleaned by NHS Digital, fully anonymised, and made available for selected users for research and monitoring, with the permission of ethical and information governance groups (NHS England, 2017).

For this project, HES will be used to quantify casemix and compare the burden of hospital activity, against NRLS reporting, adjusting for quantity and type of exposure.

Data access for both HES and NRLS was made available via University Hospital Birmingham NHS Foundation Trust's HES data licencing agreements with NHS Digital (details available on request), and NHS Improvement. This project was considered an audit, with a registered UHB audit number CARMS-13548. Outputs will be used to create an interactive data tool, delivered through the benchmarking system 'Healthcare Evaluation Data' (HED), see Chapter 10 for further details.

5.2.2 Aggregate dataset for modelling

Record-level data linkage is not possible for these two datasets, as both HES and NRLS are anonymised (individual level, but with no useful identifier to connect them). They are also collected in different units: HES at 'episode' level (a patient's time under the care of a given consultant/team) and NRLS at incident level (that may not necessarily relate to a patient, and may also relate to staff, equipment, environment etc.). Both datasets can, however, be transformed into contingency tables with aggregated counts of variables/incidents in time periods, at corresponding levels and analysed using the techniques discussed in Chapter 4.

Counts of incidents at each level of harm (no harm, low, moderate, severe and death), per organisation, were made from NRLS and collated with HES-based counts of IP bed days in age groups, sex categories, admission method categories (elective, non-elective, maternity/birth and transfer), and comorbidity score categories. The Charlson comorbidity index (Charlson et al., 1987) is a weighted index of chronic conditions, originally derived for a breast cancer study cohort in the USA. It has been re-weighted for UK data and coding schemes, with changes including reduced weight for HIV since the introduction of Highly Active Anti-Retroviral Therapy, and the inclusion of dementia. This is grouped according to the standard used in the Summary Hospital Mortality Indicator, in groups of <1, 1-4 and >4, representing health, minor comorbidities and major comorbidities (NHS Digital, 2017f). IP casemix factors were chosen as they were considered likely to affect the risk of incident, e.g. elderly patients are at higher risk of falls in general (Healey et al., 2008).

The proportion of bed-days that were admission days was also added to the dataset to examine whether admission day carried additional risk. This may be particularly pertinent for emergency admissions.

Bed-days for surgical admissions were specifically identified as surgery involves very different pathways of care and opportunities for different types of incident, compared with general medical admissions. Surgical admissions are not specifically defined in HES, and distinguishing valid surgical procedures across all specialties is a complex process, beyond the scope of this project. Admission to a surgical specialty was therefore used as a proxy. Admission to surgical treatment specialties was defined using HES field 'TRETSPPEF' with values >99 and <200, plus 502. No check was made for whether patients received a procedure, as some patients may not receive procedures if they are too ill or if other emergency cases mean there is no available capacity.

Counting HES activity in terms of episodes or discharges leads to an inequity in the exposure. If counted as the number of discharges (a common HES analysis definition), an inpatient staying in hospital for two days would appear to have the same exposure as a patient staying one day. This is not an adequate proxy for the 'opportunity for incident.' If patient occupancy can be considered as 'exposure time,' it is reasonable to consider a two-day patient stay as twice the exposure of a one-day stay.

The most accurate measures of exposure time would be the time between a patient's admission and discharge. This is not currently possible to calculate from HES, as it does not contain admission and discharge times. A common method to account for this varying exposure, given the lack of time information, is to convert the time in hospital to counts of patient 'bed-days:' the number of days inpatients are occupying beds. The highest resolution for bed-day counts in HES is whole days, as the difference between admission and discharge dates that are recorded in HES.

Transforming to bed-days is a common concept in HES analysis and several methods exist for this:

1. **Discharge date – Admission date** (Department of Health, 2013b)
2. **(Discharge date – Admission date) + 0.5:** in recent NRLS publications (Clinical Indicators Team, 2016a)
3. **(Discharge date – Admission date) + number of day-cases** (Medicines & Prescribing Team, 2015)

4. **(Discharge date – Admission date) + number of single day stays** (Bardsley et al., 2016)
5. **Quarterly midnight count of occupied beds, and day-only beds** (Macfarlane et al., 2005)

Method 1 renders single-day stays (or the sub-group of ‘day-cases’) as zero. This is clearly inadequate when considering exposure time. Short-staying patients still had exposure despite that exposure being less than a day. Method 2 adds a constant bias of 0.5, rather than zero, to account for the admission day. This is consistent but may be unrepresentative as there is no way to assess whether patient stays are longer or shorter within the frame of a calendar day i.e. 0.5 may represent, at its extremes: 00:01 to 23:59 hours, and is unlikely to be exactly half a day. Adding 0.5 is also problematic because our input is no longer an integer, conflicting with the Poisson distribution’s definition on integers (see Chapter 4). This may add bias where hospitals with differing admission and discharge practices for short-staying patients may appear the same. Method 3 specifically counts day-cases, but so-called ‘zero-day stays’, where a patient is admitted and discharged on the same day but are not considered as ‘day-cases.’ In HES terms, a day-case must be admitted with the intention of being treated as such (HES data fields: [INTMANIG] = 1 & [CLASSPAT] = 2, (NHS Digital, 2017d) with a LOS <1). Day-cases are usually admission for particular procedures, but zero-day stays may occur for other reasons (e.g. emergency admission from A&E for monitoring). Method 3 gives equal weight to day-cases and over-night stays but misses a group of patients with LOS <1 such as the short-stay emergency admissions described. Method 4 attempts to deal with unequal weighting by acknowledging any attendance as at least one day, but it is inequitable in a sense, valuing an over-night stay the same as a single day-stay. Method 5 is the NHS standard, used in national reporting of NRLS figures, however it is known to exclude critical care, ‘well babies’ and several other circumstances. It is collected through a long-standing quarterly, ‘central return,’ known as the ‘KH03’ (NHS England, 2017). KH03 data is not collected with other demographic details attached and is therefore less useful for risk-adjustment.

None of the existing methods with access to further demographic data (1-4) adequately cover time in hospital from an exposure perspective, so a new calculation is proposed here. Bed-days were counted with reference to a calendar look-up table, so all open spells on a given day are counted and summed within months, weighting all days in hospital as 1, including zero-day stays and day-cases. This method will be an over-estimate, as patients do not necessarily stay for all of this time, as described above, but the over-estimate is consistently applied across all organisations. This may particularly affect organisations with high day-case procedure rates

and increase bias. It will count more 'exposure' than was strictly true, but given the absence of time data in HES, this seems the most equitable method. IP activity was therefore considered as counts of 'bed-days' in casemix groups.

Bed-days are helpful for quantifying in-patient exposure, but OP and A&E attenders also represent major patient flows through hospitals. Since details in OP HES were sparser than IP HES, the OP patient flow was included in our models but counting attenders stratified by age group. Age was assumed to be a potential factor in mobility, medication and other incidents, but sex was assumed to be less relevant to outpatient attendances and the exposure risk. Coding rules for outpatients are less stringent than inpatients, and are not mandatory in many cases, therefore Trusts will not waste resources on this data if it is not mandatory. Outpatient comorbidity scores were therefore not considered. OP HES also contains records of appointments that patients did not attend, so an exclusion was applied to model only those records where patients attended.

A&E HES data are collected nationally, but data submission and completion of fields has varied as they were not mandatory until the recent introduction of the Emergency Care Dataset (NHS Digital, 2018a). The nature of A&E also makes data completeness an issue, as patients may not be easily identified in some cases such as being unconscious on admission, confusion in ill patients or patients leaving the department without being seen, leaving unresolved records. A&E patient flow was considered in the model as numbers of A&E attenders and waiting times, with the rationale that longer waiting times suggest more exposure for patients who are in the hospital for longer, a proxy for the pressure the department is under.

5.2.3 Organisations included

Models focussed on all data submitted to NRLS by acute trust, (excluding specialist single-specialty hospitals such as orthopaedic, or paediatric trusts) with corresponding HES data in the fiscal year 2015/16. Trust type has been identified using the NHS central Organisational Data Service, rather than location or hospital type recorded in the NRLS. In rare cases, trusts may have multiple hospital sites that include different organisations, such as mental health facilities or community hospitals. Where such organisations are part of larger acute trusts, they have been included in the model and reported as part of that trust.

NHS structural changes have occurred regularly, with organisations opening, closing, or merging, so a consistent set of processing rules for organisations was required. Changes in organisational grouping affect the number of repeated measures at each organisation and will affect estimates of random-effects structures (see Chapter 4). Mapping rules were applied as follows:

1. Organisations with a stable identity through the period pose no problems, and are identified consistently. This represented the majority of Trusts in the data.
2. Organisations who fully merged during the modelling period were mapped to the merged organisation for the entire modelling period.
3. Organisations that did not fully merge, but were split between more than one merger organisation, were mapped to the original organisation for the duration of its existence, and the new organisations after its creation. E.g. Mid-Staffordshire hospital closed during 2014/15 and was split between two organisations and could not be fully mapped to either of the merging Trusts for the whole period. Therefore Mid-Staffordshire exists at the start of that fiscal year, and North Midlands NHS Trust, and Cannock Hospital at Wolverhampton have received additional activity and incidents after the merger.

5.2.3.1 Exclusions

A variety of validation checks were performed on the data, including checks for missing values, unlikely values such as a single incident report in a month, and monthly shifts of 50% or more in reported numbers of incidents, bed-days, OP or A&E attenders. Initial modelling stages also fed back into the validation process and four data points were excluded for 2015/16:

- 03/2016 – University Hospitals North Staffordshire NHS Foundation Trust. High leverage and residual on model diagnostics.
- 03/2016 - Central Manchester University Hospitals NHS Foundation Trust: High leverage and residual on model diagnostics, due to missing A&E data.
- 01/2016 - Chelsea and Westminster Hospitals NHS Foundation Trust: large drop in the number of incidents reported, unlike previous months.
- 03/2016 – Guy's and St. Thomas NHS Foundation Trust: Substantial drop in IP bed days (particularly emergency admissions) and large rise in A&E attendances. Suggestive of change in admission policy in last month of period.

A total of 1616 data points entered the models from 135 organisations.

5.3 Parameterisations of incident and exposure data

Explanatory variables in the dataset, described in section 5.2, required a degree of aggregation or grouping to be adequately linked to the incident data. The following sub-chapter describes possible parameterisation that were considered, and in some cases tested. Final parameterisations were arrived at by considering the theoretical reason for a fitting in a particular form, parameter scaling, changes to distributional form and convergence of models.

5.3.1 Describing exposure data in the aggregated dataset

Three potential aggregation methods were examined, based on IP, OP & A&E HES data and NRLS incidents per organisation per month. Each potential parameterisation was assessed from a logic/study question perspective, and practical perspective for fitting models. Models and possible study questions were:

1. **Binned, or grouped, counts of predictors in demographic groups.** E.g. count of bed-days where patients aged 1 – 17, or count of patients admitted as elective. This parameterisation would address the question: do increases in bed-days/ attenders, in total, and in given exposure groups, increase the number of incidents reported?
2. **Proportions of predictor variables in demographic groups:** Counts, as described in '1,' divided by a relevant denominator, total bed-days for IP predictors, and total attendances for OP & A&E predictors respectively. Therefore, each demographic category would sum to 1. This parameterisation would address the question: do changes in the distributions of patient casemix factors affect the number of incidents reported?
3. **Quantile values of predictors in demographic groups:** Values for a selection of percentile values in appropriate demographic groups could be used to describe the distribution. E.g. 25th, 50th and 75th percentiles of age based on IP bed-days, without assuming a particular distribution. This is important given most predictor variables are counts. This parameterisation would address the question: do the distributions of bed-day/ attender casemix factors affect the rate of incidents, independent of the level of exposure?

Each of the proposed parameterisations has strengths and limitations in terms of model estimation and interpretation. Approach 1 asks a different question to approaches 2 and 3, namely predicting if numbers of patients in particular demographic groups affect the number of incidents reported.

Approach 2 & 3 describe the distribution of variables, but do not factor the size of organisations in the models. Knowing the percentile ranges or proportions of bed-days in groups does not inform the model as to whether an organisation reports 10 or 10,000 bed-days in a period. A measure of the magnitude of the exposure is also required in these models, and can be achieved either by adding a parameter/fixed effect for total bed-days, or by using an offset variable.

Approach 1 & 3 fit count covariates that can range (in theory) from 0 to $+\infty$. Approach 2 fits proportions as covariates, which are therefore constrained between 0 (no bed-days in a group) and 1 (all bed-days within group). Changes in these covariates will be smaller than in Approach 1, due to scale, and also suffer from multicollinearity problems. 'Multicollinearity' in models describes situations when combinations of model covariates together perfectly predict other model covariates. Models are unable to estimate these coefficients, as parameters could functionally be any combination of the affected covariates. E.g. if a variable has three groups, a, b and c, and sums to 1. When fitting this model, knowing a and b will always fix c, with no freedom to vary. Functionally, a, b & c could take any values summing to one, and therefore have no predictive ability. This can be tackled by dropping one of the multicollinear variables. E.g. fitting a and b but omitting c allows a and b to vary freely. Approach 2 necessitates dropping some covariates for estimation.

Approach 3 presented a high computational burden when preparing the dataset. The extracted IP, OP & A&E files were in excess of 1 GB, 1.5 GB and 0.4 GB respectively, with 52,374,788, IP records, 80,687,829 OP records and 19,136,957 A&E records. Percentile calculations require ordering and logic steps for ties, easily accomplished in statistical software on small datasets. Microsoft SQL Server, the data management solution used for NRLS storage, loading and preparation, does not have an in-built concept of a median as data are not stored in an inherent order. Methods exist to code this in an SQL Server environment, with its strengths in sorting and storage, but statistical software is more naturally suited to these calculations. The burden of sorting such large datasets, particularly in-memory using R, makes this challenging despite efficient median and percentile functions. Initial test estimates, based on test median and percentile calculations, suggested several days of continuous processing time would be required (given that many were required to test suitable percentiles required). Percentiles calculated were: minimum, 5, 10, 25, 50, 75, 90, 95th percentiles and maximum, in all groups for testing. This would not be considered a sustainable model for

regular NRLS/HES analysis, so an optimisation exercise was performed to reduce this, including:

- Using R on a powerful server rather than desktop machine, with 96GB memory and 24 cores.
- Reducing the memory used for data storage, freeing memory for additional computations, by loading only data columns required for iteration, month, organisation, and parameter of interest.
- Using indexed and sorted ('keyed') tables to reduce sort and seek time for both iteration and percentile operations. Although common in database systems, and in other statistical packages such as SAS, this is only implemented in R via the third-party packages, and 'data.table' (Dowle and Srinivasan, 2017) was used for this.
- Parallelising the process to allow several iterations to happen simultaneously. Parallel operation does not reduce run time in a simple multiplicative fashion. Additional steps are required to route operations to different nodes as well as return and assemble outputs. Nodes may compete for resources, and may sometimes wait for each other to complete before proceeding. Despite this, suitable operations can be returned significantly faster when parallelised. R code for parallelised loops was constructed using the 'doParallel' (Revolution Analytics and Weston, 2015) package, and run on 20 cores simultaneously.
- Use of parallelised Basic Linear Algebra Subprograms (BLAS) improves speed. R (and many other systems) rely on a BLAS to perform many calculations, particularly involving matrices. Since R represents many data structures as matrices regardless of how they are presented to the user, many operations can be improved using an optimized, parallel BLAS instead of R's default single-threaded BLAS. Several implementation of this are available, but Microsoft's R Open (MRO) using the Intel Math Kernel (MKL) BLAS is the most accessible for Windows users and regarded as one of the best performing (Microsoft, 2018). Parallel BLAS and explicit parallelisation, as per the last point, may conflict and care must be taken to assess the suitability of either approach. MRO was set to single threaded execution, but explicitly parallelised as described above, benefiting from faster MKL execution, but distributing calculations across nodes without conflict.

The combined reduction in memory use, efficient storage and sorting, and parallel operation reduce run time to six hours. This time will be further reduced once final percentiles were selected.

A final parameterisation was chosen after investigating the approaches described above and with input from academic, clinical and regulatory colleagues. Approach 2 was felt to be the most 'intuitive' version, but resulted in a more challenging model to estimate. Approach 2 shows less variation in the proportion variables, compared to counts or percentiles, and models took longer to converge. This is to be expected, as the possible range of proportions is narrower. This approach was applied to the majority of predictors using proportions of inpatient bed-days, outpatient or A&E attendances as appropriate. No offset was used, as there were three non-equivalent types of exposure in the data, but count predictors were also fitted as weights including total bed-days, OP & A&E attendances. A&E waiting times were fitted as percentiles, with longer waiting times representing increased exposure time for patients when compared to shorter waiting times. The 25th, 50th and 75th percentiles were used, as they represent the central 50% of the distribution, without being overly influenced by high outliers. Other percentiles did not appear to improve fit.

A single interaction effect between the proportion of emergency admissions and proportion of bed-days relating to admission was entered into the model to account for any additional risk from potential rushed emergency admissions.

5.3.2 Time period/seasonality

The modelling data set spans a single financial year. The NHS is assumed to be stressed over winter periods (NHS England, 2018), and seasonal patterns such as national holidays etc. are to be expected. It is likely that, during times of increased pressure, resilience against incidents (and available time for staff to report incidents) may be affected. To represent this, time period or 'seasonality', was considered for inclusion in the model.

Various methods exist to parameterise time period and each was fitted to the models to test for a preferred method:

- Categorical variables for time periods, such as fiscal quarters or months, using 'dummy variables.' The model then assumes one level to be a reference level (e.g. the first quarter or month in the period) with coefficients for the change from reference level for each additional quarter or month.

- Treating month as a numeric value, numbering 1-12. Although easiest to fit, the model would assume a linear trend (on the scale of the link function), that may not be suitable to represent seasonal effects, particularly if they are non-linear.
- Smoothed representation using splines. If we assume that time effects may be 'noisy' but have general non-linear trends fitting a smoothed version of time covariates is logical. Splines are functions that are expressed as piece-wise polynomials, continuous at their 0th, 1st and 2nd derivatives, joined at 'knot points' (Wood, 2017b). Knot points may be directly specified but are commonly placed at percentiles a distribution. Spline functions, having gradients, capture change in a function as well as effect sizes. Splines produce more stable results than fitting traditional polynomials for non-linear trends, and they do not oscillate at the extremes ('Runge's phenomenon') (Wood, 2017d). 'Natural' cubic splines are a sub-class of splines, of degree 3, with a linear constraint (second derivative set to zero) at the extreme knots, and are recommended in this context (Harrell, 2001). See Chapter 6 for more details of splines, expanded for use in Generalized Additive Models.

A natural cubic spline was used to model monthly fluctuation, using 3 knots, placed at the 10th, 50th and 90th percentiles as recommended by Harrell (Harrell, 2001). Various spline functions exist in R for fitting these models. Although estimating the same function, the numerical methods differ in terms of stability and singularity when used in model fitting. The `ns` function using a B-spline basis, rather than truncated power basis used in Harrell's own R functions, was chosen for its numerical stability, their similar performance to Harrell's functions in other studies (Govindarajulu et al., 2009), and better performance on model fitting, based on AIC in NRLS models.

5.3.3 Excluded predictors

Evidence from the literature review suggested some predictors were not significantly associated with NRLS incident reporting rates. These included mortality, patient satisfaction and numbers of staff (Howell et al., 2015). The same study suggested hospital size was not correlated with reporting rate. This may be true in their study with a particular parametrisation, but when NRLS data are modelled as counts rather than rates, for the purposes of this chapter, this conclusion seems unlikely. Hospital size was therefore included in terms of total bed-days, outpatient and A&E attendances.

5.3.4 Summary of constructed dataset

The constructed dataset was examined using univariate descriptive statistics (Table 5.1) prior to modelling to identify potential issues, errors or distributional considerations. Distributions in the constructed dataset were examined through plotting, before any transformations were applied (Figure 5.1 & 5.2).

Variables were tested for pair-wise correlations using Pearson correlation coefficients. Correlation matrices are not presented here as correlations between covariates were generally high but, given that they are predominantly binned counts of the same indicators, this was unsurprising. This did not prevent their use in modelling, but increased the suspicion of multicollinearity affecting models. Major expected correlations were confirmed including age and higher comorbidity scores, and maternity admissions and proportions of females. No decisions about the inclusion of covariates in models was taken on the strength of these summaries, as marginal associations may hide interactions, confounding, and relationships in full conditional models (Harrell, 2001, Sun et al., 1996).

Important relationships between incidents and the total counts in each dataset (IP, OP & AE) are shown in Figure 5.1. All three show a positive correlation, where the number of incident reports increases as the count of IP bed-days, OP or A&E attendances increases. This is consistent with the idea of increasing exposure increasing the risk of incident.

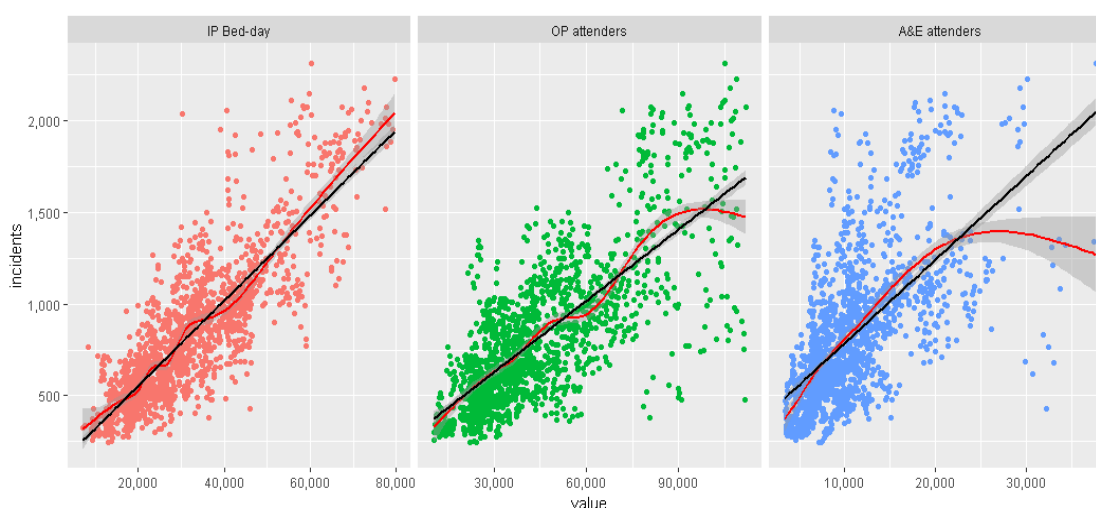


Figure 5.1 Relationship between total incidents and main IP, OP & A&E counts, reported per month, per trust.

Y-axis represents incident reports and x-axis the value of each predictor, in panels from left to right: Inpatient bed-days, outpatient attendances and A&E attendances). Red lines are smoothers fitted using generalized additive models (GAMs), to emphasise overall trends. Black lines are linear model 'smoothers.'

Data item	mean	sd	25th Percentile	Median	75th Percentile
Total Incidents					
Incidents	817.7	386.6	528.0	725.0	1,012.3
Total Bed-days					
Bed-days	31,351.6	14,073.7	20,582.8	28,687.5	39,887.0
Inpatient Age (bed-days)					
Age <1	1,836.3	1,129.2	1,094.0	1,649.0	2,312.3
Age 1-17	1,256.5	1,053.7	642.0	998.0	1,467.3
Age 18-29	2,196.3	1,228.5	1,366.8	1,882.0	2,753.3
Age 30 - 49	4,320.3	2,472.3	2,631.5	3,686.0	5,344.0
Age 50-69	7,324.3	3,769.6	4,456.5	6,293.5	9,656.3
Age 70-84	9,073.5	3,769.0	6,193.0	8,370.5	11,680.3
Age >84	5,154.9	2,022.9	3,665.8	4,821.0	6,340.3
Sex (bed-days)					
Males	14,221.6	6,782.5	8,926.0	12,711.0	17,905.5
Females	17,128.3	7,366.7	11,713.8	15,658.0	21,491.8
Co-morbidity (bed-days)					
Charlson score <1	16,110.4	7,792.1	10,419.5	14,184.0	20,232.8
Charlson score 1-4	4,068.8	1,764.3	2,742.0	3,668.0	5,185.3
Charlson score >4	11,172.4	4,949.0	7,320.0	10,319.5	14,124.0
Admission Method (bed-days)					
Elective	7,826.2	4,890.3	4,314.5	6,376.5	10,227.3
Non-elective	19,584.8	8,239.4	13,172.3	18,489.5	24,410.0
Maternity/birth	3,102.1	1,859.2	1,830.5	2,863.0	4,025.3
Transfer	832.8	918.4	256.0	493.0	1,064.0
Admitting Specialty (bed-days)					
Surgical	10,215.1	5,199.1	6,267.0	9,179.5	13,152.8
Bed-days that were day of admission					
Admission Day	9,802.0	4,704.2	6,461.5	8,657.0	12,214.5
Total Outpatient Attenders					
OP attenders	44,594.5	22,318.5	27,351.8	38,566.5	56,434.3
Outpatient Age (Attenders)					
Age <1	466.4	321.1	248.8	391.5	593.3
Age 1-17	4,193.9	2,569.6	2,481.8	3,589.0	5,184.3
Age 18-29	4,839.2	2,943.6	2,845.0	3,930.0	6,228.3
Age 30 - 49	10,128.7	6,264.9	5,618.8	8,488.5	12,415.0
Age 50-69	13,140.2	6,654.0	8,065.3	11,726.0	16,352.5
Age 70-84	9,482.3	4,243.2	6,202.5	8,712.5	12,168.3
Age >84	2,168.4	1,006.9	1,420.3	1,947.5	2,762.0
A&E					
A&E Attenders	10,658.7	5,353.4	6,810.8	9,506.5	12,599.3
A&E wait time					
25th percentile	89.9	25.6	72.0	89.0	109.0
50th percentile	151.8	29.7	133.0	153.0	174.0
75th percentile	217.6	25.0	204.0	220.0	232.0
A&E					
Ambulance arrivals	2,759.7	1,248.2	1,912.5	2,541.5	3,373.5

Table 5.1 Summary statistics for combined NRLS-HES modelling dataset

The majority of predictors, being counts in their raw form, are characteristically skewed, where a small number of high value outliers stretched the distributions out. The percentile waiting times showed different distributions to other variables, and this is to be expected, as they are essentially a sample of a distribution of percentiles, and are normally distributed. The 75th percentile waits breaks this pattern with a noticeable peak at ~240 minutes, or 4 hours i.e. the national waiting times target for A&E. Few values were observed above this line, although not entirely absent. This suggests an artefact of recording. It is not possible to know whether organisations censor data at this point, or whether the drop genuinely represents few patients waiting longer than 4 hours, but the abrupt change is suspicious (Figure 5.2). Possible reasons for this include: ending monitoring once a patient has 'breached,' deliberately reformatting the waiting times to 240 minutes that exceed this, or default values extracted from A&E administration software. This truncated distribution supports the use of percentiles, as they will be internally consistent within organisations, conditional on their reporting behaviour.

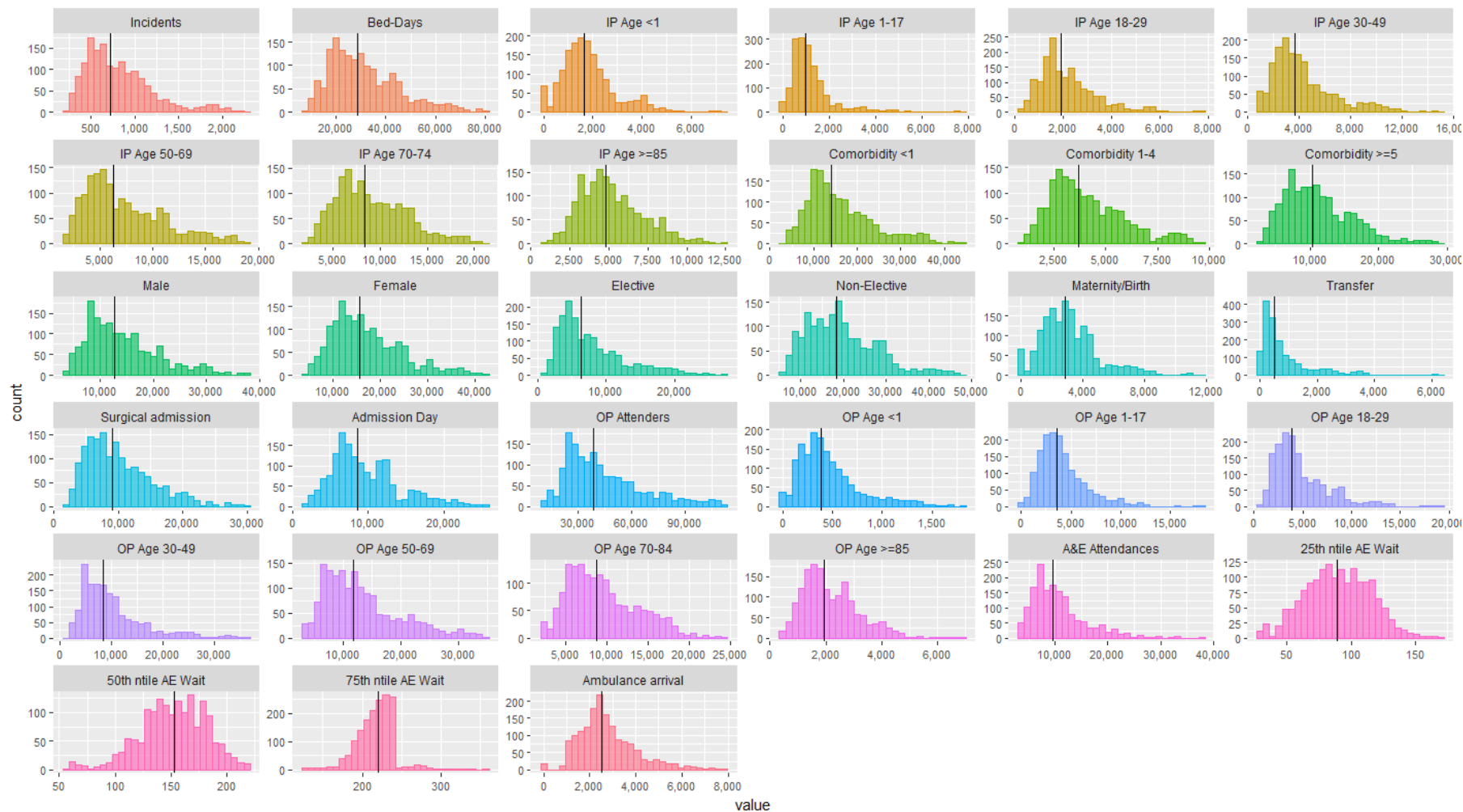


Figure 5.2 Distributions of predictor variables from NRLS-HES combined dataset

Black lines represent the median value for each variable.

5.4 Single-level model fitting and output

The constructed dataset was used to iteratively build models and test their effectiveness for predicting incident reporting.

5.4.1 Methods

To examine if initial count models were suitable, a Poisson GLM was first fitted to the dataset to assess the overdispersion before further model were examined. Parameter estimate confidence intervals were affected by overdispersion, so bootstrap estimation was also performed with 1000 samples to better estimated the error distribution for comparison.

Poisson GLMs were then refitted using quasi-likelihood, described in Chapter 4. These models do not specify a full distribution from the exponential family, as GLMs normally do, but merely specify a mean-variance relationship. The models are then estimated using estimating equations rather than traditional MLE methods. The estimation method and lack of MLE makes them ill-suited to comparisons with other MLE methods. Although we cannot calculate likelihood ratio tests or AIC, these models allow an assessment of a multiplicative overdispersion factor. This factor both verifies the overdispersion tests of GLMs and allows for variance scaling. Parameter estimates of models are the same as those in the naïve Poisson GLM, however the scaled standard errors can be used to better approximation of parameter significance.

Models were further developed mixture models, based on MLE methods, that included adjustments for scale. Mixture models, including the NB1, NB2 and GP models were fitted. Parameter estimates in these models are conditional on the scaled deviance, in contrast with the quasipoisson, allowing assessment of whether parameters are affected by OD, or simply reflected in the error structure.

The various models discussed in Chapter 4 were fitted using the statistical computing environment R (R Core Team, 2016), primarily using the base R function: `glm`. It is common practice for R users to cite the authors of specialist packages, as many of them are supported through research grants with their progress measured by citations. This thesis will explicitly acknowledge such packages when referring to them for the first time, with work in this section

using functions from the `tidyverse` suite of packages including `dplyr` (Wickham et al., 2017) and `tidyr` (Wickham and Henry, 2017) for data handling, plot functions from the `ggplot2` (Wickham, 2009) and `gridExtra` (Auguie, 2016) packages, `Anova` & `Boot` functions from `car` (Fox and Weisberg, 2011), `mse` and `mae` functions from `ModelMetrics` (Hunt, 2016) and model output tidying and format functions from `broom` (Robinson, 2017).

The standard R `glm` function may be used to fit quasi-likelihood models by specifying a `family` argument of `'quasipoisson'` rather than `'poisson'`. `glm` is unable to fit the mixture models, and the `glm.nb` function from the `MASS` packages (Venables and Ripley, 2002) was used instead. This function initially fits a Poisson GLM then iterates between ML estimation of the scale parameter θ , and refitting the model using NB2 with the estimated θ . This models the NB2 parameterisation, where variance is quadratic to the mean, but it does not allow the fitting of NB1. The recently developed modelling package `glmmTMB` (Magnusson et al., 2017) was used, allowing NB1 and GP families, but it's performance was 'sense-checked' by fitting both the Poisson and NB2 models using `glmmTMB` and the standard `glm`/`glm.nb` functions. No differences in parameter estimates over 1×10^{-5} were observed, with most estimates identical between functions, and `glmmTMB` were considered robust for use.

Model diagnostics were examined by checking that models attained their default convergence values, commonly no change in the objective functions of greater than 1×10^{-8} over several iterations. Model parameter estimates and AIC are presented in this section and prediction error examined in section 5.6.

5.4.2 Results

Poisson GLM result are presented in table 5.2 and figures 5.3 and 5.4. The naïve Poisson model reported almost all parameters as significant at 95%. Bootstrapped estimates inflated the standard error and confidence intervals accordingly (Figure 5.3) as expected. This reduced the significance of some parameters, notably for inpatient bed-days aged >84, all comorbidity score groups, non-elective admissions, sex, admission data total OP attenders and A&E waiting times. When tested for overdispersion, as described in section 4.2.2, the Poisson model was highly overdispersed with an estimated dispersion ratio of ~38.

Poisson Model					
term	estimate	Wald		Bootstrap (1000)	
		L95CI	U95CI	L95CI	U95CI
Intercept	5.036	4.797	5.274	3.434	6.554
Teaching Hospital status	-0.017	-0.023	-0.010	-0.062	0.026
Total IP Beddays	0.508	0.498	0.518	0.449	0.566
<i>Proportion IP bed-days by age group</i>					
<1 year	-2.459	-2.821	-2.097	-4.774	-0.291
1-17 years	4.818	4.561	5.076	3.234	6.421
18-29 years	3.497	3.128	3.866	1.418	5.596
50-69 years	2.786	2.529	3.042	1.230	4.347
70-84 years	1.515	1.326	1.704	0.375	2.617
>84 years	0.541	0.341	0.741	-0.617	1.685
<i>Proportion IP bed-days by comorbidity groups</i>					
Charlson score <1	-0.920	-1.071	-0.768	-1.822	0.029
Charlson score >4	-0.075	-0.234	0.083	-1.028	0.957
<i>Proportion IP bed-days by sex</i>					
IP bed-days males	0.149	0.016	0.282	-0.765	1.104
<i>Proportion IP bed-days by admission type</i>					
Non-Elective	0.224	0.055	0.393	-0.910	1.383
Maternity/birth	1.679	1.468	1.889	0.391	3.037
Transfer	-0.072	-0.179	0.034	-0.739	0.637
<i>Proportion of IP bed-days by specialty</i>					
Surgical Admission	0.122	0.085	0.159	-0.199	0.459
<i>Proportion of IP bed-day that are admission day</i>					
Admission Day	0.680	0.404	0.957	-1.113	2.471
Total OP Attenders	0.028	0.020	0.037	-0.028	0.087
<i>Proportion of OP Attendances by age group</i>					
<1 year	2.973	2.506	3.440	-0.138	5.987
1-17 years	0.050	-0.078	0.178	-0.670	0.798
18-29 years	1.517	1.317	1.717	0.384	2.673
50-69 years	0.504	0.327	0.681	-0.554	1.568
70-84 years	1.717	1.529	1.904	0.615	2.850
>84 years	-2.053	-2.458	-1.648	-4.491	0.341
<i>Seasonality Spline basis</i>					
1	0.000	-0.010	0.010	-0.066	0.070
2	-0.012	-0.022	-0.003	-0.072	0.047
3	0.104	0.089	0.119	0.004	0.203
4	0.173	0.164	0.183	0.110	0.240
Total AE attendances	0.126	0.119	0.133	0.077	0.172
<i>Percentiles of A&E Waiting Time</i>					
25th percentile	-0.024	-0.037	-0.012	-0.101	0.059
50th percentile	0.087	0.069	0.105	-0.033	0.201
75th percentile	-0.036	-0.044	-0.027	-0.095	0.026
<i>Proportion of A&E Arrival Type</i>					
Ambulance	0.417	0.391	0.443	0.259	0.558
<i>Interaction of Admission Day and non-elective</i>					
Admission Day * Non-Elective	-3.769	-4.198	-3.339	-6.539	-1.006

Table 5.2 Model coefficients for NRLS-HES Poisson GLM with Wald and bootstrapped confidence intervals

Green cells represent statistically significant values ($\alpha = 0.05$). 'L95CI' is the lower bound of the 95% confidence interval, and 'U95CI' the upper.

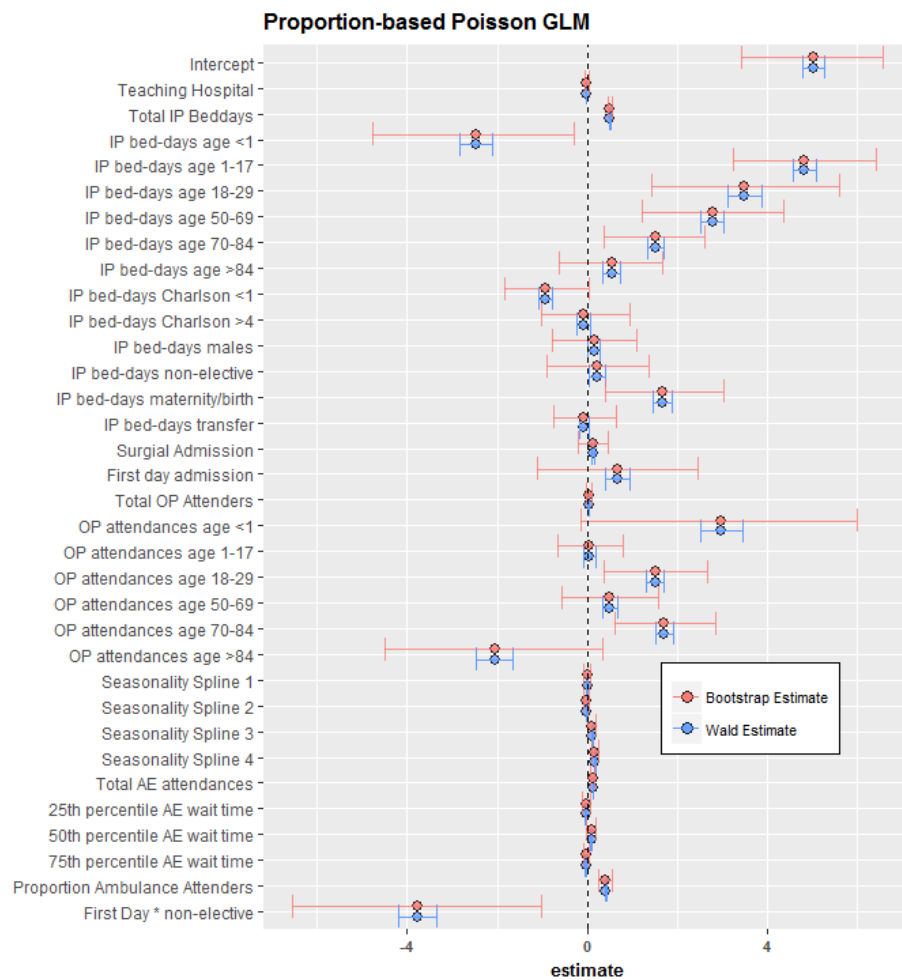


Figure 5.3 Comparison of estimated model coefficients and 95CIs for NRLS-HES Poisson regression models

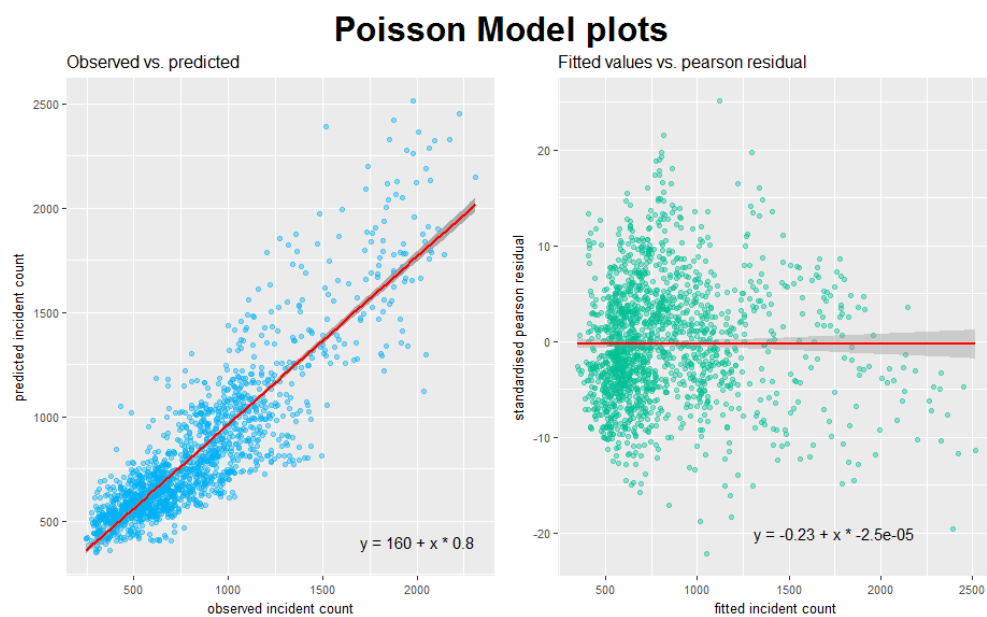


Figure 5.4 Model diagnostic plots for NRLS-HES Poisson regression models

Diagnostic plots showed some patterns, suggesting most points were of low observed and predicted values. Larger negative residuals were seen at higher fitted values suggesting that the model over-predicts at higher observed values. The overall line of best fit is approximately zero, but this may be misleading with the pattern described above.

Goodness of fit was tested using chi-squared tests on the residual deviance versus null deviance, suggested models were significantly under-specified. When combined with the overdispersion test, this suggests overdispersion is dominating the model. The model is likely under-specified, or parameterisation is not specific enough to represent the variance in the data, particularly given the aggregation.

Parameter estimates in the quasipoisson model remained unchanged compared to the Poisson model but errors were adjusted using the scale factor. Comparisons of tables 5.2 & 5.3 show quasipoisson and bootstrapped estimates of standard error to be broadly comparable. As coefficients remained the same as the Poisson model, model diagnostic charts for quasipoisson in Figure 5.5 are similar to the Poisson model in figure 5.4, showing similar under-estimation at higher values, but the residuals are better scaled.

Outputs for mixture models are presented in table 5.3 and figure 5.5. Mixture models, NB1, NB2 and GP all improved upon the Poisson and GP models. Significance of parameter estimates was broadly stable across QP, NB1 and GP models, with NB2 showing more variation. Comparisons of QP and mixture models between their bootstrapped and Wald intervals showed small differences, but similar significance levels, indicating that overdispersion was not adversely shrinking standard errors in these models and the scaling had been successful. Therefore, Wald intervals were used rather than bootstrapped intervals for ease of calculation. Each scaled model showed similar parameter significance to the bootstrapped Poisson model, and suggested that the scale factors are taking reasonable account of the overdispersion in the underlying distribution. Residual plots and predictive value plots still demonstrated the pattern observed in the Poisson model, under-predicting the highest predicted values (figure 5.5). The NB2 model showed some differences to others for parameter significance, and a notably lower intercept, but also showed the strongest trend in residuals suggesting a bias. This stands in contrast to the NB2 model showing the lowest AIC value.

Quasi-Poisson and Mixture Models												
term	Quasi-Poisson			Negative Binomial 1			Negative Binomial 2			Generalized Poisson		
	estimate	L95CI	U95CI	estimate	L95CI	U95CI	estimate	L95CI	U95CI	estimate	L95CI	U95CI
Intercept	5.036	3.548	6.523	5.177	3.727	6.627	4.580	3.105	6.055	5.260	3.818	6.703
Teaching Hospital status	-0.017	-0.059	0.026	-0.009	-0.050	0.033	-0.049	-0.094	-0.005	-0.003	-0.044	0.038
Total IP Beddays	0.508	0.448	0.568	0.520	0.461	0.578	0.584	0.520	0.648	0.527	0.469	0.586
Proportion IP bed-days by age group												
<1 year	-2.459	-4.714	-0.203	-2.354	-4.553	-0.156	-0.899	-3.131	1.334	-2.261	-4.448	-0.074
1-17 years	4.818	3.211	6.426	4.680	3.113	6.247	5.261	3.643	6.879	4.607	3.045	6.168
18-29 years	3.497	1.195	5.799	3.374	1.148	5.600	4.633	2.414	6.853	3.270	1.070	5.471
50-69 years	2.786	1.187	4.384	2.639	1.105	4.172	3.354	1.822	4.887	2.534	1.027	4.041
70-84 years	1.515	0.335	2.694	1.384	0.237	2.531	2.193	1.049	3.337	1.320	0.181	2.459
>84 years	0.541	-0.703	1.785	0.516	-0.681	1.714	1.348	0.164	2.532	0.507	-0.674	1.689
Proportion IP bed-days by comorbidity groups												
Charlson score <1	-0.920	-1.863	0.024	-0.801	-1.722	0.121	-0.214	-1.133	0.705	-0.701	-1.619	0.217
Charlson score >4	-0.075	-1.064	0.913	0.091	-0.879	1.061	0.677	-0.287	1.640	0.211	-0.756	1.179
Proportion IP bed-days by sex												
IP bed-days males	0.149	-0.680	0.978	0.082	-0.727	0.891	0.176	-0.621	0.972	0.017	-0.788	0.823
Proportion IP bed-days by admission type												
Non-Elective	0.224	-0.830	1.278	-0.049	-1.069	0.971	-0.584	-1.627	0.459	-0.258	-1.271	0.755
Maternity/birth	1.679	0.364	2.993	1.660	0.385	2.935	1.236	-0.018	2.489	1.638	0.376	2.901
Transfer	-0.072	-0.737	0.593	-0.086	-0.726	0.555	-0.229	-0.867	0.408	-0.100	-0.733	0.532
Proportion of IP bed-days by specialty												
Surgical Admission	0.122	-0.109	0.354	0.095	-0.143	0.333	0.321	0.091	0.552	0.079	-0.166	0.324
Proportion of IP bed-days that are admission day												
Admission Day	0.680	-1.043	2.404	0.261	-1.413	1.935	-0.551	-2.318	1.215	-0.052	-1.721	1.618
Total OP Attenders	0.028	-0.024	0.081	0.017	-0.034	0.069	0.002	-0.053	0.057	0.010	-0.042	0.062
Proportion of OP Attendances by age group												
<1 year	2.973	0.062	5.885	2.342	-0.504	5.188	1.776	-0.968	4.519	1.986	-0.858	4.829
1-17 years	0.050	-0.747	0.847	0.130	-0.633	0.893	-0.302	-1.056	0.451	0.180	-0.569	0.929
18-29 years	1.517	0.271	2.762	1.507	0.326	2.688	1.492	0.305	2.679	1.549	0.399	2.700
50-69 years	0.504	-0.599	1.607	0.491	-0.565	1.548	-0.099	-1.110	0.912	0.507	-0.528	1.543
70-84 years	1.717	0.548	2.886	1.760	0.634	2.886	1.645	0.520	2.769	1.796	0.685	2.907
>84 years	-2.053	-4.576	0.471	-2.169	-4.604	0.266	-2.884	-5.282	-0.485	-2.231	-4.638	0.175
Seasonality Spline basis												
1	0.000	-0.064	0.064	-0.004	-0.066	0.058	0.005	-0.059	0.070	-0.007	-0.069	0.054
2	-0.012	-0.070	0.045	-0.006	-0.062	0.050	-0.023	-0.081	0.035	-0.001	-0.057	0.055
3	0.104	0.009	0.198	0.107	0.015	0.199	0.130	0.035	0.225	0.111	0.020	0.203
4	0.173	0.113	0.233	0.168	0.110	0.226	0.194	0.134	0.254	0.163	0.105	0.221
Total AE attendances	0.126	0.084	0.169	0.115	0.075	0.156	0.135	0.087	0.182	0.108	0.068	0.149
Percentiles of A&E Waiting Time												
25th percentile	-0.024	-0.103	0.055	-0.015	-0.092	0.062	-0.040	-0.119	0.038	-0.009	-0.086	0.068
50th percentile	0.087	-0.025	0.200	0.082	-0.027	0.191	0.077	-0.036	0.190	0.079	-0.029	0.188
75th percentile	-0.036	-0.091	0.019	-0.037	-0.091	0.017	-0.018	-0.075	0.039	-0.038	-0.093	0.016
Proportion of A&E Arrival Type												
Ambulance	0.417	0.254	0.581	0.370	0.217	0.523	0.459	0.302	0.616	0.341	0.192	0.490
Interaction of Admission Day and non-elective												
Admission Day * Non-Elective	-3.769	-6.448	-1.089	-2.856	-5.458	-0.254	-2.184	-4.887	0.520	-2.196	-4.790	0.398
Scale Parameters												
	38.86			36.60			21.14			38.70		
AIC												
	-			21,172			21,146			21,162		

Table 5.3 Model coefficients for NRLS-HES Quasipoisson, Negative Binomial and Generalized Poisson GLMs

Green cells represent statistically significant values ($\alpha = 0.05$). 'L95CI' is the lower bound of the 95% confidence interval, and 'U95CI' the upper.

One-year Quasi-Poisson and Mixture Models

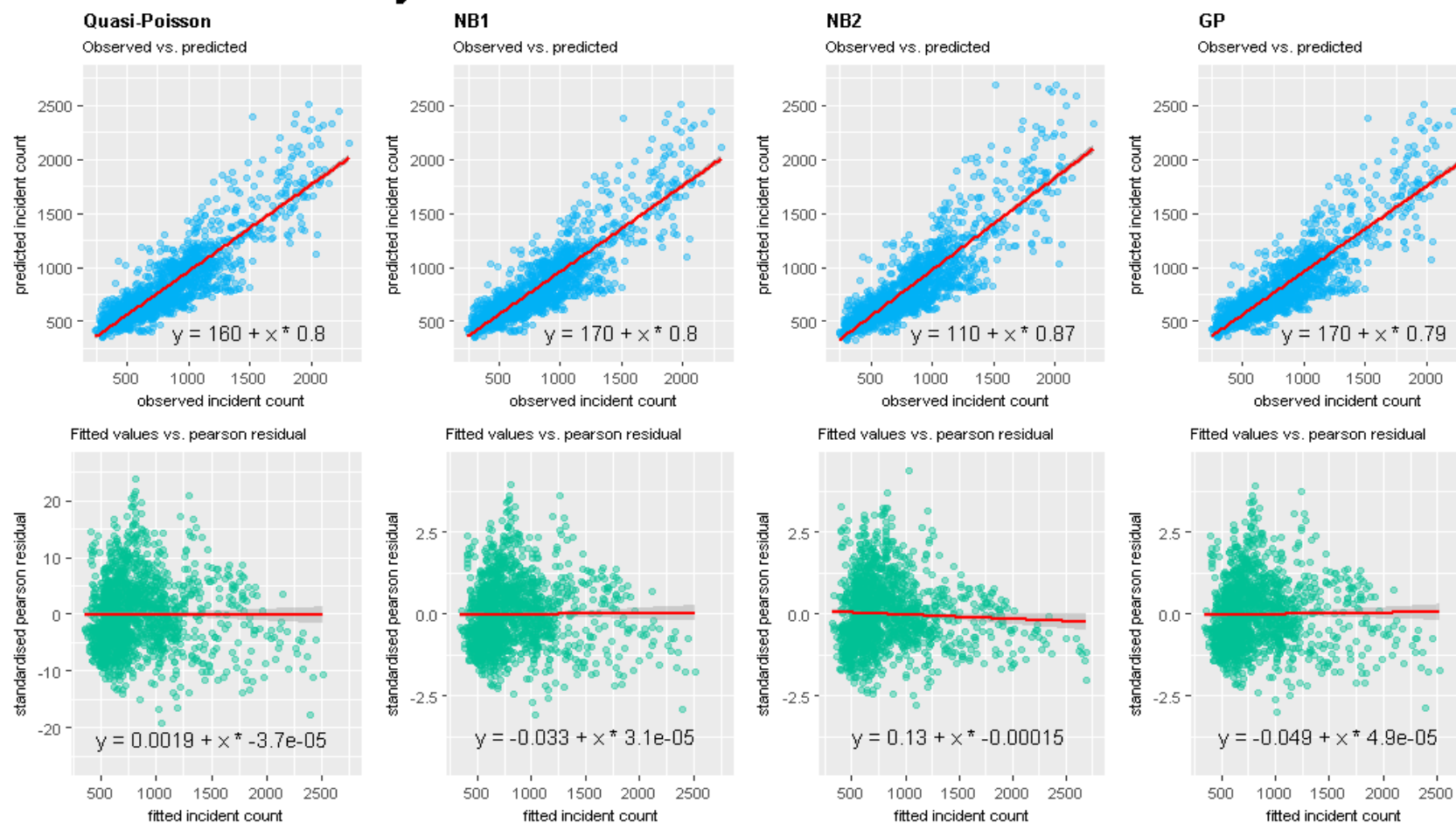


Figure 5.5 Model diagnostic plots for NRLS-HES scaled Poisson models: Quasipoisson, Negative Binomial and Gen. Poisson

Columns represent each model, with the top row showing plots of predicted vs. observed, and the bottom row the fitted vs. standardized residuals

5.4.3 Discussion

The results suggest that Poisson models, although a useful start, are not appropriate choice for this dataset as they grossly underestimate the error due to overdispersion. Each scaling method showed great improvements on the Poisson model alone, regardless of the method. Mixture models appeared more appropriate for the dataset than quasipoisson models and suggested there was a latent structure in the dataset. All models notably over-predicted for larger input values, suggesting the relationship to size is non-linear and/or there is missing information related to the submitting organisation that might better describe this size element.

NB1 and quasipoisson models show similar significance for covariates, despite parameter estimates in NB1 being adjusted. This is to be expected, as they use similar variance-scaling methods (see Chapter 4). In this application, the NB1 could be considered preferable, both for updating parameter estimates given the overdispersion, and for having a fully specified distributional form, allowing the use of LRT and AIC.

The GP model showed good performance for the scaling, and judged by the AIC alone, improved upon the NB1 model by losing the less information. The GP scale factor is similar to the NB1 and quasipoisson models, with parameter significance similar to NB1.

NB2 shows different parameter estimates to NB1, with a notably different scale factor, due to its quadratic scale term. NB2 has a noticeable smaller intercept estimate and was the only model to estimate the proportion of bed-days for patients admitted to surgical specialities as significant. AIC comparisons of NB1 and NB2 support NB2 as a better fit, losing less information than NB1, but NB2 showed a stronger trend than any other model on residual plots. This suggested the proportional overdispersion assumption may be wrong, with the simple multiplicative scaling of NB1 placing too much emphasis on high counts when scaling the variance. The NB2 parameterization gives more weight to low expected counts when scaling for overdispersion and supports an assumption that organisational size/organisational specific factors are important in driving the overdispersion and model fit. It could also be suggested that the NB2 model showed reduced variance at the expense of increased bias.

All mixture models showed sufficient improvements to be considered better than the naïve Poisson model, but the undescribed differences between large and small organisation are the most important missing feature of the models examined so far.

5.5 Mixed/random-effects models

The models presented above have not explicitly reflected the known repeated-measures structure in the data and applied simpler multiplicative scaling terms. This known clustered structure in the datasets violates the independence and homoscedasticity of variance assumptions of typical regression models. A more explicit modelling approach is discussed in the following sections to account for this structure. The models fitted in the previous section were therefore re-fitted with an additional random-intercept for each hospital submitting data to NRLS.

5.5.1 Methods

Initial multilevel structures were fitted as Poisson GEEs to test for differences in residuals compared to the naïve Poisson model. GEE models more adequately scaled the variance of data but were abandoned due to their lack of full distributional assumption and an MLE. Results for GEEs are not presented here, as they could not be as easily compared to MLE methods.

The standard R package for fitting multilevel models, in a frequentist framework, is `lme4` (Bates et al., 2015) and the function `glmer` was used for this. Despite being the standard function in R for these models, `lme4` does not allow the fitting of NB1, GP models, and quasi-likelihood models (which are incompatible with the mixed modelling paradigm). As in section 5.4, `glmmTMB` was used to fit these models NB1 and GP models. Poisson and NB2 models fitted with `lme4` were compared with `glmmTMB` fits. Again, these fits showed minimal differences in parameter estimates, with identical AICs observed, so `glmmTMB` was considered to be robust.

Multilevel model estimation is computationally challenging, with parameterisation and scales of predictors having major implications convergence. The models presented below have, in some cases given scale warnings, or have displayed convergence warnings during their

development. Various testing procedures were performed to determine if convergence warnings were problematic, or if they simply reflected the maximum number of iterations for the non-linear optimisers used during fitting. These tests have been conducted by consulting the literature surrounding the `lme4` package, following published diagnostic processes (Bolker, 2017), and in direct correspondence with the package's primary authors, via the Cross-Validated statistics web forum.

Changes and tuning options used for convergence issues, compared to default `glmer` settings include:

- Trying a variety of optimizers, including a custom optimizer recommended for speed and convergence (Bolker, 2017). Diagnostic procedures suggest that, if most optimizers converge with similar results, but the default does not, the fit may be sound, but limited by a given optimizer.
- Increasing the number of iterations of the optimiser beyond the default
- Centring and scaling variables where possible, as discussed in section 4.3
- Double checking gradient and Hessian calculations using more exact (but computationally costly) derivative functions.

Solutions were achieved mainly through the centring and scaling of predictors (avoiding large gradients for the optimizer to estimate), increasing the number of iterations to 2×10^8 , and for the NB2 model in particular, using the 'bobyqa' optimizer (Powell, 2009) from `lme4`. Little to no difference in gradients was observed when double-checked. `glmmTMB` did not raise any convergence errors, require additional iterations, or need specific optimizer settings beyond default.

The significance of the random-intercept term was not directly tested for the models presented below as the `lme4` package does not provide standard errors for the random-effect variances. The package's authors (Bates & Bolker) suggest that the sampling distribution of the variance of random-effects is usually asymmetric, and asymptotic normal standard errors are biased (Bolker, 2018). The preferred option in `lme4` is to use the likelihood profiles to construct a confidence interval (Bates et al., 2015), and this has been presented below for the random-effect variances. Wald intervals and profiled intervals were similar for the fixed effects in the models, so Wald intervals were reported.

5.5.2 Results

Model coefficients are displayed in table 5.4, with diagnostic plots in figure 5.6. The Poisson random-intercept model retained more significant factors than the scaled random-intercept models, mirroring the single-level models. Parameters estimates varied from model-to-model, but larger differences were observed between the Poisson model and others, notably in the global intercept term.

Estimates of the random-intercept variances were similar across models, with estimated variance highest in the Poisson random-intercept model, although the confidence intervals were narrower, presumably due to overdispersion. Overdispersion was reduced in the Poisson random-intercept model when compared to the single-level Poisson model, but not entirely eliminated. When tested with the chi-squared test described in section 4.5, overdispersion was still present, with a dispersion ratio of ~ 7.3 . This is a notable reduction from the dispersion of ~ 38 , but still substantially overdispersed. The same tests on the NB1, NB2 and GP random-intercept models, were not significant, indicating that scaling removed the residual overdispersion, with dispersion ratios ~ 1 .

Residual plots (figure 5.6) for all random-intercept models show points symmetrically distributed above and below the mean, although the size-based clusters were more obvious. The majority of data points were part of the large cluster of lower predicted values, and a second smaller cluster at higher predicted values. The NB2 model appears to show a tighter group of residuals around zero for the higher group when compared with the other models. The spread of the residuals in these plots was wider (between -6 and 6) for NB1, NB2 and GP models compared the single level models (between -3 and 3).

The random-intercept models showed improved fit in all model families compared with single-level models. The NB1, NB2 & GP random-intercept models further reduced the AIC when compared to the Poisson random-intercept model, but residual plots suggested more spurious/outlying residuals in all fits (figure 5.6). In AIC comparisons, the random-intercept NB1 model performed best, with GP and NB2 better than the Poisson model.

Random-Intercept Models												
term	Poisson			Negative Binomial 1			Negative Binomial 2			Generalized Poisson		
	estimate	L95CI	U95CI	estimate	L95CI	U95CI	estimate	L95CI	U95CI	estimate	L95CI	U95CI
Intercept	5.343	4.742	5.943	4.927	3.434	6.421	4.971	3.470	6.471	4.948	3.453	6.444
Teaching Hospital status	0.236	0.132	0.339	0.162	0.043	0.280	0.127	0.008	0.246	0.162	0.043	0.281
Total IP Beddays	0.198	0.161	0.236	0.262	0.177	0.348	0.288	0.194	0.382	0.262	0.176	0.348
Proportion IP bed-days by age group												
<1 year	0.820	0.146	1.493	0.602	-1.233	2.438	0.833	-1.033	2.700	0.566	-1.271	2.402
1-17 years	1.286	0.624	1.948	2.016	0.344	3.689	2.339	0.624	4.054	2.038	0.364	3.711
18-29 years	1.660	1.014	2.306	2.078	0.293	3.863	1.798	-0.014	3.610	2.041	0.254	3.828
50-69 years	1.979	1.571	2.386	1.953	0.843	3.064	2.031	0.915	3.147	1.920	0.809	3.032
70-84 years	1.409	1.024	1.794	1.276	0.254	2.297	1.455	0.434	2.477	1.248	0.225	2.270
>84 years	1.508	1.118	1.899	1.326	0.301	2.350	1.392	0.379	2.405	1.295	0.270	2.319
Proportion IP bed-days by comorbidity groups												
Charlson score <1	-0.501	-0.790	-0.211	-0.485	-1.283	0.313	-0.319	-1.108	0.470	-0.504	-1.302	0.295
Charlson score >4	-0.424	-0.730	-0.118	-0.243	-1.088	0.602	-0.139	-0.968	0.691	-0.238	-1.084	0.608
Proportion IP bed-days by sex												
IP bed-days males	-0.067	-0.282	0.148	-0.070	-0.669	0.529	0.039	-0.554	0.633	-0.086	-0.686	0.513
Proportion IP bed-days by admission type												
Non-Elective	-0.186	-0.565	0.194	-0.343	-1.372	0.686	-0.371	-1.410	0.669	-0.327	-1.358	0.704
Maternity/birth	0.482	0.040	0.923	0.249	-0.935	1.434	0.156	-1.024	1.336	0.263	-0.923	1.450
Transfer	0.363	0.037	0.689	0.259	-0.582	1.099	0.092	-0.753	0.937	0.269	-0.573	1.111
Proportion of IP bed-days by specialty												
Surgical Admission	-0.298	-0.426	-0.170	-0.223	-0.555	0.108	-0.290	-0.639	0.059	-0.219	-0.551	0.113
Proportion of IP bed-days that are admission day												
Admission Day	-1.095	-1.712	-0.477	-0.889	-2.579	0.801	-0.908	-2.645	0.828	-0.864	-2.558	0.830
Total OP Attenders	0.131	0.101	0.160	0.131	0.056	0.207	0.151	0.067	0.235	0.131	0.056	0.207
Proportion of OP Attendances by age group												
<1 year	-0.443	-1.571	0.685	0.727	-2.258	3.711	1.058	-1.827	3.942	0.762	-2.221	3.744
1-17 years	1.017	0.468	1.565	1.523	0.335	2.711	1.565	0.390	2.740	1.497	0.309	2.685
18-29 years	-0.877	-1.665	-0.089	0.712	-1.111	2.535	0.483	-1.329	2.295	0.708	-1.116	2.533
50-69 years	1.864	1.288	2.441	2.291	0.921	3.660	1.726	0.373	3.080	2.306	0.936	3.676
70-84 years	-1.537	-2.104	-0.969	-0.648	-1.995	0.699	-0.011	-1.361	1.340	-0.646	-1.994	0.702
>84 years	1.230	0.135	2.325	1.156	-1.637	3.949	-0.783	-3.553	1.987	1.155	-1.640	3.951
Seasonality Spline basis												
1	0.022	0.008	0.035	0.013	-0.024	0.050	0.015	-0.024	0.054	0.013	-0.024	0.050
2	-0.062	-0.075	-0.048	-0.056	-0.093	-0.020	-0.055	-0.093	-0.017	-0.056	-0.093	-0.019
3	0.052	0.034	0.071	0.053	0.002	0.104	0.076	0.022	0.130	0.054	0.003	0.105
4	0.039	0.023	0.055	0.048	0.004	0.092	0.067	0.022	0.113	0.046	0.002	0.090
Total AE attendances	0.360	0.330	0.390	0.335	0.266	0.405	0.319	0.243	0.395	0.335	0.265	0.405
Percentiles of A&E Waiting Time												
25th percentile	-0.033	-0.061	-0.005	-0.045	-0.120	0.030	-0.021	-0.101	0.058	-0.047	-0.122	0.028
50th percentile	0.094	0.061	0.128	0.097	0.006	0.189	0.084	-0.012	0.181	0.100	0.008	0.191
75th percentile	0.029	0.018	0.041	0.028	-0.005	0.061	0.020	-0.015	0.056	0.028	-0.005	0.061
Proportion of A&E Arrival Type												
Ambulance	0.394	0.253	0.535	0.464	0.144	0.784	0.464	0.152	0.776	0.462	0.141	0.782
Interaction of Admission Day and non-elective												
Admission Day * Non-Elective	2.127	1.187	3.066	1.453	-1.122	4.028	0.936	-1.697	3.570	1.435	-1.146	4.016
Variances												
Organisation code	0.243	0.217	0.276	0.223	0.197	0.255	0.220	0.195	0.249	0.224	0.198	0.256
Scale Parameters												
	1.00			7.06			104.59			8.10		
AIC												
	26,375			19,249			19,281			19,259		

Table 5.4 Model coefficients for NRLS-HES random-intercept Poisson and mixture GLMMs

Green cells represent statistically significant values ($\alpha = 0.05$). 'L95CI' is the lower bound of the 95% confidence interval, and 'U95CI' the upper.

One-year Random-Intercept Models

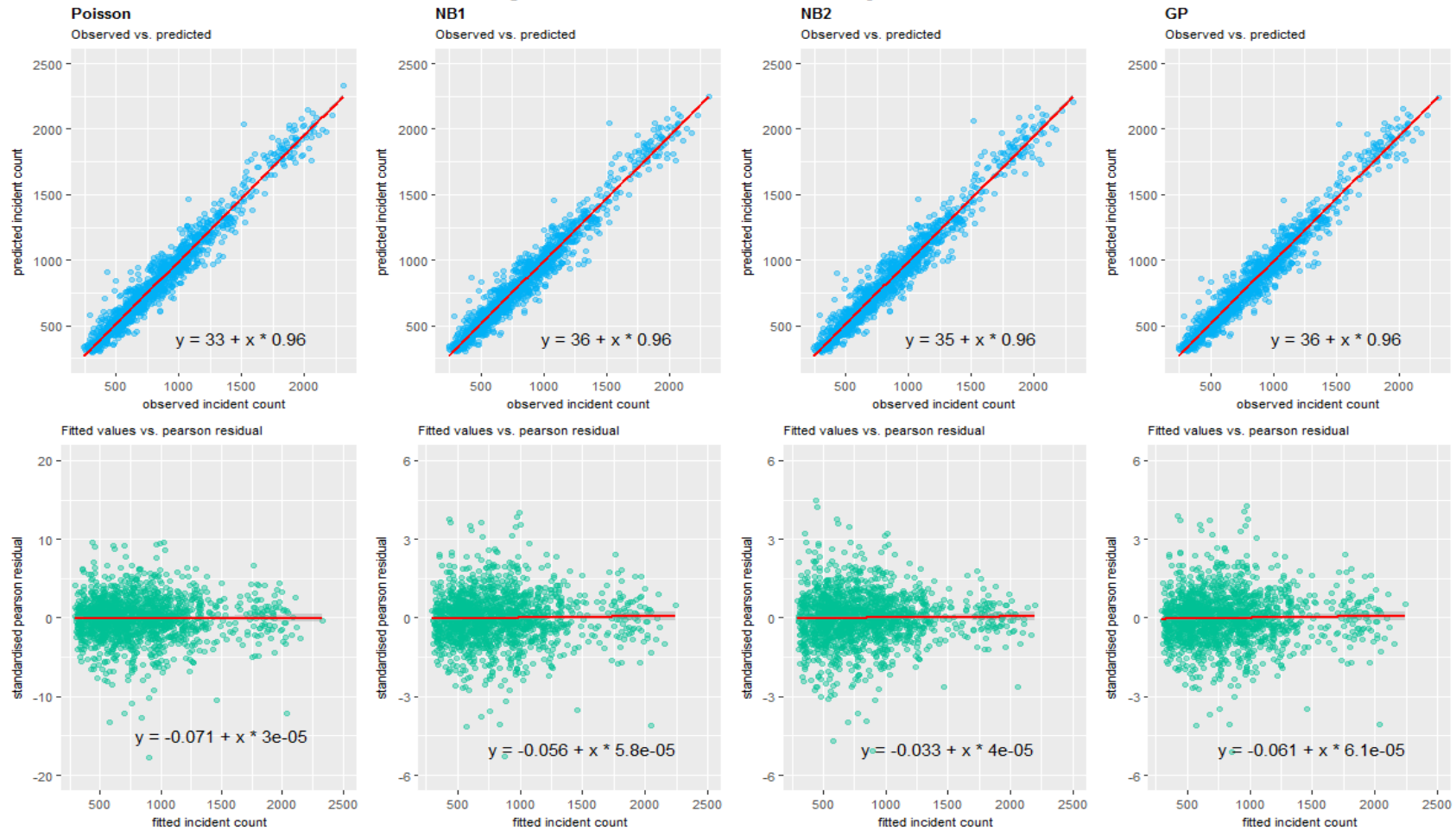


Figure 5.6 Model diagnostic plots for NRLS-HES random-intercept models

Lines represent linear fits, with equations expressing the fit as simple linear model.

5.5.3 Discussion

The significance of parameters was broadly consistent between models despite different scaling and variance structures. This consistency for age, total in-patient bed-days, seasonality and proportion arriving by ambulance suggests they are important predictors across all units. The random-intercept Poisson model was still affected by overdispersion and, as such, the significance of parameters was overstated.

The clusters in residuals and predictions highlight that, once clustering is reflected in the models, there are two distinct reporting behaviours: low and high. This is most likely to be due to differences between organisations, as seasonality and size were reflected in the fixed-effects. Introducing the random-intercept term appears to have reduced the bias seen in all single-level models, and is consistent with the known data structure. The tighter residual pattern in NB2 may suggest that the NB2 model better predicts reporting at larger organisations. The wider spread of residuals in the random-intercept models should be viewed in light of the tighter cluster and scale of the plots. This suggests that random-intercepts better model the data as a whole, but generate more extreme residuals, and suggest the data are more volatile than single-level models may convey.

Based purely on AIC, the NB1 random-intercept model gives the best overall performance, but NB2's adjustment for larger organisations is a useful property. As these models have only been compared using the training dataset, it is not possible to easily examine the extent of overfitting. Bootstrapping or cross-validation are possible on these types of models, but given there are only 12 data points for most organisations, creating a hold-out sample that reflects the clusters/stratification is challenging and simple bootstrapped approaches were dismissed due to this concern.

5.6 Model selection

Having prepared models with differing sets of assumption, the performance of each model required testing to establish generalisability. Models were tested on new data, outside the training dataset, to identify the models that best fitted the relationships between outcome and predictors without overfitting.

5.6.1 Methods

A testing dataset was prepared using the same data processing rules as those used for the training set, described earlier in this chapter, but applied to data from the 2016/17 fiscal year for both NRLS and HES. Data were centred and scaled according to the means and variances in the training set rather than testing sets. This was chosen to correspond to the same scaling used when the model was trained and prevent additional bias from mismatched scaling. No new organisation codes were created for 2016/17 in the underlying HES data, allowing easy comparison without coding new organisations as zero random-effects estimates (the global average).

Model comparisons were made across all single-level and random-intercept models using MAE, RMSE and AIC (described section 4.5). Both MAE and RMSE were expressed as both as raw values and as percentages of the average reporting rate, using the median number of incident reports as a denominator. Although model comparisons on training data can be clear with raw error values, percentages allow additional intuitive comparisons of the scale of error. For a baseline error comparison, we can also calculate the national average number of incident reports per bed-day (≈ 0.0261) as it is a model. This can then be used to predict an expected number per trust without any further casemix-adjustment.

A comparison of significance of the statistical significance of model predictors was also made by tabulating and examining 95% confidence intervals based on each models' variance calculations. Although examined in sections 5.4 and 5.5, this collation aids further interpretation.

5.6.2 Results

Training and testing MAEs of all models are presented in table 5.5. The training MAE of this baseline 'model' (national average value per bed-day) was 154.94, and testing MAE 201.538. All models, even without random-intercepts, improved upon these figures.

Introducing random-intercepts, for all model classes, reduced both prediction error and AIC. When using testing prediction error, the NB2 random-intercept model showed the best performance of the mixed models, but Poisson random-intercept showed the lowest error

Model	Training (2015/16) Prediction error				Testing (2016/17) Prediction error				AIC	Degrees of Freedom
	MAE	% of median (725)	RMSE	% of median (725)	MAE	% of median (745)	RMSE	% of median (745)		
NB2 Random Intercept	55.34	7.6%	78.61	9.6%	101.68	14.0%	149.77	17.9%	19268	36
NB1 Random Intercept	55.29	7.6%	77.87	9.5%	102.73	14.2%	150.66	18.1%	19237	36
GP Random Intercept	55.38	7.6%	78.03	9.6%	103.07	14.2%	151.11	18.1%	19247	36
Poisson Random Intercept	55.02	7.6%	77.26	9.5%	104.85	14.5%	152.25	18.2%	26366	35
Poisson	138.79	19.1%	196.94	24.1%	143.46	19.8%	196.94	23.6%	74394	34
Quasi	138.79	19.1%	196.94	24.1%	143.46	19.8%	196.94	23.6%	-	34
NB1	138.79	19.1%	181.79	22.3%	143.78	19.8%	197.11	23.6%	21158	35
GP	138.96	19.2%	182.14	22.3%	144.16	19.9%	197.61	23.7%	21148	35
NB2	141.72	19.5%	194.78	23.8%	144.96	20.0%	209.10	25.1%	21132	35

Table 5.5 Prediction error and AIC summary NRLS-HES models

Rows represent models fitted, ordered by training MAE in descending order. Columns are grouped in relation to training and testing datasets.

Significance of model covariates under all models									
term	Single-level					Random Intercept			
	Pois +	Quasi	NB1	NB2	GEN	Pois	NB1	NB2	GEN
Intercept	+	+	+	+	+	+	+	+	+
Teaching Hospital Status									
Teaching Hospital				-		+	+	+	+
IP Bed Days									
Total IP bed-days	+	+	+	+	+	+	+	+	+
Inpatient Age (bed-days)									
Age <1	-	-	-		-	+			
Age 1-17	+	+	+	+	+	+	+	+	+
Age 18-29	+	+	+	+	+	+	+		+
Age 50-69	+	+	+	+	+	+	+	+	+
Age 70-74	+	+	+	+	+	+	+	+	+
Age >84				+		+	+	+	+
Co-morbidity (bed-days)									
Charlson score <1					-	-			
Charlson score >4						-			
Sex (bed-days)									
Males									
Admission Method (bed-days)									
Non-elective									
Maternity/birth	+	+	+		+	+			
Transfer						+			
Admission Type									
Surgical						-			
Proportion of IP bed-days that are admission day									
Admission Day						-			
Outpatients									
OP Attenders						+	+	+	+
Outpatient Age (Attenders)									
Age <1		+							
Age 1-17						+	+	+	+
Age 18-29	+	+	+	+	+	-			
Age 50-69						+	+	+	+
Age 70-74	+	+	+	+	+	-			
Age >69						+			
Time-period Spline									
Spline portion 1						+			
Spline portion 2						-	-	-	-
Spline portion 3	+	+	+	+	+	+	+	+	+
Spline portion 4	+	+	+	+	+	+	+	+	+
A&E									
A&E Attenders	+	+	+	+	+	+	+	+	+
A&E Waiting Times									
25th percentile AE wait time						-			
50th percentile AE wait time						+	+		+
75th percentile AE wait time						+			
A&E Attendance type									
Ambulance	+	+	+	+	+	+	+	+	+
Interaction of Admission Day and non-elective									
Admission Day * Non-Elective	-	-	-	-	-	+			

Table 5.6 Comparison of predictors significance for NRLS-HES models

Significance based on profiled 95% confidence intervals. 'Pois+' refers to single-level Poisson model with bootstrapped confidence intervals. Green '+' represents significance and positive model coefficient, Red '-' represents significance and negative model coefficient.

rates in training. This stands in contrast to the AIC figures that suggest the NB1 to be the 'better' model in terms of information loss, compared to the NB2 or Poisson random-intercept models. The Poisson, NB1, NB2 and GP random-intercept were all very close in terms of their prediction error, varying by only half a percent of the average error reporting rate.

Parameter estimates across all models were also compared in Table 5.6. Global intercepts were significant in all models, as was total IP bed days, inpatient age, seasonality, outpatient age and proportion arriving by ambulance. In both single and random-intercept models, NB2 models showed some differences in significance of parameters compared to other models.

5.6.3 Discussion

As all models improved upon the baseline error rate, for average incident reports per bed-day, we can be certain that exposure/casemix predicts incident reporting. The consistency of predictors, including IP bed days, inpatient age, seasonality, outpatient age and proportion arriving by ambulance, indicate they are major predictors of the exposure risk at hospitals. This makes them useful casemix variables for predicting incident reporting at a hospital.

The best testing performance was given by NB2 models. Given the residual plots of section 5.5.2 showed tighter clusters at larger organisations, this effect may be responsible for the better performance in 2016/17. The NB2 is the only model where 'Inpatient Age 18-29,' and the 50th percentile A&E waiting time are not significant. It is not possible to determine whether this is a function of a bias/variance trade-off or if these predictors may be more susceptible to noise induced by aggregation.

AIC can be used to judge between fixed effects-only models or between the random-effects models but, with the NB1 random-intercept model performing best on this criterion, AIC is misleading in this case. The comparison of the suitability of NB2 with and without random-intercept is interesting, as NB2 without random-intercept performed worst on training and testing error, but with a random-intercept, it gave the best testing error and one of the best training error rates. This suggests that the NB2 variance structure is the most suitable for these data once the correlations of the repeated measures are explicitly modelled by the random-intercept terms.

When comparing random-effect models with fixed-effects models, considering the fixed-effect model nested within the random-effect model, we are implicitly comparing a likelihood on the boundary of the parameter space (i.e. random-effect distribution is constrained to be ≥ 0 , and it cannot be normally distributed about zero to satisfy the asymptotic chi-squared distribution with one degree of freedom). AIC is biased in this application (Greven and Kneib, 2010), so random-effect models were considered justified, from a data structure perspective with the known clustered structure of the data. This was further supported by the profiled confidence intervals of the random-effects models, the drastic improvement in prediction error, and the different weighting for high predicted values in the single-level NB2 model compared with other single-level models.

Information theoretic approaches such as AIC, are broadly useful, but suffer two flaws in this application. Firstly, that AIC is based on the log-likelihood of the model and therefore related to the likelihood ratio test (LRT). Model comparisons based on LRT assume that models are nested, and this is not the case if comparing across classes like Poisson and NB2. There are, however, differing opinions with some authors claiming nesting is required due to the way they derived the AIC (Ripley, 2004), and others suggesting (as a measure of approximate Kullback-Leibler information) this is not the case provided they are on the same dataset (Anderson and Burnham, 2006, Burnham and Anderson, 2004).

The similar prediction error rates versus differing AIC values suggest that there is residual overdispersion affecting the AIC. The scaling in the mixture models appears to introduce more error into the process, although the differences are small. The Poisson random-intercept model could be described as the most accurate, but a degree of overdispersion remains. The residual overdispersion is best scaled with the NB2 random-intercept, but the un-biased nature of the Poisson distribution, even with overdispersion, may be sufficient to model incident reporting in this case.

5.7 Extending models to longer time periods

The models described above were extended across longer periods of time, to see if predictors of reporting remained consistent with more events. A 5-year period was used for fiscal years 2010/11 – 2015/16.

Data preparation was more challenging, as the earlier years contained more spurious entries than 2015/16, with missing data, and a number of unlikely values (such as a single incident report). Merging of organisations (and their influence on random-effects) became more of an issue that hindered estimation. 119 data points were excluded (1.5%) from 137 organisations, a notably higher rate than the 0.2% for the single year model, detailed in section 5.2.3.1.

The models were adapted to include more time dimensions, extending the techniques in sections 5.3 – 5.5, to cover multiple years. Step changes in reporting were considered likely, both within and between organisations each year. ‘Between’ variation, or national-level change, was modelled with both categorical factors for fiscal years and fiscal quarters as well as with spline functions. Fiscal year categories were used with the assumption that organisations have annual plans and quality improvement programmes that alter each fiscal year, and changes are not necessarily linear. Non-linear fluctuations over the entire time period was also modelled as a natural cubic spline with varying numbers of knots, with reduction in AIC of the full model used to distinguish improvement. This fitted the data well, but model selection always favoured higher numbers of knots and risked overfitting. Fiscal year terms were rendered non-significant when spline functions were used, but this is again a symptom of potential overfitting, multicollinearity or correlation.

Within organisation variance over time was modelled by including a random-slope with the random-intercept term. Random-slopes allow random-intercepts variances to change in response to another parameter (Gelman and Hill, 2006b). The random-intercept term remained as ‘Organisation’ with a slope allowed for fiscal year. Random-slopes are commonly applied over continuous variables, but can also be applied to categorical variables such as fiscal year, modelling the change in the intercept in relation to the reference, in this case period 1 (2010/11), see figure 5.7 for an example with a small number of organisations. Fiscal year was also included in the model as a fixed effect as recommended in the `lme4` support documentation (Bates et al., 2015). This forces the model to fit fixed effects for the global changes per fiscal year, allowing the random slope to estimate cluster-specific changes in variances only.

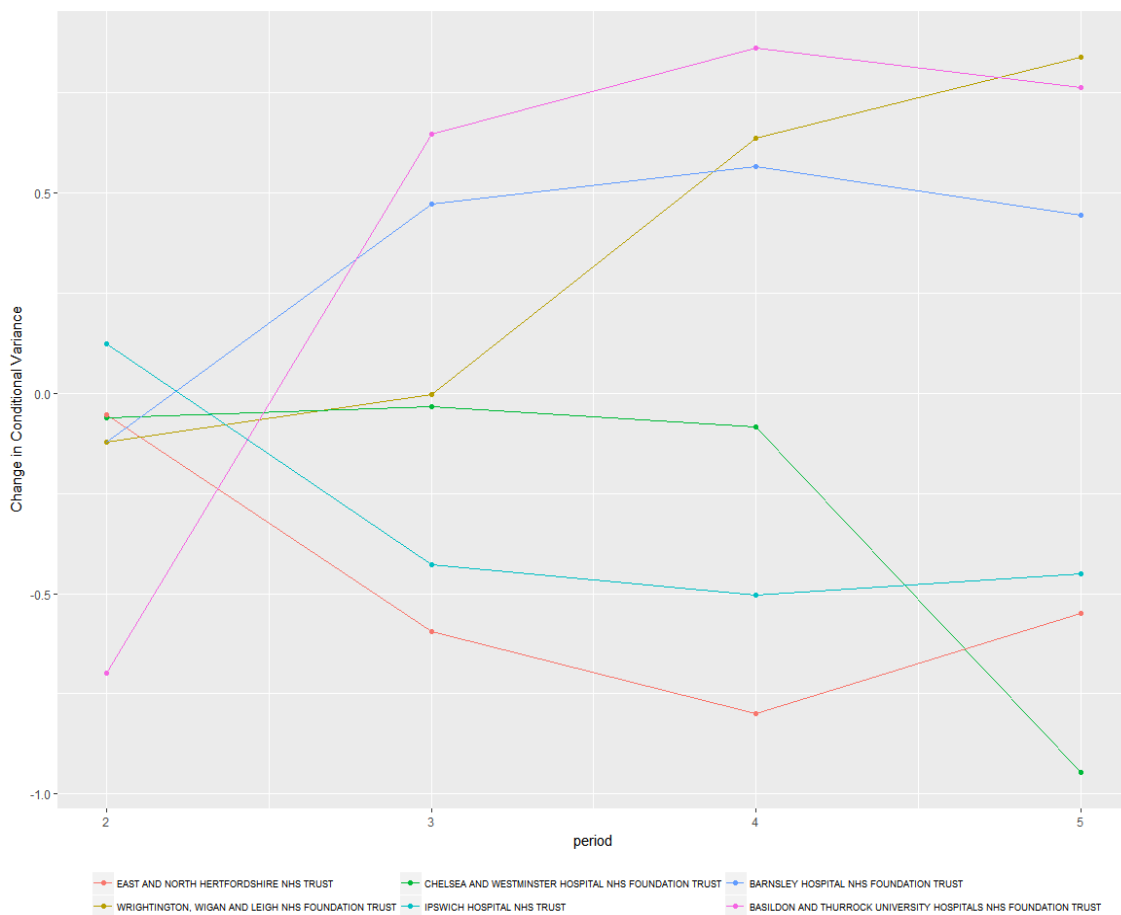


Figure 5.7 : Example of variation in slopes of random-intercepts for six outlier organisations

‘Period’ represents years 2, 3, 4 and 5 of a 5-year model, vs. random-intercept variance in period 1.

It is unclear how much benefit a random-intercept and slope model brought to the dataset, other than confirming the heterogeneity of random-intercepts over time. The effects on overdispersion, and the differing variance parameters over the modelling period make global assumptions challenging. Reducing models to simpler random-intercept only models did not represent this heterogeneity well, and models struggled to converge. Figure 5.7 illustrates the change in random slopes compared to initial estimates of the random-intercept in the first time period. Basildon & Thurrock NHS Foundation trust (pink trend) demonstrates how random-effects variance can both increase and decrease over the modelling period.

Outputs of these models are not presented, or further examined, as model fit and convergence were questionable.

5.8 Discussion and conclusions

This chapter has addressed aim 3 of the thesis: to examine what statistical modelling processes are appropriate for these data. The chapter established a theoretical model of incident reporting, separating 'exposure' risk from 'culture' of reporting. It has discussed the creation of a count dataset using NRLS incident reporting data and hospital-based exposure variables from HES. It has tested appropriate methods for analysing count data in a regression framework and applied these modelling approaches to the dataset, assessing the predictive ability based on exposure variables. It established that using case-mix adjustment on exposure variables gives better predictions of incident reporting rates than using simple averages, and identified how best to adjust for overdispersion due to clustering and aggregation.

Model outputs suggest that, whilst exposure can be used to predict incident reporting, poor parameterisation increases overdispersion. The choice of predictors, based on proportions of IP bed-days, OP & A&E attendances, better reflects the exposure risk compared to other methods of quantifying hospital activity such as discharges. Models have demonstrated that increasing exposure increases the number of incident reports. This message may be intuitive without modelling, but modelling provides empirical estimates that will be taken forward in chapter 8 and be applied to tools for monitoring and regulation. The mean-centring of the models allows us to predict with all predictors set to zero to derive the average number of reports. This fitting method suggested average expected incident reports per month of ~760 for single level models, and ~740 when using random-intercept models.

The substantial overdispersion encountered in these models is evidence of both the poor specification of the fixed effects and random error. Some of this overdispersion was due to a lack of independence in the repeated measurements at organisations, creating a correlated structure. Organisation-specific random-intercepts, factored a degree of this overdispersion, adjusting for local factors that cause an organisation to the average reporting behaviour. Dispersion ratios for Poisson models decreased when random-intercepts were introduced. Similar effects occur in the scaled models, but the additional scaling renders the dispersion ratio useless as an assessment of this. This is in line with other adverse event analyses that have also required random-effects structures to account for clustering (Baines et al., 2013, Landrigan et al., 2010).

NB2 models are commonly used solutions for overdispersion and in this case, when combined with a random-intercept term to account for the clustering, the weighting assumptions used for scale factor are a good fit. NB2 gives more weight to small means when scaling the variance (Ver Hoef and Boveng, 2007), in a sense, assuming proportionally more ‘noise’ at the smaller site. This fits with the idea that the aggregation of the constructed dataset leads to overdispersion, where a single incident report is a larger proportion of the total incidents for smaller sites. The scaling of NB1 and GP random-intercept models was also an improvement on prediction error in Poisson model. All perform similarly, to scale the residual variance, and are all improvements on unscaled models. This may, however, be a quirk of the test set used and a logical next step is testing on multiple sets to verify if this holds, as cross-validation is complicated by the need for cluster stratification with only 12 measurements at each site.

Whilst variance scaling improved single level models in the case of NB1, NB2 & GP, they had little effect on prediction error when compared the Poisson model, increasing it only slightly. Variance scaling models, when applied to the single-level regressions, increased the apparent performance in AIC terms. Given that models were aimed at prediction, the scaling of residual overdispersion is less important than the adjustment for the known clustering. Scaling could be seen as a further refinement, once the clustering was modelled. This suggest that the models are performing their task of adjusting the effects of ‘exposure’, and the effects of ‘culture’ are being absorbed by the random-effects.

A useful next step would have been to treat organisations as fixed effects and examine their model coefficients as an explanatory model. The large number of clusters, combined with current predictors, would have increased the chance of overfitting, as the number of predictors would have been too large for the degrees of freedom of the model (Harrell et al., 1996). A random-effects approach, whilst it does not directly quantify the culture effects with a parameter estimate (as fixed-effects do), allows adjustment and marginal prediction (see Chapter 8).

Residual overdispersion may also be due to further multilevel structures within the data, e.g. departments within hospitals (Pham et al., 2010). This was not investigated in these models, as the NRLS descriptions of departments/locations (see chapter 2) is not directly comparable with the HES data dictionary concept of treatment specialties (NHS Digital, 2017d), and exposure could not be reasonably attributed.

Models were aimed at prediction, but examination of parameter estimates, given the overdispersion, was an important step to allow models to be understood by future users. Parameters that were significant across all random-intercept models such as age, total bed-days, OP and A&E attendances can be considered major predictors and aid understanding by NHS staff.

The proposed models chosen from this chapter are therefore the NB2 random-intercept model and the Poisson-random-intercept model despite, residual overdispersion, is also a useful comparison. These models can be used to predict in two ways: 'conditionally' where the random-effects are included in the prediction, or 'marginally' where the random-effects are used to construct the model and improve fixed-effect estimates, but not used in prediction. The conditional method would allow organisations to monitor their incident reporting rates based on their casemix, with a compensation for the culture of their organisation. The marginal model can be used to predict expected average rates for organisations and assess which organisations deviate from this (see Chapter 8 for more details).

Application of the modelling method to an extended 5-year period showed questionable model fit. It did, however, provide evidence for a key recommendation of this chapter: models should be constructed within fiscal years, and if required, presented as stratified single-year models whose predictions can be combined to cover several years, e.g. three models for a three-fiscal year period. This allows the non-linear effects of national and organisational priorities to be well adjusted within years, rather than poorly averaged over many. This will reduce the heterogeneity in both fixed effects and random-effects. Year-on-year change in incident reporting rates will be challenging to define using such a metric, due to changes in organisational and nation priorities.

Potential limitations of this approach include poor predictive ability from aggregated predictors, lack of measures of organisational culture. Aggregated predictors were unavoidable in this modelling approach, as NRLS does not have information on exposure (only on events) and the different focus of the dataset (NRLS may not necessarily relate to patients). Critical Care units are high intensity areas of hospital activity and high-risk patients, suggesting that a day in critical care unit would represent additional risk compared with a day in a normal

hospital bed. Critical care data were not available as part of UHB's HES subscription at the time these models were constructed but would be an area for future model development.

The next chapter will continue to address aim 3 by examining if more intricate modelling methods, such as models of smoothed covariates, non-linear tree-based models, and latent-variable models, better predict NRLS incident reporting. The models in this chapter will be used in chapter 8 to develop methods for regulators and hospitals to analyse incident reporting, addressing aim 4 of the project.

Chapter 6 Non-parametric modelling

techniques: GAMs, Trees & Neural Networks

6.1 Introductions

The models presented in Chapter 5 applied generalized linear and generalized linear mixed models to predict incident reporting at NHS hospitals, using an aggregated dataset. Random-intercepts allowed the models to reflect repeated measurements at organisations, but residual overdispersion remained in the final Poisson models. This was further adjusted using a negative binomial mixed model. Overdispersion was a major issue in Chapter 5, and one potential source is poorly characterised predictors. This chapter continues to address aim 3, by testing models that may better reflect non-linear relationships between predictors and the response, or better address correlations between predictors. Generalized Additive Models (GAMs) use smooth function of predictors rather than the predictors themselves, tree-based regressions (with ‘bagging’/‘boosting’) allow decorrelation and estimation of unknown functions that can be averaged across models, and neural networks can be used to learn complex patterns from data. These techniques have strengths over those presented in Chapter 5, and this chapter examines their use on NRLS, but their increased complexity may present barriers for NHS or regulatory staff without specialist skills.

6.2 Generalized additive models (GAMs)

In the modelling techniques presented in Chapter 5, the relationships between predictors and response have been assumed to be linear on the scale of the link function. This approximation is reasonable in some cases, but it may be inappropriate when relationships are non-linear, such as bimodal data. A common first step is to discretise/categorise data, referred to as ‘factors’ in R, e.g. groups for bed-days of 1, 2-5 or greater than 5. This loses information and is not continuous but allows local means to be used for each group rather than averaging across the full range (Figure 6.1). One method to map non-linear relationships is to use polynomial expansions of model terms, such as expanding x to a polynomial of degree 3 in a linear model:

$$Y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Polynomials may be sufficient to represent relationships in some cases, but they tend to oscillate wildly in some circumstances (Figure 6.1), so piecewise-structures such as spline functions (as introduced in Chapter 5) may be used to construct better approximations over the range of a predictor.

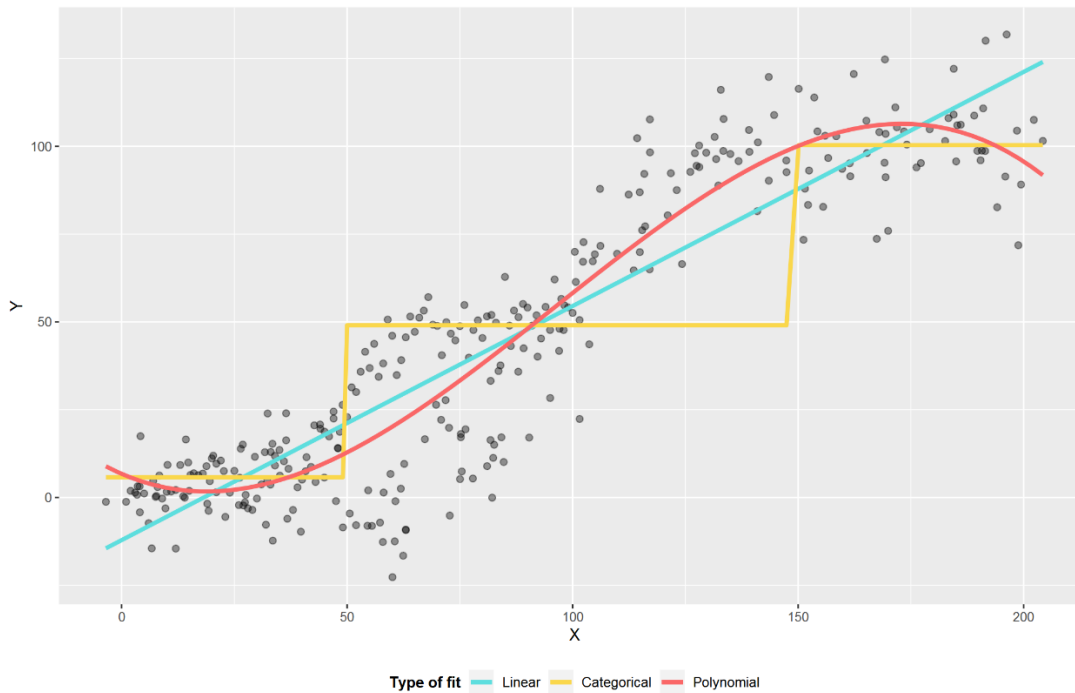


Figure 6.1 Approximation examples for non-linear relationships

Simulated data with a sigmoidal relationship, fitted with a linear model, categorical (<50, >49 & <125, >124) binned variables, and orthogonal polynomial smoothers of degree 3.

6.2.1 Structure of GAMs

A Generalized Additive Model (Hastie and Tibshirani, 1986) follows from the structure of a Generalized Linear Model (GLM) presented in Chapter 4, by adding additional ‘smoothing’ structures to the GLM. In the case of incident reporting models, ‘noisy’ predictors, non-linear relationships or clusters of data points, may be better represented by smooth terms that reflect these relationships. They may, in turn, reduce the overdispersion encountered in GLM/GLMM models.

To achieve this, we can replace the linear predictor of the GLM: $\sum \beta_j X_j$, with the sum of smooth functions $\sum s_j(X_j)$. The $s_j(\cdot)$ ’s are unspecified functions that are estimated from the data in various ways depending the type of smoother used (Hastie and Tibshirani, 1986, Wood, 2017c).

GAMs are strictly additive models of smooth functions, but also allow the inclusion of parametric (non-smoothed) predictors. Smooth functions may represent individual predictors or combinations of predictors.

An example GAM could have the following structure (Wood, 2017c):

$$g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots$$

Where:

- $\mu_i \equiv \mathbb{E}(Y_i)$, the expectation of Y
- $Y_i \sim EF(\mu_i, \phi_i)$, Y_i a response variable, distributed according to an exponential family distribution with mean μ_i and shape parameter ϕ
- A_i is a row of the model matrix for any strictly parametric model components with θ the corresponding parameter vector
- f_i are smooth functions of the covariates x_k

In a similar fashion to GLMs, GAMs may be estimated using maximum likelihood techniques, but they add two further complications: representing smoothed terms in a manner that can be estimated, and choosing how smooth these terms should be in relation to our data (Wood, 2010). Ideally, smooth terms should be estimated from the data.

Hastie and Tibshirani's work on GAMs focused on the use of scatterplot smoothers, initially using just local scoring techniques. Since early publications, their technique (and own R code) has offered both smoothing splines and locally estimated regression smoothers. An alternative framework developed by Wood (2017b), using various reduced rank smooth functions based on regression splines or similar basis functions, has gained wide acceptance. It has become to be considered an essential tool in R and is now included in the default R installation for all users. The two paradigms contrast in their approach to smoothing and penalization, and both have been used to fit GAM models to incident data in this chapter.

To describe GAMs further, a summary of the common smoothers, methods for penalization, estimation, and effective degrees of freedom follows below.

6.2.2 Smoothers

Any smooth function could be used to fit GAMs in principle, but in application, there are three popular types: locally estimated smoothing splines, smoothing splines and regression splines.

6.2.2.1 Locally Estimated Smoothing Splines

Locally Estimated Smoothing Splines ('loess' or 'lowess' in some statistical packages) compare local fit in a region of a plot to the mean. Loess is widely used in statistical software such as Stata and R, and is usually implemented based on methods by Cleveland et al. (1992). Loess smoothers are related to nearest-neighbour techniques and can be understood as an improvement on a 'moving average'. A moving average smooths the data based on a window of its neighbours, and the smoothness can be altered by changing the width of the window (how many points to average over). This approach is simple but problematic at the boundaries of the data, whether the window is truncated. Loess smoothers retain the moving window ('span') concept but use weighted polynomial fits, scoring local points related to the average with data outside the window weighted as zero. The weighting gives a smooth fit, and reduces the influence of extreme points/outliers, but can be computationally intensive as it involves a regression for every data point in a smooth (Cleveland, 1979, Larsen, 2018).

6.2.2.2 Smoothing splines

Splines are classes of smooth functions named after draftsman's splines, thin bendable measures held in place by weights, that are used to draw curves. Smoothing splines (Reinsch, 1967) are piece-wise polynomials that join at knot points, and have a knot point at every datum. They do not work in local regions as loess does, but rather minimise a penalized sum of squares across the whole range of the data, that can be described as:

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int (s''(x))^2 dx,$$

Where red represents the residual sum of the squares, and blue represents a penalty term that penalizes 'wiggleness' (the term used in the literature). If a curve is 'wiggly,' it's second derivatives (the 'slopes of the slopes,' or the change in the rate of change in the function) will be large. If data form a straight line, the second derivative will be zero. A penalty term, λ , acts as a multiplier to the integrated square of the second derivative, and allows the smoothness to be altered. Smoothness is therefore controlled by changing the penalty, not by adjusting the number of knots or in relation to a span (Figure 6.2).

Natural cubic splines can be considered the ideal functions for minimising the penalized sum of the squares (Wood, 2017d) described above, but this comes at a high computational cost due to the knots at every data point.

6.2.2.3 Regression Splines (and associated techniques)

An alternative to the costly approaches for loess and smoothing splines is to estimate reduced rank versions of smoothers, such as the natural cubic splines, provided these splines adequately describe the relationships in the data. This reduction in rank comes by using only sufficient knots to estimate the function, rather than at every datum. This approach is attractive from smoothness and computational perspectives. If a good reduced rank function exists, loess or smoothing splines could be considered ‘wasteful’ by comparison.

Regression splines can be expressed as a set of basis functions that do not depend on Y . Basis functions span the regions between knot points but are continuous up to and including second derivatives at knots. A regression spline can be written as:

$$f(x) = \sum_{i=1}^k b_i(x)\beta_i$$

Where b_i is the i th basis function for $k-1$ knots, and β_i its corresponding coefficient (Wood, 2017d).

For use in GAMs, our aim is to select sufficient knots that the smooth relationship reflects the trend in the data, but is not overfitting or ‘wiggly’. Selection of the number of knots is challenging using likelihood ratio tests or AIC, as a spline of $k-1$ knots is not necessarily nested within a spline of k knots across the same region. However, the ‘right’ number of knots is considered less critical for smoothing when we are combining this with a penalty term (Figure 6.2).

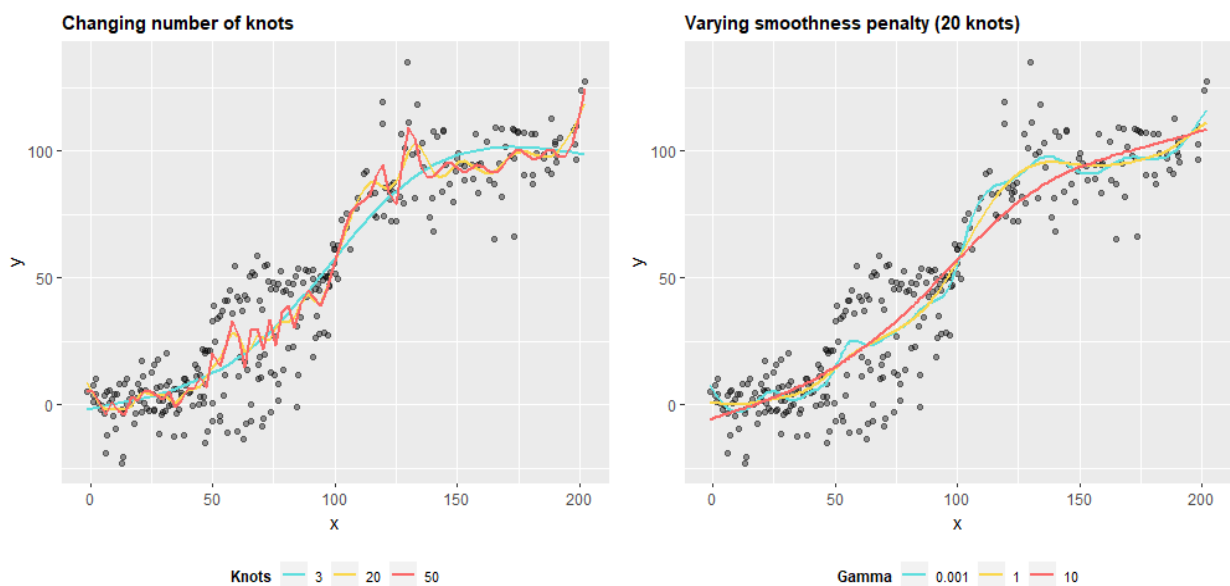


Figure 6.2 Examples of controlling spline smoothers

Approximation to simulate sigmoidal data using natural cubic splines, with smoothness control by number of knots (left panel) and gamma penalty (right panel).

Though higher numbers of knots will allow a larger number of sub-spaces within the model, the penalty will adapt to smooth appropriately. Numbers of knots should therefore be chosen to be slightly higher than required to describe the function, with an appropriate penalty term (Wood, 2017c).

GAMS using smoothers of this type are not restricted to splines and may easily include other smoothers that can be expressed as basis functions. Notable examples include:

- **Thin plate regression splines (TPRS):** TPRS are not splines per se, but can be used as similar low-rank isotropic smoothers, that can be expressed as basis functions. The term isotropic, here, refers to the fact that a smoother produces the same predictions under any rotation or scaling (Wood, 2017d). The term ‘thin plate’ describes an analogy where a three-dimensional surface could be fitted in the smoothest possible sense by bending a flexible plate with just enough tension to minimise the flex in the plate. They are adapted from thin plate splines (Duchon, 1977), and penalise the *“wiggly components of the thin plate spline”* (Wood, 2017d). They can be applied to any number of variables (where an isotropic smoother makes logical sense, e.g. related variables on the same scale). They don’t require the specification of knots, but use an analogous Eigen-decomposition, and have been referred to as an *“optimal smoother”* (see (Wood, 2003, Wood, 2017d) for further detail).
- **P-Splines:** Combinations of B-Spline bases, with penalties on the basis coefficients are referred to as P-Splines (Eilers and Marx, 1996). P-splines can be advantageous in situations where different orders of bases and penalties are required, but in general, do not perform as well as TPRS or cubic regression splines (Wood, 2017b).

6.2.3 Estimation of GAMs

When estimating GAMs, we must simultaneously estimate all smooth functions, parametric predictors and covariance’s between smoothers (Larsen, 2018). This can be achieved in two ways:

- **Local scoring algorithm:** used for loess and smoothing splines, and applicable to other functions. This uses a backfitting algorithm (Hastie, 1992), that iteratively smooths partial residuals.
- **Penalized iteratively re-weighted least squares (PIRLS):** used for regression splines. Models are reparametrized as parametric, penalized GLMs with smoothness selection

by Generalized Cross-validation (GCV), Restricted Maximum Likelihood (REML) or similar techniques. Numbers of knots are not estimated, but the penalty term is.

6.2.4 Model selection and degrees of freedom

Model/parameter selection is more complex with GAMs than GLMs, but λ estimation is commonly performed using cross-validation (Hastie et al., 2009a). Wood (2017a) suggest that much of what is considered parameter selection in GLM/GLMMs is performed by the smoothing penalties. Backwards selection functions are included in Hastie & Tibshirani's `R` code, whereas Wood has not implemented this approach as it is at odds with other theory. A method of applying the LASSO penalty to GAMs has been proposed (Chouldechova and Hastie, 2015), but this has only been implemented for Gaussian and binomial model to date and is not applied here.

Shrinkage methods can also be applied to regression spline GAMs, either by adding a penalty to the identity matrix of each smooth, where strong penalization will shrink coefficients to zero, or by adding an additional penalty to the null space of each smooth so that functions of zero 'wiggleness' are penalized out of the model (Marra and Wood, 2011).

Degrees of freedom of fixed effect models are quite easily defined as the number of parameters to be estimated, which can also be calculated as the trace of the hat matrix. Penalized smoothers add an extra complication, as counting all basis functions would equate to an unpenalized model. Hastie and Tibshirani suggested that, as a trace of the hat matrix in a GAM is on the data projected to basis functions rather than the underlying data, it could therefore be used as the *effective degrees of freedom* (EDF) (Hastie et al., 2009b). The EDF will change as the smoothing penalty changes or with the number of knots (in the case of regression splines). In the `mgcv` framework, it is possible to specify a value to multiply the effective degrees of freedom by, for GCV/REML estimation. This fixed penalty increases 'cost' of wiggly data when estimating smooths, leading to smoother functions (Wood, 2017a). The combination of numbers of knots and penalty selection can mean that it is possible to have the same EDF, despite changing the knots, as the penalty adapts.

Use of AIC is reasonable with the regression spline and smoothing spline approaches, but the effective degrees of freedom must be considered. Uncertainty in the estimates of smoothers should also be considered, and conditional AIC (Grevén and Kneib, 2010, Wood, 2017a) used in place of marginal AIC to counter for this. The `mgcv` package uses conditional AIC as a default, due to smoothing uncertainty, that could also be viewed in the same way as a normally

distributed random-effect around a parameter, sharing the same issue alluded to in section 4.5.1.

6.2.5 Random-effects models in GAMs

Random-effects models can be fitted in both the major GAM implementations in R. In Hastie's `gam` package, a 'random' class of smoother is implemented that allows a shrunken mean-fit within clusters for categorical variables. This is formally equivalent to fitting a mixed model by generalized least squares. In `mgcv`, random-effects can be fitted in two ways, the first approach is as a class of smoother than creates parametric interactions of predictors, penalized by a multiple of the identity matrix (corresponding to the assumption that they are independent and identically distributed). A second approach uses an interface with other mixed model packages (including `lme4` or `nlme`) and reparametrizes the entire model with all smooth terms converted to fixed and random components (Wood, 2017d). Both `mgcv` approaches have been used when fitting the respective models below, but only the former reached convergence.

6.2.6 Fitting GAM models to incident data

6.2.6.1 Model Structure

When fitting a GAM model using the framework described above, the `mgcv` package was used to fit individual smoothers using both natural cubic splines and thin-plate splines. Thin plate splines were also applied as multi-dimensional smoothers where predictors described the same units, e.g. a three-dimensional smoother of AE waiting times, constructed from the 25th, 50th and 75th percentile waiting times. Loess and smoothing spline models (Hastie & Tibshirani's approach, using Hastie's 'gam' package) are not presented in the following section due to poorer performance in terms of MAE, and an inability to fit to a new dataset for comparison. Initial models were based on the Poisson and NB2 GLMMs from Chapter 5. Performance of models was assessed using Mean Absolute Error (MAE).

6.2.6.2 Estimation of smooths

Smooths were initially fitted with the `mgcv` default of 10 knots per smooth. Smoothness estimation was examined using ML, REML, penalized REML and GCV, with REML appearing the more robust to overfitting and giving the best MAE performance. Performance for smoothers was tested using the `gam.check` function, which creates residual plots, marginal smooth

plots and tests effective degrees of freedom against the reduction in deviance. Visual inspection of the marginal smooths was used primarily, as the overdispersion affected deviance tests. The principle of these models is to model underlying (potentially non-linear) trends without overfitting the noise/overdispersion/wiggleness in a variable. Smoother fits were therefore preferred. Numbers of knots were increased where required, for total bed-days, proportions with comorbidity score = 0, 25th and 50th percentiles of A&E wait-times. Figures 6.3 and 6.4 show the plots of the marginal smooths against the predicted values, showing overdispersion gives different smooths in the Poisson and NB2 models. Smooths were noticeably less wiggly with the NB2 model compared to the Poisson and suggested much of the wiggleness related to the overdispersion.

The categorical variable for teaching hospital was entered as a categorical term, as it does not make sense apply smoothing to this sort of term. Organisation was included as a random-effect as described in section 6.2.5.

6.2.6.3 Penalization and Selection

MAE performance was best for REML selection of smoothing penalty, although the standard penalty was subject to the overdispersion in the data. An additional fixed penalty, based on BIC (Schwarz, 1978), was added to all terms (Wood, 2010). The value of the penalty was ≈ 3.69 , and corresponded to the following formula, where k = the number of parameters:

$$penalty = \frac{\log(k)}{2}$$

This reduced function wiggleness, increased the training MAE slightly, but greatly reduced the MAE when fitted to the testing dataset. Shrinkage penalties as described in 6.2.4, were examined but no improvement in MAE was observed over the BIC-type penalty models. BIC-type penalization notably altered the shape of smoothers, but shrinkage penalties showed little change.

6.2.6.4 Model Output

The conditional AIC, and prediction error for training and testing sets are presented for candidate models in table 6.1. Testing and training sets were not scaled as described in earlier chapters, as the transformation to spline bases achieves this and pre-scaling can be problematic for the construction of multivariable TPRS, as variables should be on the same scale.

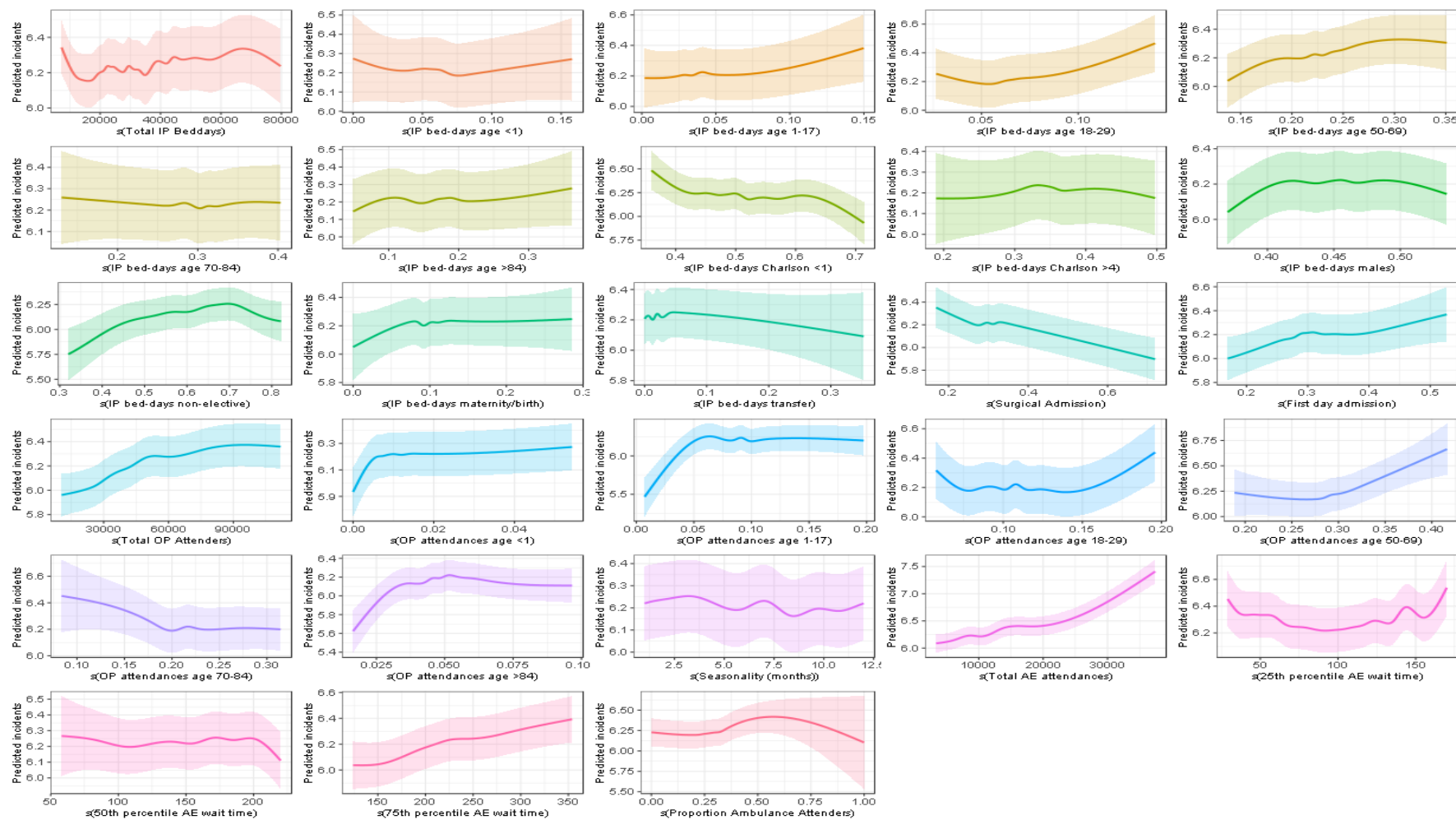


Figure 6.3 Marginal smooth plots for NRLS-HES Poisson GAMs

GAMS fitted without additional penalties using single-dimensional cubic splines, plotted against predicted incidents.

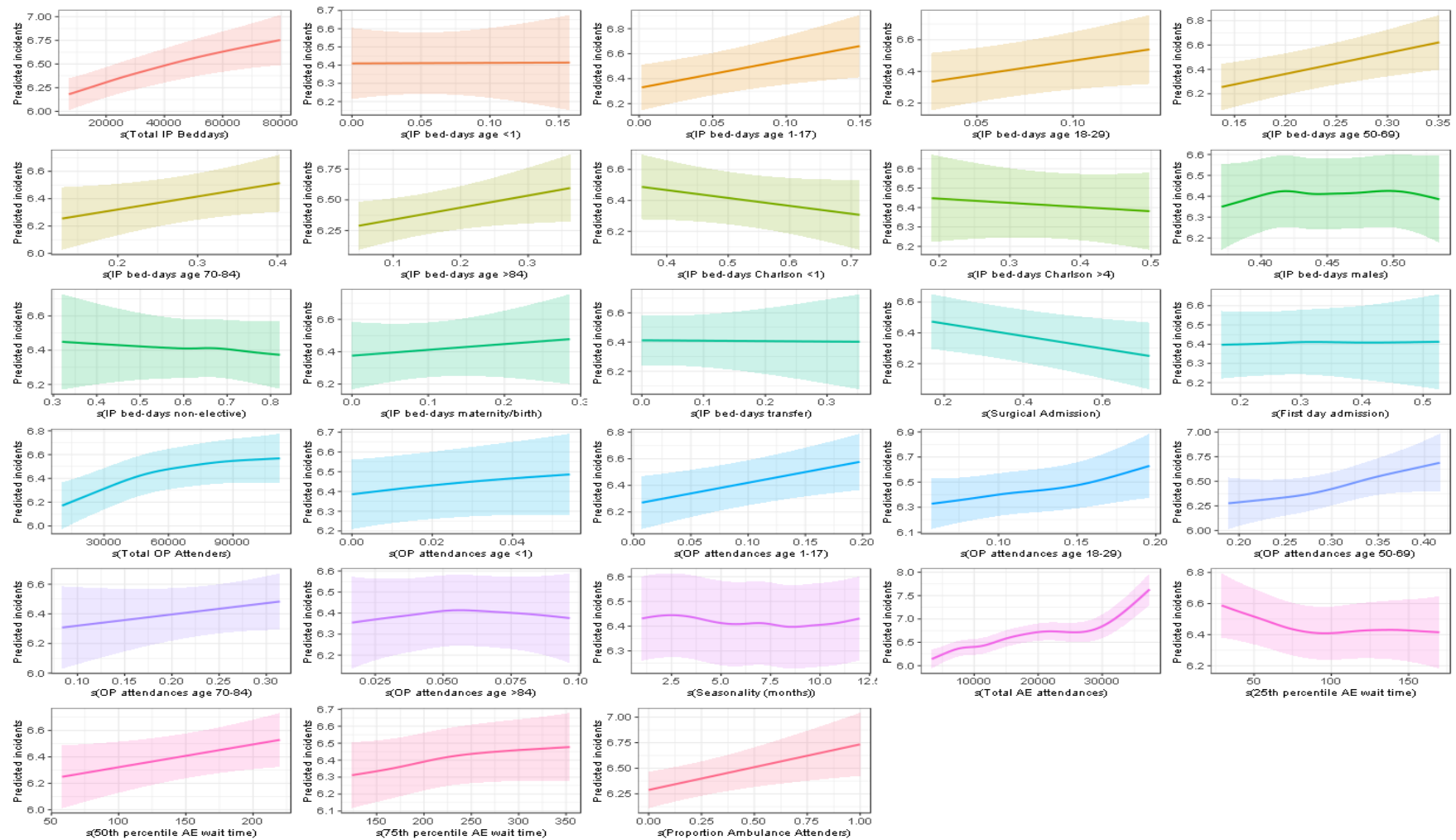


Figure 6.4 Marginal smooth plots for NRLS-HES Negative Binomial (NB2) GAMs

GAMS fitted without additional penalties using single-dimensional cubic splines, plotted against predicted incidents.

Model	Shrinkage	Fixed BIC penalty	Poisson					NB2				
			AIC	Mean Absolute Error (MAE)				AIC	Mean Absolute Error (MAE)			
				Training		Testing			Training		Testing	
				Raw	% of median (725)	Raw	% of median (745)		Raw	% of median (725)	Raw	% of median (745)
Single cubic splines	N	N	23035.2	48.61	6.7%	112.55	15.1%	18834.1	54.23	7.5%	104.19	14.0%
Single cubic splines	Y	N	23065.8	48.78	6.7%	112.37	15.1%	18808	57.94	8.0%	97.43	13.1%
Single cubic splines	N	Y	24073.3	51.22	7.1%	103.36	13.9%	19070.3	54.26	7.5%	104.59	14.0%
Single cubic splines	Y	Y	24079.3	51.24	7.1%	103.92	13.9%	19055.4	58.35	8.0%	97.22	13.0%
Single TP splines	N	N	22859.6	50.40	7.0%	120.68	16.2%	18814.2	53.56	7.4%	109.15	14.7%
Single TP splines	Y	N	22928	47.88	6.6%	118.78	15.9%	18791.5	57.95	8.0%	97.48	13.1%
Single TP splines	N	Y	24073.3	51.22	7.1%	103.36	13.9%	19078.3	53.62	7.4%	109.06	14.6%
Single TP splines	Y	Y	24032.2	51.16	7.1%	108.95	14.6%	19058.9	58.33	8.0%	96.87	13.0%
Multiple TP splines	N	N	20274.9	37.10	5.1%	8.61E+22	1.16E+20	18862.4	48.72	6.7%	8.02E+16	1.08E+14
Multiple TP splines	Y	N	20488.6	38.19	5.3%	134.92	18.1%	18752.2	62.70	8.6%	2.41E+18	3.23E+15
Multiple TP splines	N	Y	21789.8	43.53	6.0%	2.28E+16	3.05E+13	19780.9	57.98	8.0%	96.68	13.0%
Multiple TP splines	Y	Y	22742.6	46.95	6.5%	111.58	15.0%	19040.7	52.19	7.2%	111.11	14.9%

Table 6.1 GAM model outputs for NRLS-HES models

Rows represent model smoother types, with ‘shrinkage,’ and ‘fixed BIC penalty’ referring to model settings. Poisson and Negative Binomial (NB2) models are column groups, with training and testing mean absolute error (MAE) displayed as raw values and as a percentage of the median incident report value. ‘Training’ refers to the 2015/16 data used to build the models and ‘Testing’ refers to the predictions from the model for new data for 2016/17.

Final models were affected by the overdispersion present within the data, rendering AIC comparison less helpful than comparisons of MAE. Cubic spline and single-dimensional TPRS smoothers gave good performance, but multi-dimensional TPRS smooths led to very high testing error in some cases.

6.2.7 GAM model conclusions

GAM models, including random-intercept terms, have shown a reduction in mean absolute error compared to Poisson and NB2 GLMMS in Chapter 5. Overdispersion and overfitting have, again, been issues in GAMs. Model selection via AIC, even using conditional AIC, would have selected overfitted models that did not generalize to the testing dataset well. Shrinkage penalties aided reductions in MAE in some, but not all cases. The same could be said for the fixed, BIC-like penalty. The best performance, with the exception of the Poisson single-dimensional TPRS model, was seen when using both the shrinkage and BIC-like penalty. When used together, the combined penalties showed further improvements over the fixed BIC-like penalty alone, suggesting that some of the model covariates are of low predictive value and may be confounding other predictors. The most plausible explanation for the combined effects is that the regularization from both penalties reduced the influence of the poor quality predictors that may contribute to overdispersion/confounding, and the increased fixed penalty reduced the 'wiggleness' of the smoothers. The scale of the overdispersion/noise in the data would have been reflected to some extent in the estimation of smoothers by conventional means, but enforcing a smoother fit represents the expected relationships in a more general fashion. Smoother models then fit testing data more appropriately, as they have minimised overfitting in the training set.

Multidimensional TPRS showed very good fit to the training data, reducing the MAE substantially, but did not generalize well, suggesting they were overfitting. This is likely due to the high number of basis functions, giving a complicated surface to the smoothers, and reflecting much of the noise in the data. These complex surfaces, although composed of similar lower-order terms, would be a poor fit to testing data due to the high dimensionality of the smooths. The additional penalties and shrinkage gave smoother surfaces and reduced testing MAE. Given the better MAE performance of single-dimensional smooth models, and the lower complexity for both computation and interpretation, the cubic spline models may be considered the best option. The model selection for GAMs suggest that the NB2 models gave

lower testing error, but lowest training error was seen in the multidimensional TPRS smooths, due to overfitting.

6.3 Algorithmic methods

Whilst ‘machine learning’ (ML) can be considered a field of statistical analysis, the term is frequently used in computational settings and tends to imply a focus on prediction. Statistical modelling typically assumes a data model (usually a distribution or data generating mechanisms) before estimating parameters according to the assumed model, whereas machine learning techniques commonly *“avoid(s) starting with a data model and use(s) an algorithm to learn the relationships between the response and it’s predictors”* (Elith et al., 2008). Breiman (2001b) described this difference, saying: *“There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models.”* The following section discusses and applies several common algorithmic methods to the dataset to assess whether they provide a better fit.

6.3.1 Tree-based methods

Methods presented so far have followed classical frequentist statistical estimation techniques, based on maximum likelihood and Generalized Linear Models. An alternative approach to regression is common in machine learning applications, based on ‘trees’ (Breiman et al., 1984).

Decision Tree models are algorithmic, and agnostic as to the data generating mechanism, but are often suitable for predictive models where we do not necessarily need to examine the relationship with each predictor (Shmueli, 2010). Regression trees/classification trees do not use maximum likelihood estimation, and a full parametric distribution is not required, they are considered good options for non-parametric models. Regression trees in-particular tend to use a loss functions such as root-mean squared error, but can also use many other loss functions such as accuracy, KL divergence, MAE, Poisson etc.

Methods exist for both Classification And Regression Trees (‘CART’) that are conceptually simple, with simple trees easy to visualise (Figure 6.5). “Leaves” or “nodes” represent the groups after partition, and “branches” describe the paths of split. Tree models recursively partition data based on predictors that explain the most variance, with each split proceeding separately, using an appropriate loss function to judge error/deviance and stop at an appropriate point. For categorical predictors, the splits simply map to categories, but for

continuous variables, both the variable and the cut point are estimated (Breiman et al., 1984). In the case of regression, a common splitting method is to find the variable and cut point that maximises the between group sum of the squares, equivalent an analysis of variance (commonly referred to as the `anova` method) and is therefore a mean-squared error loss function. For event rate data, a Poisson method may be used, that uses likelihood ratio tests between the two nodes, with the loss function then becoming the deviance contribution for a new observation, using the average event rate of that node (Therneau and Atkinson, 2018).

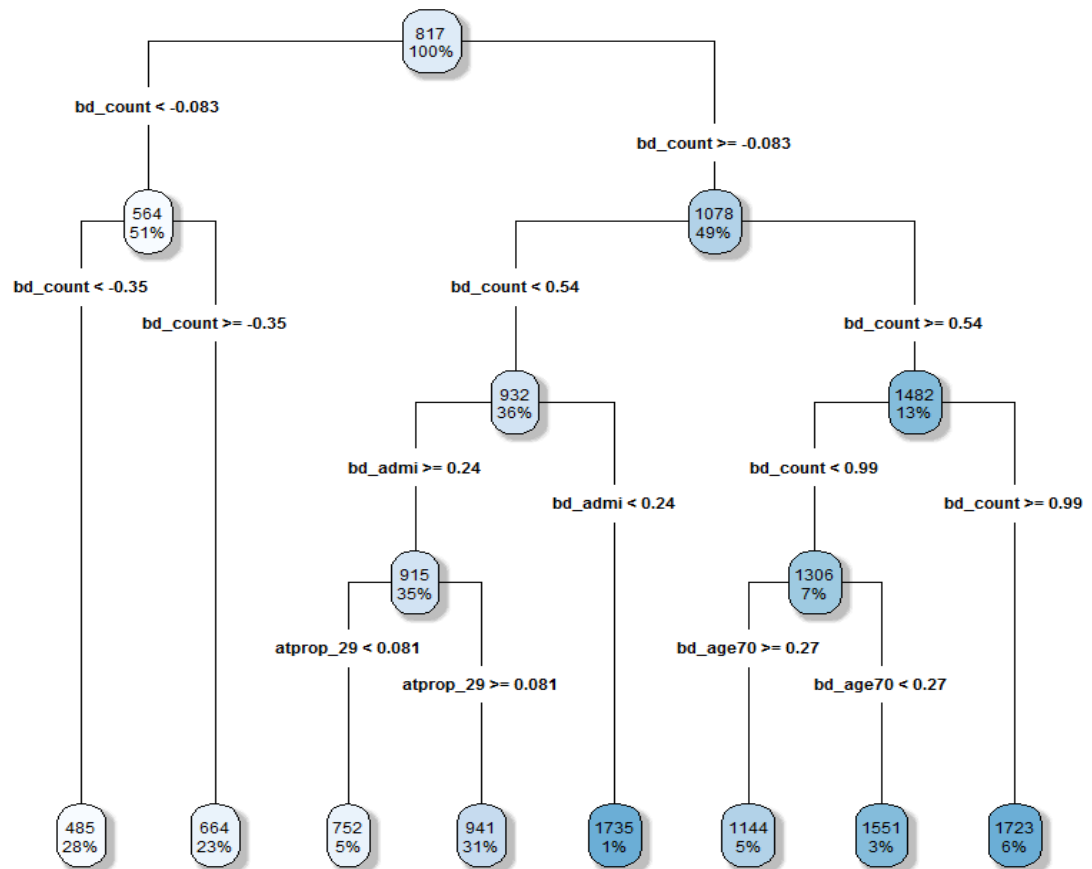


Figure 6.5 Example structure for NRLS regression tree models

NRLS regression tree model from section 6.3.2.1. Trees recursively partition the data at nodes, splitting at points that explain the most variance and stopping based on appropriate rules.

Tree-models are known to overfit training data. They will continue to partition variance in the training set, even if it is noise, down to nodes containing individual results. To counter this, early stopping rules are often implemented, such as: a maximum number of nodes/leaves, minimum variance required for partitioning, minimum number of results in terminal nodes etc. (Breiman et al., 1984) suggested a ‘one standard deviation’ rule, where the simplest model within one standard deviation of the best model is chosen.

A common alternative to pre-specified rules is to allow full estimation of a tree and “prune” it. Pruning involves removing terminal nodes and comparing trees on a complexity function. The optimal complexity function can be controlled by using cross-validation.

Overfitting is the dominant problem for regression trees. Further developments on methods presented in Chapter 4 can be applied, such as cross-validation and bootstrapping to improve tree-based models. Two common improvements based on resampling are: boosting, and bootstrap aggregation (‘bagging’).

6.3.1.1 Boosting

A regression/classification tree may be ‘boosted’ by following an adaptive refitting algorithm, such as the popular ADABOOST algorithm. Boosting generally iterates through the following steps:

- Fit an initial model (usually a ‘weak learner’, see below)
- Estimate a cost function such as root mean squared error
- Re-weight the data using the cost function, with higher weighting on points with highest cost/error (encouraging the algorithm to bias towards these results)
- Re-fit the weighted model
- Iterate the previous steps until convergence criteria/stopping rule is met.

A boosted regression tree can therefore be understood as an additive regression model in which individual terms are each simple tree, fitted in a forward, stage-wise fashion (Friedman et al., 2000). Friedman’s ‘gradient boosting machines’ (GBM) are these additive regression models, estimated by gradient descent techniques. These models sequentially fit least squares estimates to ‘pseudo-residuals’ that form a gradient of the loss function that can be minimised by descent techniques (Friedman, 2002). Gradient descent takes steps down the gradient until it finds the local minimum, with the step-size referred to as the ‘learning rate.’ A high learning rate converges on the minimum more rapidly, but risks overshooting the true minimum, and a lower rate takes longer to fit but is more likely to converge on the true local minimum.

Shortly after his initial publications on boosting, Friedman proposed an alternative approach referencing Breiman’s work on ‘bagging’ (see below) to fit ‘weak learners’ to sub-samples of training data. A weak learner is a model that performs only slightly better than random guessing. In his landmark paper, Schapire (1990) showed that weak learners can be used to construct ‘strong’ learners, and have many useful properties including lower computation burden and reduced correlation effects.

The question of overfitting in boosting is not fully understood. Some proponents have shown boosting to rarely overfit data (Schapire et al., 1998), and this has led to a misconception, prevalent in some online ML and statistics communities, that boosting cannot overfit data. Boosting has, indeed, been shown to overfit data (Freund and Schapire, 1996, Breiman, 1999), particularly when there is a high degree of noise (Long and Servedio, 2010, Dietterich, 2000). Where the stopping rules are unclear because of noise such as overdispersion, boosting will continue to iterate until the rules are met. Boosted classifiers appear to be more robust to overfitting than boosted regression trees, and removal of spurious results has been suggested as a remedy for classification models with overlapping classes (Vezhnevets and Barinova, 2007), but regularization is the main option for GBM regressions. Cross-validation can be used to choose the appropriate regularization parameter that acts to limit the number of explanatory variables used at each step (Friedman, 2001). This effectively acts like a testing/training split, with the loss assessed on the hold-out sample, and acts more appropriately than rigid stopping rules.

Whilst boosting is an effective technique, it can be a complex to apply as it has many tuning parameters, including the learning rate, number of trees, boosting iterations etc.

6.3.1.2 Bagging

An alternative to boosting, and a common next step for regression trees, is bootstrapping (Efron, 1979). As previously described in Chapter 4, bootstrapping is a random resampling and replacement approach where a sample of training data is used to calculate a statistic (or a regression tree in this case) with the process repeated multiple times on new samples. The parametric nature of the resampling allows models to be averaged over many repeats. Bootstrap aggregation (“bagging”) (Breiman, 1996a) has been proposed as a solution for reducing overfitting. The aggregation element means that bootstrapped models can then be averaged to determine the final parameters. This technique can be applied to many modelling techniques, but has proved particularly useful for regression/classification trees where small perturbations in training datasets can cause large differences in tree structure. Bagging can be extremely effective in reducing the variance in unstable estimates like trees (Hastie et al., 2009b). Bagging has been shown to be more successful than boosting in noisy datasets (Dietterich, 2000), but improvements in bagging may be limited due to correlations between prediction, and between models.

6.3.1.3 Random Forest

A further adaptation to bagged trees is the ‘Random Forest’ (Breiman, 2001a). This maintains the idea of using bootstrapped samples of the training dataset, but also randomly selects a subset of the available predictors (commonly referred to as ‘features’ in relevant publications). It is therefore averaging across weak learners, similar to boosting.

This process reduces the correlation between trees, as trees do not all contain the same predictors. The de-correlated trees are, in general, more robust to overfitting than bagged trees alone and, when trained appropriately, perform similarly to boosted trees or better. The major advantages of Random Forests are the comparative simplicity of training compared to GBM, where only the number of predictors in a random forest (the ‘complexity’ or ‘mtry’ parameter) and the number of trees grown are tuned. They also perform comparatively well over a range of tuning parameters.

Random forests can use various stopping rules, but commonly the out-of-bag (OOB) error is used, in preference to cross-validation, as it requires no additional computation. If bagged trees/boosted trees or random forests are fitted on bootstrapped samples, the portion of the dataset not used to train the model, can be used as a testing set instead. Models are built on the bootstrapped training sets, fitted to the remaining testing data, and the prediction error in this set drives a more robust fit (Breiman, 1996b). OOB can be used when tuning the model to select the number of trees or the complexity parameter.

6.3.2 Applying regression trees and extensions to NRLS data

The approaches described above have been used to fit tree-based models, improving these model’s performance using bagging and boosting approaches, and finally, using Random Forests. Model performance was, as per chapter 5, best summarised in terms of prediction error using Mean Absolute Error. This was used for final model comparisons, but relevant loss functions (such as OOB) were used for training each model as required.

The models discussed above have various implementations in R, but the most common implementations were used for each. All models were fitted using two different approaches; direct use of the modelling functions themselves, and through the model fitting framework ‘caret’ (Kuhn, 2008). *Caret* is a suite of standardised model building functions, interfacing with other modelling procedures, that are all called through the same interface. They allow for identical pre-processing, fitting, training and optimization processes. One of *caret*’s many strengths is the ability to optimize models over a grid of hyper-parameters (model tuning parameters). In some cases, there are no exact analytical solutions to models, and

systematically testing combinations of tuning parameters is an acceptable, if computationally expensive, approach to select the lowest prediction error. *Caret*'s cross-validation procedures were applied in parallel on a 4 CPU desktop, using Microsoft R Open with its optimized BLAS as described in Chapter 5.

6.3.2.1 Regression trees

The *rpart* package was used to build regression trees. Default optimization options were initially used, with the tree structure visualised using the *rpart.plot* package (figure 6.5). A stopping rule of at least 20 observations per split was used, with a Poisson loss function. As Poisson loss functions perform likelihood ratio tests between nodes, there was potential for overdispersion to affect results. The *Anova* method was also examined using RMSE as the loss function, with poorer results. As an alternative to the stopping rule, trees were also allowed to grow fully and were pruned on the complexity parameter (*'cp'* – another term for the loss-function), but both options resulted in the same tree structures and depths. Cross-validation was used to train trees and select an optimal complexity parameter giving the lowest prediction error. *Caret* was used with 10 repeats of 10-fold cross-validation across a *'cp'* range 0.0001 – 0.1. Complexity parameter of 0.01 gave the best performance based on cross-validation and was used for final tree model.

6.3.2.2 Boosted trees

Friedman's gradient boosted trees were implemented using the *gbm* package, originally by Greg Ridgeway (Greenwell et al., 2018). There are four mandatory tuning parameters within the *gbm* function:

- **Interaction depth:** The number of split nodes for trees grown at each iteration. E.g. an interaction depth of three allows up to three levels to trees. A range of 1 – 5 was tested.
- **Number of trees grown:** The number of trees grown. A range of 1,000 – 15,000 in increments of 1,000, were tested.
- **Shrinkage:** The step -size for the gradient descent ("learning rate"). Values tested were 0.001, 0.005, 0.01, 0.05, 0.1, 0.5 and 0.1. Shrinkage and number of trees interact, as more trees are needed with smaller learning rates.
- **Minimum observations in node:** Akin to an early stopping rule, this aims to prevent overfitting. This was set to 20, similar to the regression trees.

Boosted trees were fitted directly using the `gbm` package, and using parameter grid search and 10-fold cross-validation in `caret`. Cross-validation showed best performance with an interaction depth of 5, number of trees set to 3,000 and a shrinkage/learning rate set to 0.01.

6.3.2.3 Bagged trees

Bagged trees were examined using the `ipred` package using a complexity parameter of 0.01, as identified from the cross-validation of the `rpart` trees. 25, 50 and 100 bootstrapped resamples were performed, leading to reductions in both training and test error.

6.3.2.4 Random Forest

Random Forests were fitted to the dataset in R using an implementation of Breiman and Cutler's original Fortran code, ported to R (Liaw and Wiener, 2002) in the `randomForest` package, and also using `caret`. Random forest tuning is based on the 'mtry' value, or number of predictor variables to be randomly sampled, and was compared on OOB using 1000 trees. This was performed in the native package using the `tuneRF` function, and in `caret` by grid search across a range of 1 – 15. Both methods refit using different values of mtry, with `tuneRF` following a search algorithm, and `caret` searching all values in the grid. Both approaches agreed in output and selected an mtry value of 13.

6.3.3 Tree-based Model Results and Conclusions

MAE from prediction for the various tree-based methods is shown in table 6.2. Regression trees fitted the data quickly, but MAE was comparatively high in training and testing sets. Bagged trees showed an improvement on single regression trees, but this was limited to small decreases in MAE. Boosting showed the most marked improvement in the training set, reducing training MAE to the smallest value seen. Whilst testing MAE was also dramatically reduced for boosted trees, Random Forests showed the best testing set MAE. Random Forests also reduced training error to approximately half that produced by simple trees, and appeared to give the best performance.

Method	Package		caret	
	2015/16	2016/17	2015/16	2016/17
Tree	141.48	155.60	131.91	146.75
Boosted Tree	15.99	109.99	16.28	108.30
Bagged Tree	131.46	142.47	122.68	140.76
Random Forest	73.75	102.92	23.94	102.60

Table 6.2 Mean Absolute Error (MAE) for regression tree-based NRLS-HES models

Rows represent model type, with columns showing MAE for training (2015/16) and testing (2016/17) data. 'Package' refers to models fitted from their standard package functions, and 'caret' refers to the model fitting framework with hyper-parameter grid-search and 10-fold cross-validation

The boosted tree models with extremely low MAE in the training set suggest overfitting, but no adjustment was made for overdispersion and trees were based on a Poisson (log-likelihood) loss function. The unbiased nature of the Poisson distribution allows accurate prediction in this case, but the stopping rules based on the loss function will be compromised by the overdispersion. These techniques could be further developed by creating a suitable custom loss function, such as a robust Poisson or quasipoisson function, if either of these approaches is feasible in an iterative process like boosting.

The tree-based frameworks do not explicitly model the repeated measures structure in the data. The success of random forest suggests that correlation between the predictors is a major source of overdispersion. The bootstrapping and aggregation appear to have dealt with this better than other strategies. Random Forests appear to be a competitive method for developing predictive models in this dataset, but the lack of repeated measures adjustment is of concern. Although when compared to the GLM models (i.e. no random-effects), these models could be considered to have out-performed them.

6.4 Artificial Neural Networks

6.4.1 Artificial neural network structure and estimation

Neural networks have been in development for many years (Yadav et al., 2015), and were somewhat eclipsed by techniques such as boosting in the 1990s (Efron and Hastie, 2016), but have regained popularity in the fields of machine learning, Artificial Intelligence (AI) and 'Deep learning.' They are high-profile in popular culture, perhaps with an over-emphasis on what they can achieve (Chollet and Allaire, 2018). Neural networks essentially extract linear

combinations of inputs, as derived/projected features, and use them to model non-linear functions of these features (Hastie et al., 2009a).

The term 'Neural Network' applies to a large class of related models and learning methods. They are, at their heart, two stage regression or classification systems that have particular feedback, regularization or other structures that can be combined in sequences. (Hastie et al., 2009a)

Neural Networks have a conceptual link to how neurons are arranged in human brains, although this is a loose link (Goodfellow et al., 2018). Each neuron in an artificial neural network is a function that receives inputs, and 'activates' based on their values. It then transmits the signal forward, analogous to a transmission of a human synapse. In this section, the term 'node' is used interchangeably with 'neuron' to refer to these artificial neurons.

Neural Network techniques are an expansive subject, and the depth of their structures and applications is beyond the scope of this chapter. For this application, a single architecture: the 'feed-forward' neural network or 'Multi-layer perceptron', has been examined.

An example of this type of feed-forward network, with a single hidden layer, is illustrated by Figure 6.6. A series of inputs ($X_1 \dots X_p$) are fed into the network (blue). Derived features Z_m (red) are created from linear combinations of the inputs and used to predict Y_k (yellow) (Hastie et al., 2009a).

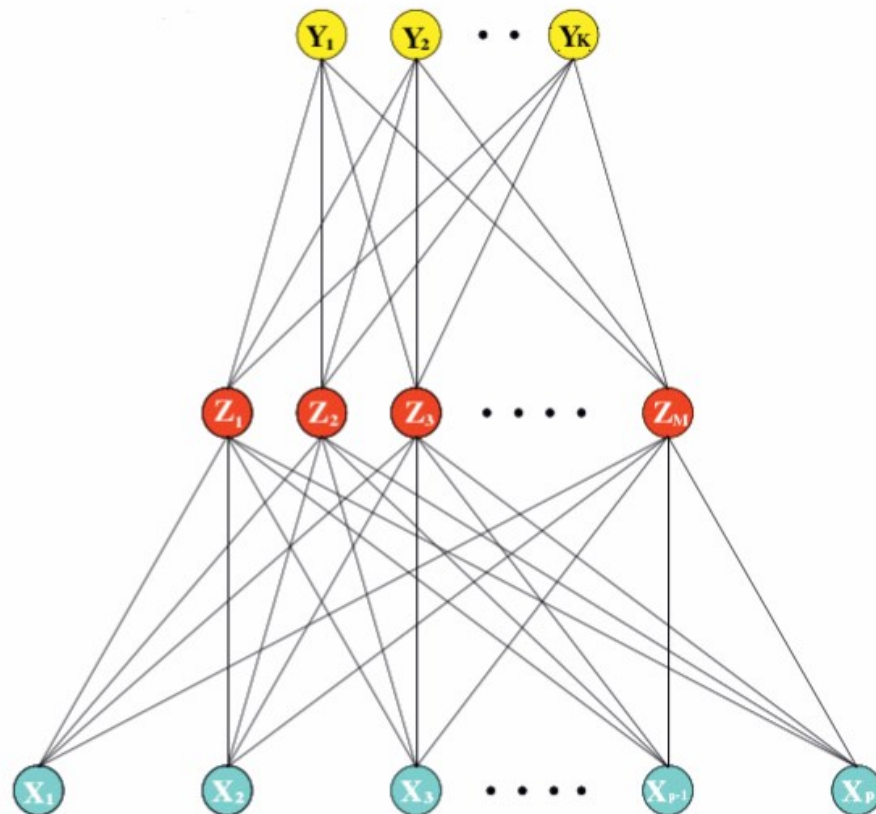


Figure 6.6 Schematic of a single hidden layer, feed-forward artificial neural network

Taken from Hastie et al.(2009a). Inputs ($X_1 \dots X_p$) are fed into the network (blue), with derived features Z_m (red) are created from linear combinations of the inputs and used to predict Y_k (yellow). Neurons are fully connected to all other in the next layer.

Z_m are referred to as ‘hidden layers’ because they are not directly observed. Network architectures can have a single hidden layer or many layers, with the term ‘deep learning’ commonly applied to networks with more than one hidden layer (although exactly what constitutes ‘deep’ is not well defined (Goodfellow et al., 2018)). These layers are described as ‘densely connected,’ as all nodes in our input are connected to all neurons in the next layer, and so on, from one layer to the next. Our inputs could now be considered as basis expansions, feeding a linear model in the case of a regression, or a logistic for binary classification (Hastie et al., 2009a). The structure and design of hidden layers is an active area of research, with few established ‘gold standards’ (Goodfellow et al., 2018) regarding the number of layers, or strategies for selecting the number of nodes in each layer. It is common, however, to have a first hidden layer that corresponds to the number of input parameters. Increases in node numbers from one layer to the next expands the capacity of the model into higher dimensions, and fewer nodes reduces the dimensionality, often towards a desired number of output parameters.

Each node within a layer receives inputs from one or more nodes, either from the input or the previous layer. The nodes multiply the inputs by a weight and sum them, before passing them to an activation function. The goal of training a network is therefore to set these weights appropriately, iterating towards the best solutions.

The activation function is another key concept of the neural networks that regulates the output of each layer. Activation functions may differ between layers and are commonly different at the output layer. The simplest activation function is the identity function, where inputs are weighted and summed but no further action is applied. For binary classification, a sigmoidal activation function can be used that is similar to the logit function, and saturates at zero and one. For multiclass classification the 'softmax' function (Efron and Hastie, 2016), a generalization of the sigmoid function, can be used. In many cases, for continuous data or transitions between hidden layers, a rectified linear activation function (ReLU) is suitable (Hastie et al., 2009a). ReLU is similar to a linear activation function, but truncates at zero and prevents negative signals being passed. Activation functions are commonly applied with a small bias that ensures that most units are initially activated for most inputs, allowing the signal to pass through for training. A bias node may also be included in each layer, that is not connected to the previous layer, and is permanently open. Bias nodes allow the output of an activation function to be shifted, conceptually similar to an intercept term in a regression equation.

Once a network architecture has been specified, the weights of the nodes in each layer can then be tuned (the 'learning' element of the neural network) from the training data, and an appropriate loss function (such as mean squared error (MSE) or mean absolute error (MAE)) is calculated to assess the fit. 'Back-propagation,' a method of differentiation across layers using the chain rule (Rumelhart et al., 1986), is the standard method for feed-forward networks, used to estimate the gradient of the loss (Chollet and Allaire, 2018). This error, per data row is then fed-back through the network in reverse and used to alter the weights associated with each link between the neurons, attempting to minimise the prediction error.

A major difference between neural networks and traditional linear models is that the loss functions tend to become non-convex due to the high dimensionality. As such, they are usually solved by stochastic gradient-descent optimisers (SGD), as opposed to least squares or finding a global optimum for convex functions such as the log-likelihood (Goodfellow et al., 2018). Adaptive optimizers, such as ADAM (Kingma and Ba, 2014) often improve training speed, but do not perform as well as stochastic gradient descent optimizers in terms of generalization to new datasets (which is of importance for application to NRLS) (Wilson et al.,

2017). Neural networks trained in this manner therefore share some of the properties of gradient boosting machines (section 6.3.1.1).

Overfitting with Neural Networks, as with other machine learning techniques in this chapter, is a major problem with this type of model. As with previous models, training and testing sets are used for evaluation, but the training set is further split to create a validation set. This is usually smaller than the remaining training data (e.g. 80% training and 20% validation). The loss function is then monitored in both training and validation sets, commonly by plotting. In both sets, training and validation loss functions should reduce (or increase, depending on the loss functions used) with training loss dropping below validation loss. As training continues, training loss may plateau or continue to reduce, but validation loss will begin to increase as overfitting starts to occur. This is the ideal stopping point for the training process. As with many regression models, or tree-based techniques, neural networks benefit from standardised inputs, but various elements of neural network construction can improve fit and reduce the chance of overfitting:

- **Training time:** models trained for too long will start to fit noise. A reasonable number of epochs (one forward and backward pass through the network for all data points), can be controlled.
- **Samples/Batches:** Within each epoch, the number of training samples used in batches (often referred to as 'mini-batches') can be varied. A neural network may pass individual data items, small-batches of data, or the whole data set in a single pass. E.g. with a data set of 200 points, and a mini-batch size of 50 data points, an epoch would represent four mini-batches (equating to four passes through the network and four updates to the weights). The larger the batch size, the quicker the training will be and the larger the gradient for optimization. The main issue with this is that the gradient tends to become less linear (Goodfellow et al., 2018), may get stuck in local minima, or converge to sharp minima that tend to be associated with overfitting and poorer generalization (Keskar et al., 2016). Smaller batch sizes increase training time but may offer a small regularizing effect (Wilson and Martinez, 2003).
- **Drop-out:** A portion of the nodes, and their connections, can be randomly dropped out of training at each pass through the network. This produces 'thinned' models that can be averaged and prevents networks adapting 'too much' (Srivastava et al., 2014). This is conceptually similar to the process used by Random Forests, where correlations are reduced due to the sub-sampling of predictors.

- **Regularization:** Goodfellow et al.(2018) define regularization as ‘...any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error.’ Various strategies exist to do this, such as constraining parameter values or use of penalties use as ℓ^1 or ℓ^2 norms (Hastie et al., 2009a).
- **Batch regularization:** conceptually the same as mean-centring and scaling by the standard deviation within each batch (i.e. scale parameters so their mean is zero and variance is one), can reduce scale shifts between layers (Ioffe and Szegedy, 2015). It has also been shown to allow the use of a higher learning rate and reduce training time, sometimes removing the need for drop out altogether.
- **Altering the learning rate:** as mentioned in section 6.3.2.2 in relation to boosting, learning rate is the step-size used by the gradient descent algorithm. A higher learning rate will converge more quickly, but risks overshooting the local minimum and having similar effects to large batches regarding the sharp minima mentioned above (Efron and Hastie, 2016).
- **Early stopping rules:** stopping rules can be applied but differ somewhat from the approach for tree models. These rules are commonly loss function related and aimed at stopping prior to overfitting. Various rules exist, including stopping if the loss function does not improve by a certain threshold for a certain number of epochs.

6.4.2 Fitting neural networks to NRLS data

A feed-forward neural network, like the structure described above, was applied to NRLS data. Various frameworks exist for fitting Neural Networks, but a current leader is Google’s TensorFlow (Abadi et al., 2016), originally developed for Google’s internal use but released as Open Source software under an Apache Foundation licence. TensorFlow has major strengths in its flexibility for model building and its ability to run on CPUs or GPUs on multiple environments/operating systems. TensorFlow has its own syntax, but can be accessed through interfacing with other languages including Python, Java and R. The R package *Keras* (Allaire and Chollet, 2018), based on the ‘keras’ Python interface to TensorFlow, has been used to fit models. The training dataset was the same data fitted for models earlier in this chapter and in Chapters 5. The scaling method described in Chapter 5, mean-centred and divided by two standard deviations, was maintained for the input layer for comparison against the GLMM models.

An initial simple network with an input layer, a single hidden layer with the same number of nodes as the inputs, and a single node output layer was used as a first step. This architecture

was then augmented by adding one additional hidden layer at a time. Numbers of nodes in each layer were increased and decreased by 25% and 50% respectively and MAE loss function evaluated. A training/evaluation split of 80% to 20% was used for validation, and final MAE was tested on the 2016/17 testing dataset. Reduction in the loss function can be seen in Figure 6.7.

Additional model settings were chosen by iteratively fitting and increasing or decreasing each setting in turn. Model options and final settings included:

- Batch normalization was used and improved both training and testing error in all models, so was retained for the final model.
- ℓ^1 and ℓ^2 regularization did not improve model fit further than batch normalization and were not used in the final model.
- Early stopping rules did not appear to improve generalization to the test set and were not used in the final model.
- Drop-out was tested but did not improve performance in any setting and was not used in the final model.
- Learning rate was set at the default values of 0.01 (also the threshold that `gbm` optimization selected in section 6.3.2.2). `Keras` settings allowed the model to reduce its learning rate when the loss function reached a plateau. This allows smaller steps for the gradient descent and allowed higher resolution for identifying the local minimum.
- Stochastic Gradient Descent ('SGD') optimizer was used.
- 60 epochs with mini-batch sizes of 32 provided the best performance on testing data.
- Models were fitted both with and without a constant from a bias node. The bias improved model fit and was retained in the final model.
- Final models used five hidden layers and one output layer, all using ReLU activation functions, with node numbers 37 (36 plus intercept), 24, 12, 6, 3, 1 from first layer to output. An initial layer of 48 nodes, expanding inputs to higher dimensions, did not improve fit.

Final prediction error (MAE) was 59.40 for training set and 114.83 for testing. Plots of model error suggested a plateau from 30 epochs.

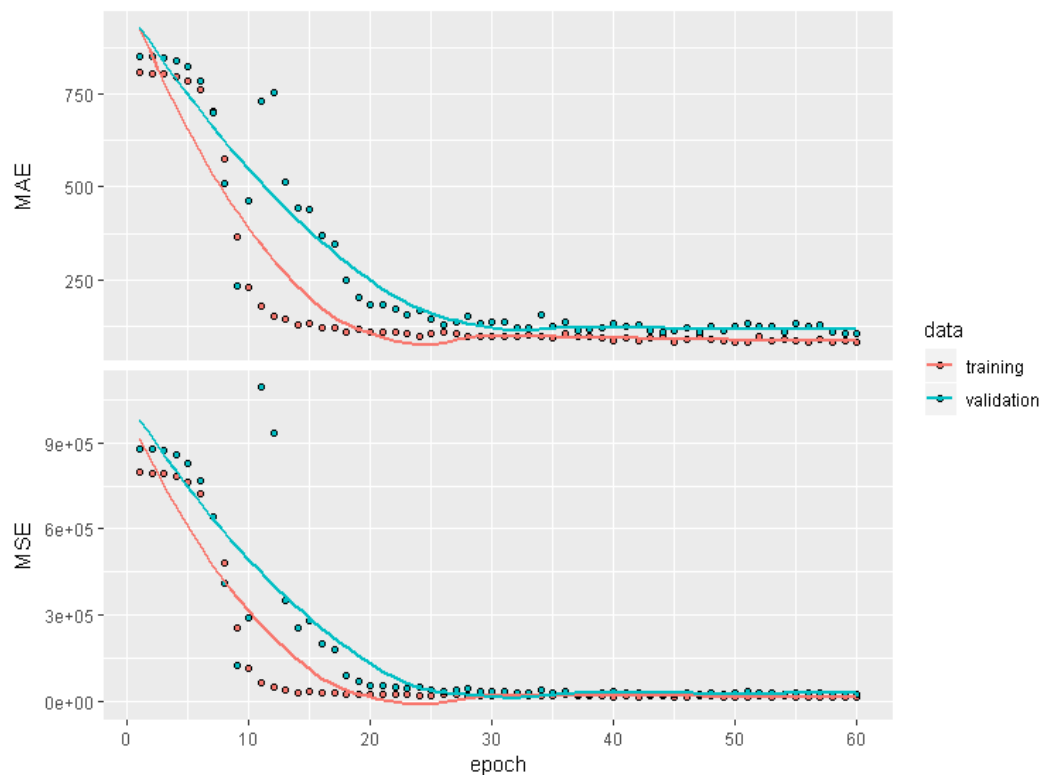


Figure 6.7 Training and validation error for NRLS-HES Neural Network

Mean absolute error (MAE) and mean-squared error (MSE), with MAE used for training via backpropagation

6.4.3 Conclusions for neural networks

The fitting process for the feed -forward network was relatively straight-forward, took minimal time and showed reasonable performance. Neural networks did not show better performance than the boosting, random forests or GAMs, but this may be due to the limited training data. Neural network commonly perform best on large datasets (Chollet and Allaire, 2018) and the 80% training set may be sub-optimal in terms of size. Models performed best with several layers decreasing in node numbers with each layer. This suggest that the network was able to successively simplify non-linear relationships down towards final predictions. Training error was higher when numbers of nodes increased from one layer to the next, suggesting that projection to higher dimensions did not aid predictions and data may have become too sparse. This may have overfitted data in a similar fashion to the multi-TPRS smoothers.

The training of neural networks, although simple to code using `Keras`, is essentially a trial and error process with several options to improve fit. High performing Neural Networks take time to train and test a given architecture/layers/neuron numbers. The time taken to optimize a network like this must be justified by the scale of the data, complexity of the data/question, or specific requirements for non-linearity. The NRLS aggregated data, whilst non-linearity is of concern, does not appear to justify this approach.

6.5 Conclusions and comparisons with GLMM

The models fitted in this chapter further address aim 3 of the thesis, and were used to improve upon predictions presented using GLM and GLMM methods presented in Chapter 5. Table 6.3 compares the best MAE performance in each of the model classes, excluding the single-level mixture models due to their increased bias compared with Poisson GLMs. Random-intercept GLMMs clearly reduce prediction error and strongly support the clustering assumptions related to the hospital-level repeated measures, and this effect is also included in the GAM models. GAM showed improved prediction error compared with GLMMs, suggesting that the non-linear/smooth terms better capture the relationships within the data. They are more likely to reflect the underlying relationships within the data, but given the noise and the aggregation, this may be an artefact. With a larger dataset, relationships may have been smoother and the difference between the models minimised. A larger dataset would also have been advantageous for the Neural Network models. This was difficult to examine in practice as, although prediction error was tested on 2016/17 data, step-changes in the organisation-level random-effects were large, and using a five-year training set led to noisier data and poorer models (see Chapter 5.8).

Our aim in these models is to fit the underlying ‘average’ relationship of incident reports and predictors to allow generalization to other datasets without overfitting. GAMs smooth out noisy predictors so are suitable for this task. Trees are not concerned about fitting a theoretical distribution and split on covariates that explain the most variance. If variance is constant, the noise is random (normally distributed) or proportional to the exposure size (size of hospital), trees will fit regardless.

Method	MAE	
	2015/16	2016/17
Poisson GLM	138.78	143.46
NB2 GLMM	55.34	101.68
NB2 GAM CR	58.35	97.22
NB2 GAM TPRS	57.98	96.68
NB2 GAM Multi TPRS	46.95	111.58
Boosted Tree	16.28	108.30
Random Forest	23.94	102.60
Neural Network	59.40	114.83

Table 6.3 Comparison of GLM, GLMM, GAM, Tree and Neural Network prediction errors

The best performing models based on testing MAE are presented. 2015/16 is the training set and 2016/17 the testing set.

The NB2 GAMs appear to show the best fit of all the models applied, as they:

- Reflect the clustering with an organisation-level random-intercept
- Smooth noisy predictors in a non-linear fashion, towards a more general relationship
- Account for the aggregation in the NB2 variance scaling, giving more weight to the assessment of overdispersion at organisation with smaller conditional means. This fits with the aggregation assumptions as a single incident is a larger proportion of the mean at an organisation with 100 bed-days than one with 1000 bed-days.

Algorithmic methods, particularly random forest and boosting showed good performance without explicitly reflecting the clustered nature of the dataset. Random Forests showed only marginally higher testing error than the NB2 GLMM. Some of this may be due to de-correlation effects of the random forest but may also suggest that, given the noise in the data, there are several possible solutions. Random Forest models may produce biased predictions on the testing set if the correlation structure is notably different to the training set. A potential solution to this is to combine both the random forest and random-intercept approaches. This has been demonstrated in one proof of concept paper (Hajjem et al., 2014), but has yet to gain acceptance and wider implementation in standard statistical software. The authors of the paper kindly provided R code to attempt this approach, but its non-standard implementation was not fault tolerant and could not be applied within the timespan of this project. Adapting random forests to sample within strata, and make use of their bootstrapped structure, is a subject for further research.

A useful next step with algorithmic models was to include organisation as predictor, similar to the inclusion of organisation as a fixed effect as discussed in the conclusion from Chapter 5. Regression trees fitted in this way simply split by total bed days and organisations with distinct effects. These models increased overfitting and degraded the performance of boosted models. It is likely, however, that bed days are a reasonable proxy of the organisation code and that correlation is the issue. Random Forests were then investigated to attempt mitigate the correlation problems. RFs presented a further complication, as Breiman and Cutler's fortran code is limited to categorical predictors with no more than 32 levels, corresponding to 2^{32} splits. This is computationally intractable in this form, and but an alternative representation in Java-based machine learning library 'H2O.ai' (H2O.ai Team, 2018) allows this by representing predictors as histograms. H2O random forest were compared with the models discussed in this chapter as a validation, but their application with organisational predictors did not increase predictive ability of the models.

Many of these methods discussed in this chapter share similar principles, as described by Hastie & Tibshirani (Hastie et al., 2009a). Methods included projections into higher or lower dimensions, choosing and minimising an appropriate loss function, data splitting, and validation on an out-of-sample set (or cross-validation). These techniques, to a lesser or greater extent also suffer from the same pitfalls: being somewhat 'blackbox' to many practitioners. This can be overcome in some senses by simulation methods or the use of frameworks such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016). Overfitting and lack of representation of the clusters structure in the data is also an issue. However, given the lack of cluster representation these methods perform admirably, in terms of minimising prediction error through their own mechanism, leading to an important question about whether the cluster representation is required?

In Chapter 8, GLMM, GAM and random forest techniques will be applied to the data to derive standardised indicators for use in surveillance and monitoring from both regulator and organisational perspectives, addressing aim 4 of the project. Differences in these models, and the organisations they identify, will be discussed and reveal potential biases of random forest models. Models will first be applied to a specific sub-set of severe harm or death incidents in Chapter 7.

Chapter 7 Development of death or severe harm models.

7.1 Introduction

The models developed in the previous chapters have focussed on total incident reports. These are a mixture of reports at different harm level, only some of which are mandatory. The current national publications on NRLS (see Chapter 8) focus on total incident reports, but also on the incident reports that are mandatory: those that lead to severe harm or death. This chapter continues to address aim 3 by applying the model structures used in previous chapters to the death and severe incident reports only, discusses the issues relevant to repurposing the models in this fashion, and interpretation of the outputs.

7.2 Development of death/severe harm incidents model

The analytic framework based on the total incident reporting models was used as the starting point for developing a model for death and severe harm incident reports (DS). Using the same predictor dataset is a reasonable starting point, as model training would identify different model coefficients for DS incidents, compared to total incidents.

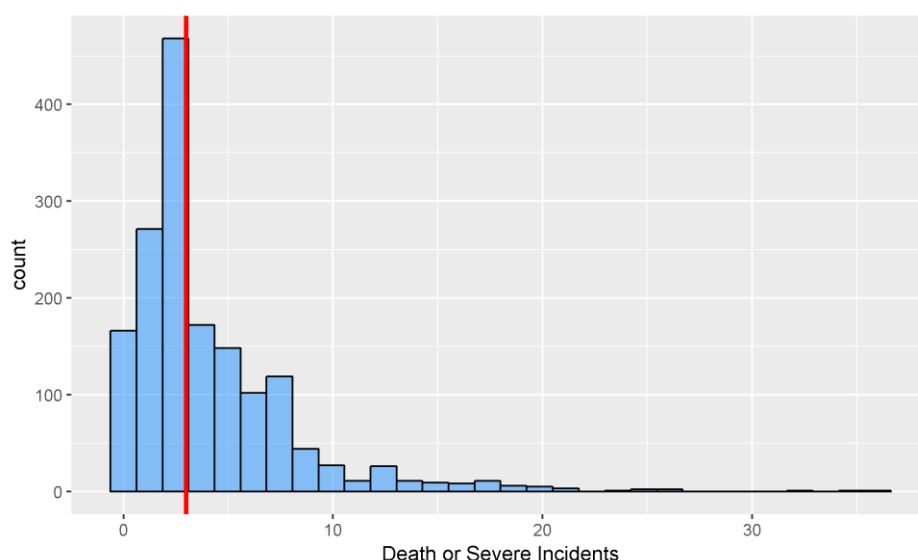


Figure 7.1 Histogram of counts of death or severe incident reports per trust, per month

Death of severe harm incident reported in fiscal year 2015/16. Red line represents the median value (3).

Figure 7.1 shows a summary of the response variable. DS incidents are very rare in comparison to all incident reports, representing <1% of the annual reports. Breaking them down by trust and by month gives a median value of 3 DS incidents per trust per month. This poses the problem of sparseness for a predictive model, as there is very little evidence to model. The impact of prediction error is also proportionally higher given that a single case represents 33.3% of the median value.

Poisson models were initially considered, using the random-intercepts described in previous chapters for clustering. Two additional predictors were included in the model: total incidents and non-mandatory incidents, both centred and standardised as discussed in Chapter 4. In all cases, the models with total incidents as a predictor out-performed those with non-mandatory reports, suggesting that non-mandatory was an incomplete picture compared to using all incidents reports. The non-mandatory incident predictor was then dropped from the model.

Overdispersion was also present in these models, with a dispersion ratio of 1.43 for the Poisson GLMM base model. This was much lower than the dispersion ratio of the total incident reports Poisson GLMM model, and suggested less noise in the data. Though events were sparse, these reports are likely to be more objective and consistently observed due to their 'extreme' outcomes. Alternative model distributions for NB1, NB2 and Tweedie (Dunn and Smyth, 2005) (with power estimated by modelling procedures) were fitted, but with such low counts, it was logical to consider zero-inflation (as briefly mentioned in Chapter 4, in the context of overdispersion) (Cameron and Trivedi, 2013e). In zero-inflated Poisson (ZIP) models, a Poisson model is fitted for non-zero counts and an alternative binomial model is fitted for the probability of zero counts. These models therefore assume a specific mechanism dictates the probability of a zero count, and it is not simply a Poisson mean. This may be due to reporting system quirks or other cultural issues at organisations that dictate the likelihood of reporting an event. However, this is unlikely due to the mandatory reporting of these incidents. It is possible that an unknown mechanism exists for zero-inflation, so this was formally tested in the data with the score test described by van den Broek (1995). The test suggested zero-inflation in the dataset, but the mechanism that might cause zero-inflation was unclear, and may be an artefact of the low counts, noise, or aggregation. ZIP models were fitted with three alternative zip formulae: all predictors, total bed-days, outpatient and A&E attenders, and just total bed-days. ZIP models were no clear improvement on the Poisson models in terms of prediction error, with training MAE ranging from 1.71 – 1.73 and testing error ranging from 2.26 – 2.36.

Inspection of parameter estimates in these full models suggested most parameters were not statistically significant at 95%. Given that the effects of overdispersion on these estimates

would have been to overstate their significance, it suggests that many of the predictor variables were not aiding the model at all. The models were therefore over-parametrised, and performance may have degraded because of this. Model bias and variance are commonly described as a 'trade-off' (Hastie et al., 2009b), where decreasing variance commonly increases bias and vice-versa. E.g. NB models reduce variance in comparison to Poisson models, but at the cost of some bias in predictions. This trade-off suggests that the models may be biased from overfitting parameters to reduce variance. Simplification of the models was the next logical step to see if bias could be reduced. Rather than eliminate parameters using automatic step-wise techniques or shrinkages methods such as ridge regression or LASSO (Tibshirani, 1996), models were instead rebuilt pragmatically using only the strongest predictor variables.

Simplified model parametrisations were constructed as GAM and GLMM models using Poisson, NB1, NB2. The following parametrisations were applied to all models, with cubic regression splines used for GAMs. Parameterisations will be referred to by number for the rest of this chapter. Parameterisations were:

1. Random-intercept for trust, teaching hospital status, seasonality spline & total incidents counts
2. Random-intercept for trust, seasonality spline, total incidents count, total outpatient attenders & total A&E attenders.
3. Random-intercept for trust, teaching hospital status, seasonality spline, total incidents count, total outpatient attenders & total A&E attenders
4. Random-intercept for trust, teaching hospital status, seasonality spline, total incidents count, total outpatient attenders, total A&E attenders, count of bed-days of admission, count of bed-day for patient admitted by transfer, count of bed-day for non-elective admissions, count of bed-day for maternity/birth admissions, count of bed-day for patients aged 70-84, count of bed-day for patients aged 85 and over, count of bed-day for patients with Charlson comorbidity score of 0, count of bed-day for patients with Charlson comorbidity score of 5 or greater, count of outpatient attenders age 50-69 and count of outpatient attenders aged 70 or over.

Model prediction errors for the best performing models are shown in table 7.1, with the simplified Poisson (parametrisation 1) shown at the bottom of the table for contrast. A full table of model prediction error outputs for 60 models is presented in Appendix C.5. Prediction error on testing sets was very close between all the models tested, varying by 0.52 between highest and lowest MAE values. Variation between models was higher in the training set, with 0.92 between the highest and lowest MAE values.

Description	Model Class	Family / Distribution	Training MAE	Testing MAE
Full Model	GAM	NB2	1.908	2.055
Full Model	GAM	Poisson	1.772	2.078
Parameterisation 1	GLMM	NB1	1.742	2.092
Parameterisation 2	GLMM	NB1	1.736	2.098
Parameterisation 1	GLMM	Poisson	1.733	2.121

Table 7.1 Mean absolute prediction error for death or severe harm NRLS-HES model

Columns relate to model class and distribution in training (2015/16) and testing (2016/17) datasets. Model descriptions indicate model parameterisation. 'Full' indicates all predictors from earlier incident models plus total incidents. Bottom row of the table shows this simplified Poisson parameterisation as a benchmark for comparison.

The similarity of MAE performance between the full and simplified models suggest two things: the full models captured a degree of variation that is not well represented in the simpler models, but also that the major predictive value in the data rests with seasonality, the random-intercept and the total incident reports. The added complexity of including all predictors may not be justified. The full models that performed best were GAMs with selection and additional BIC-like smoothing penalties (detailed in Chapter 6), whilst the most predictive reduced models were the simplest (parameterisations 1 & 2) GLMM models. Figure 7.2 plots the relationship of the major predictors in the simplified models. The effects of low number are stark, and although a slight upward trajectory is discernible as total incidents increase, the variance is high. It also appears that teaching hospitals are generally larger in size and reporting more incidents in total, but death or severe harm incidents show only a slight increase with size. When 'total incidents' and 'total inpatient bed-days' were both tested in the simpler parameterisations 'total bed-days' was no-longer significant once 'total incidents' entered the model. When just total bed-days was fitted without 'total incidents', it appeared significant. This suggests that there is a confounding/proxy effect for total incident reports that is approximated by the size of the organisation/total bed-days. This is an intuitive result, as we would expect larger organisations (with more bed-days) to report more incidents to their greater exposure.

The relationships in Figure 7.2 do not give a clear picture on increasing DS incident reports. On the one-hand, they could be interpreted that the number of DS reports increases with the total number of reports. However, the low rate of change in the smoother suggests that there is little correlation between total incidents, and more erratic changes tend to occur in smaller non-teaching hospitals.

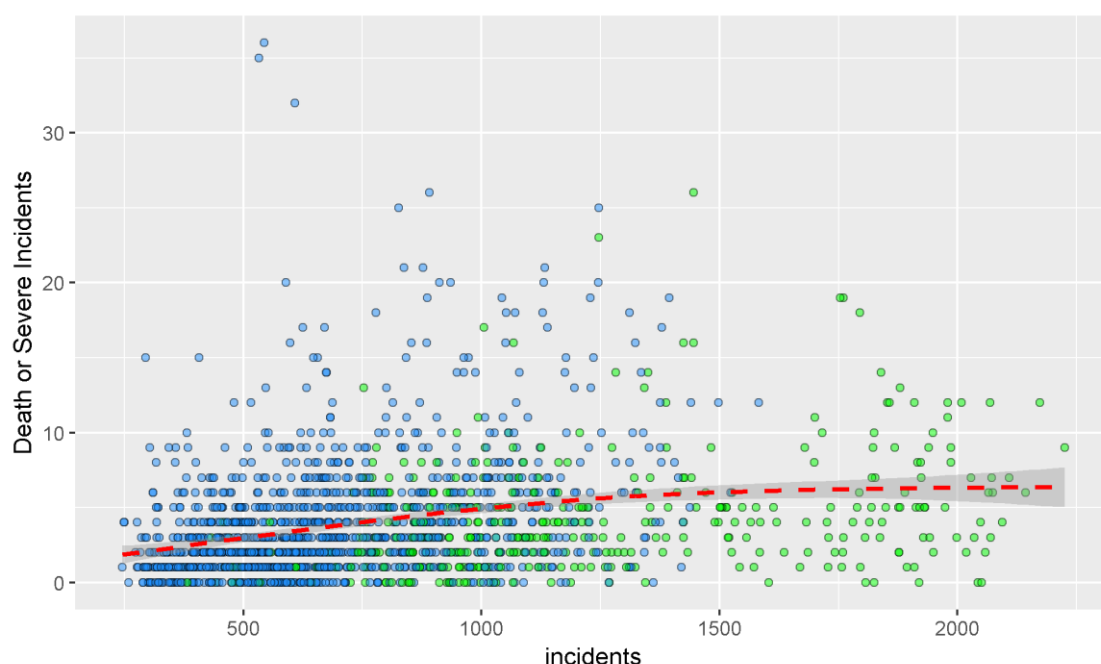


Figure 7.2 Relationships between death or severe, and total, NRLS incident reports

Blue points represent non-teaching hospitals and green represent teaching hospitals. Red line is a smooth GAM fit of the relationship. NRLS data for incidents reported in 2015/16.

7.3 Conclusions

DS incidents can be modelled in a similar manner to the total incidents reports models, but fewer of the predictors are required to model them well. Good performance can be achieved through replicating the important structural features, namely the clustering/repeated measurements at trusts, seasonal fluctuations and the total number of incidents reported (that can be assumed to be a proxy of the size of an organisation). It is likely that, given the sparsity in the models, fewer of the predictors can be calibrated due to the low signal-to-noise ratio. Other predictors are likely to be dominated by the size predictor. Reduced models showed similar performance to models with large number of predictors and a case for ‘Occam’s razor’ could be made, selecting the simpler models.

NB1 distribution showed good performance in these models. When compared to the performance of NB2 models for all incidents, it suggests conclusions about overdispersion and scaling. Overdispersion was much lower in DS models. NB2 models worked well for all incidents as they weighted the effects of overdispersion proportionally higher in small organisations, whereas NB1 models weight all trust in proportion to their conditional mean. The ‘size’ in question here is the ‘ γ ’ value in the model. DS incidents are rare occurrences, with a median of 3 per organisation per month. Secondly, the range of scaling applied by NB1 is much smaller than the range of the ‘ γ ’ in the total incident reports models. NB2 GAM

models appear to have performed best on testing set, but may have introduced bias on the training set.

Although testing error was poorer in Poisson GLMM models, the Poisson's unbiased nature is a useful benchmark for comparison. The Poisson GLMM displayed lower training error, but slightly higher testing error. This is somewhat unexpected, as it suggests that the Poisson model is less biased in the training set but may be mis-calibrated for the testing set. The reduction in the overdispersion from the NB1 parameterisation may have reduced training error in terms of reduced variance, but at the cost of bias.

These results in all models should also be viewed in the light of the sparsity in the model, and the differences in MAE results being minimal. This leads to several question: will services have any ability to learn from such rare events, and does it add any value to monitor them in such a way? Lilford et al.(2010) suggested that *"Changes in clinical outcome (such as mortality or infection rates) can never be bigger than changes in the clinical error rates on which they depend and are usually much smaller; it is rare for the risk of an adverse outcome to be wholly attributable to clinical error."* The ability to learn from such incidents can only be assessed over larger samples, and national level analysis would seem to be the more appropriate resolution for this. This is very different take to that of learning from each error in clinical audit.

The models reported in table 7.1, as the best performing models on testing set error, will be taken forward for fitting as standardised ratios in Chapter 8.

Chapter 8 Developing a risk-adjusted indicator for NHS regulatory use

This chapter focuses on how the risk-adjusted prediction models, developed in previous chapters, can be used by current UK healthcare regulators. It addresses aim 4 of these, and first describes the current regulatory use of NRLS, major stakeholders, and proposes an alternative indicator, the Standardised Incident Reporting Ratio (SIRR), and shows how this fits with current systems for cross-sectional analysis and longitudinal modelling.

8.1 Methods for UK health regulators

8.1.1 Regulators

Healthcare, and the NHS, are highly regulated industries in the UK with multiple lines of accountability and a degree of overlap between regulators. This chapter focuses on the role and remit of these organisations and their use of NRLS for learning and safety monitoring. The NHS is an evolving collection of organisations and regularly subject to political intervention. It has had several regulatory bodies and functions during the lifespan of the NRLS, that have been merged or been replaced over the period. When starting this PhD project, the NRLS was managed by the National Patient Safety Agency (NPSA), who held responsibility for NHS patient safety development and monitoring. The NPSA was abolished and the functions combined with NHS England for a period (Parliament, 2012). Patient safety functions were later transferred to NHS Improvement. At the time of writing, and in terms of national focus on the acute hospital sector, the major NHS regulators are:

- **NHS Improvement (NHSI):** Formed by a merger of the former economic and quality regulators: Monitor and the Trust Development Authority; NHSI has roles in setting national payment tariffs, leadership in areas of productivity and efficiency, collecting NHS costing data, and a role providing intensive support and oversight for struggling organisations. Patient safety is also a key theme of their portfolio, with dedicated teams focussed on patient safety interventions and learning from error. NRLS is 'owned' and managed by NHSI, who have provided data access for this project.
- **Care Quality Commission (CQC):** An independent non-governmental organisation, and successor to the Healthcare Commission, with a statutory duty to monitor and

regulate health and social care providers in the UK. Their work applies to social care, primary care, secondary care, ambulance services and private organisations providing health or social care. The CQC's purpose is to assure care is safe, effective, compassionate and high-quality. Their roles include registering, monitoring and inspecting providers, rating organisations and taking action to protect service users. The CQC makes regular inspections of organisations, but also runs an intelligence-led, targeted inspection programme, and a specific programme for mortality 'outliers.' CQC have 'Intelligence and Insight' teams who produce packs of information for CQC central functions, inspectors and trusts. They therefore have a major role in monitoring information to identify variation and potentially trigger inspections.

- **NHS England (NHSE):** The national NHS executive organisation, headed by the chief executive of the NHS. NHSE has a wide variety of responsibilities for planning and overseeing the NHS' work including emergency planning and future strategy. This includes managing national and regional teams, 'specialised commissioning' (for complex, high-cost or specialist tertiary services), monitoring of national targets such as waiting times, leading or developing new models of care, and technological development (although a tension exists between them, NHS Digital, and the newly formed 'NHSX' on the later).
- **The Department of Health and Social Care (DH):** The DH is the civil service organisation supporting health ministers to implement the Government's agenda for the NHS. Much of its former responsibility was devolved to NHSE in the Health and Social Care Act 2012, but it holds influence over funding and policy agenda in the NHS. Under the tenure of the previous Secretary of State for Health, the Rt Hon. Jeremy Hunt, the patient safety agenda was highlighted, encouraging incident reporting, human factors analysis and wider safety programmes.
- **Medicines and Healthcare Authority (MHRA):** The MHRA regulates the use, trial and implementation of medicines, devices, and more recently, computer/mobile apps. They regulate packaging, labelling and legal compliance. They run the yellow card scheme for drugs based adverse reaction (discussed in Chapter 2).
- **National Institute of Health and Care Excellence (NICE):** NICE's role is to produced evidence-based guidance on treatments and interventions, including safety and cost effectiveness, developing quality standard (such as safe staffing levels) and providing information to commissioners and practitioners.

8.1.2 Current users of NRLS incident reporting data

Whilst the NRLS is used for research and special requests, such as this project, the major national users of these data, and their associated publications and indicators, are:

- **NHSI** (with all indicators available on their website: www.improvement.nhs.uk):
 - Six-monthly data releases referred to as the National Patient Safety Incident Reports (NaPSIRs). Published as spreadsheets, these releases contain numbers of incidents reported monthly, presented in various categorical groupings including 'care setting' and 'incident type' (NHS Improvement, 2017b).
 - Six-monthly Organisational Patient Safety Incident Reports (OPSIRs). Published as spreadsheets, these books contain organisational reporting rate data, based on incidents per 1000 bed-days (using the KH-03 bed-days, as described in Chapter 5, or similar activity units for other care settings). Reported figures include the median time between incident occurrence and reporting to NRLS and reporting rates in harm-level categories. Data quality notes are provided, and incident data are presented in relation to both the date they were reported to the NRLS and the date incidents occurred (NHS Improvement, 2017c).
 - NRLS Explorer tool is a web interface to generate organisational reports that detail splits of incident harm levels, changes in incident reporting rates, and potential under-reporting of incident reports at local organisations.
 - NRLS data may be used as part of special publications to support/develop particular patient safety programmes.
 - Patient Safety Alerts are sent to organisations via the central alerting system (CAS) as emerging themes are identified. These alerts may be driven by, or investigated using, NRLS data.
 - General patient safety work at NHSI may use the NRLS, but this does not necessarily lead to specific publications.
- **Care Quality Commission** (some indicators externally published, others extracted from 'Insight' reports sent to UHB):
 - CQC's 'Insight' program publishes data packs that are used to inform hospital inspection, based on their 'Key Lines of Enquiry' (KLOE) (Care Quality Commission (CQC), 2017). No formal methodology publication has been released for the Insight programme, but Insight packs contain a limited set of notes to describe indicators. The inspection programme prior to 'Insight' was

referred to as 'Intelligent Monitoring,' and most indicators appear to be common between the two (Care Quality Commission (CQC), 2014b, Care Quality Commission (CQC), 2014a). Insight packs are sent to trusts as regular monitoring information, and include the following indicators as part of KLOE 5 and 6:

- **KLOE 5 NRLS** - Proportion of reported patient safety incidents that are harmful (%);
 - **KLOE 6 NRLS** - Consistency of reporting, National Reporting Learning System (NRLS) – National;
 - **KLOE 6 NRLS** - Potential under-reporting of patient safety incidents resulting in death or severe harm;
 - **KLOE 6 NRLS** - Potential under-reporting of patient safety incidents.
- 'Intelligent Monitoring' the predecessor to 'Insight' with published formal methodologies (Care Quality Commission (CQC), 2014a) included the following indicators with more detail than 'Insight':
- **NRLSL03:** Proportion of reported patient safety incidents that are harmful (%), with numerator of total incident reports that cause harm, and denominator of the total incident reports.
 - **NRLSL04:** Potential under-reporting of patient safety incidents resulting in death or severe harm, a standardised ratio with numerator of the count of severe harm or death and denominator an expected number of incident reports based on trust's bed-days multiplied by the national average ratio. Bed-days are calculated from HES, with day cases treated as 0.5, but no description of how 'zero-day' (non-daycase) stays are counted (see Chapter 5). Specialist trusts, such as Children's hospitals, are excluded due to their radically different casemix.
 - **NRLSL05:** Potential under-reporting of patient safety incidents, as per NRLS04 but applied to all incident reports, not just severe harm or death.

The major uses of NRLS data at regulatory level are therefore: analysing the free-text for specific themes and comparisons of reporting rates with some adjustments for the sizes of organisations. The models developed in the previous chapters have shown an approach for total incident reports and death or severe incident reports that will be used later in this chapter to develop standardised ratios that can be used by NHSI, CQC, and trusts to examine

their reporting behaviour. Current indicators do not make adjustments for case mix or “exposure”, and SIRR models can fill this gap.

8.2 Creating a standardised incident reporting ratio (SIRR)

Comparisons of incident reporting at organisations cannot be done directly using parameter estimates, as models built in Chapters 5-8 did not include ‘organisation’ as a fixed effect/predictor. Comparisons of random-effects estimates are possible with the GLMM and GAM approaches used in earlier chapters, but these techniques are not common-place in current NHS monitoring schemes. Many current indicators, such as mortality, readmission or LOS can, instead, use the model predictions as outputs for comparison (Bottle and Aylin, 2008). We have used predictive values to assess the error of models, but they can also be used for reporting. Ratios of the observed events to predicted events are described as indirectly standardised ratio (Breslow and Day, 1980):

$$\text{Standardised Ratio} = \frac{\sum \text{observed events}}{\sum \text{predicted events}}$$

A ratio of one indicates that observed and expected are equal, below one indicates fewer observed than expected, and above one indicates more observed than expected. This technique does not explicitly represent clustering and assumes that data points are independent. Later sections of this chapter will address clustering and the conflict inherent in using these methods with outputs from multilevel models.

In the case of NRLS models, two standardised ratios will be calculated, one for the ‘all incident report’ models and one for the ‘death or severe harm’ (DS) models. These ratios are a good fit with the CQC monitoring methods discussed in the following section.

8.3 Monitoring techniques used by CQC

The CQC's role is to regulate organisations using inspections and a risk/compliance framework. The aim of the risk/compliance framework is to inform/trigger an inspection and monitor organisations between regular inspections (Bardsley et al., 2009). The Insight reports mentioned above form the basis of this, generating 'key lines of enquiry.' Much of the history of this framework comes from the tragic circumstances surrounding hospital failures such as events at Mid-Staffordshire hospitals (Francis, 2013) and cardiac surgery at Bristol (Spiegelhalter et al., 2002). Some of the techniques associated with mortality monitoring, including the use of process control methods, and charts for comparing indicators, and monitoring change over time, have been developed in an authoritative source, presented to

the Royal Statistical Society, and adopted for CQC's monitoring (Care Quality Commission (CQC), 2014b, Care Quality Commission (CQC), 2014a, Spiegelhalter et al., 2012c). These techniques are inspired largely by meta-analysis methods for dealing with heterogeneity in clinical trials as described by DerSimonian & Laird (1986). In meta-analysis, a reviewer is aiming to summarise the effects of several different clinical trials, usually of different sizes, without the original data. This is conceptually similar to comparing many hospitals across different indicators using aggregated indicators, rather than record-level data.

These processes use several techniques that are relevant to SIRR models and will be applied in the rest of this chapter. The techniques include (Care Quality Commission (CQC), 2014b, Bardsley et al., 2009, Spiegelhalter et al., 2012c):

- Comparison against expected ratios using Funnel Plots
- Transformation to z-scores
- Estimation of overdispersion using an additive model and adjusting z-scores and funnel plots accordingly
- Time series monitoring using 'Cusum' charts

Much of the following section is adapted from these sources, as the aim here is to fit with the current regulatory systems.

8.3.1 Comparison using funnel plots

Comparisons of mortality indicators such as the Hospital Standardised Mortality Ratio(HSMR) (Jarman et al., 1999), have been criticised for (amongst other things), their misinterpretation. Some of this stems from early presentation of mortality ratios as league tables, based on indirectly-standardised indicators (Lilford et al., 2004) and the tendency to interpret them as summary markers of 'good' or 'bad' hospitals (Lilford and Pronovost, 2010, Black, 2010).

The two major problems of indicators presented in this fashion are that they do not reflect the uncertainty in the data, nor do they inform readers whether a given organisation requires investigation or not. Funnel plots have been proposed as an alternative for representing this in a visual manner (Spiegelhalter et al., 2012c, Spiegelhalter, 2005a, Egger et al., 1997). These plots are rooted meta-analysis techniques comparing studies of different sizes (Goldstein and Spiegelhalter, 1996). These charts plot the indicator of choice on the y-axis and a measure of the size of the unit on the x-axis (Figure 8.1).

In application to hospital mortality or other measures, the x-axis commonly presents the number of expected events from the indirect-standardisation, or some other measure of

organisation size such as the number of discharges. This measure of size is commonly a count so can be regarded as Poisson distributed. The fixed variance=mean relationship of the Poisson distribution (see Chapter 4) allows the construction of standard control limits that come together in a funnel shape as the standard error decreases as the sample size increases. These limits can be overlaid onto the scatter of the ratio against the size measure (Spiegelhalter, 2005a).

The term 'Control limits' refers to a concept from statistical process control (Mohammed et al., 2001, Benneyan et al., 2003) that are used to define boundaries between 'common-cause variation' (natural variation) or 'special-cause variation' (systematic differences/greater than natural fluctuation). Control limits are designed to drive action, and a data point outside of the funnel is showing special-cause variation and should trigger investigation of that data point. A data point within the funnel is showing common cause variation and should be regarded as 'in-control' or within the expected natural variation.

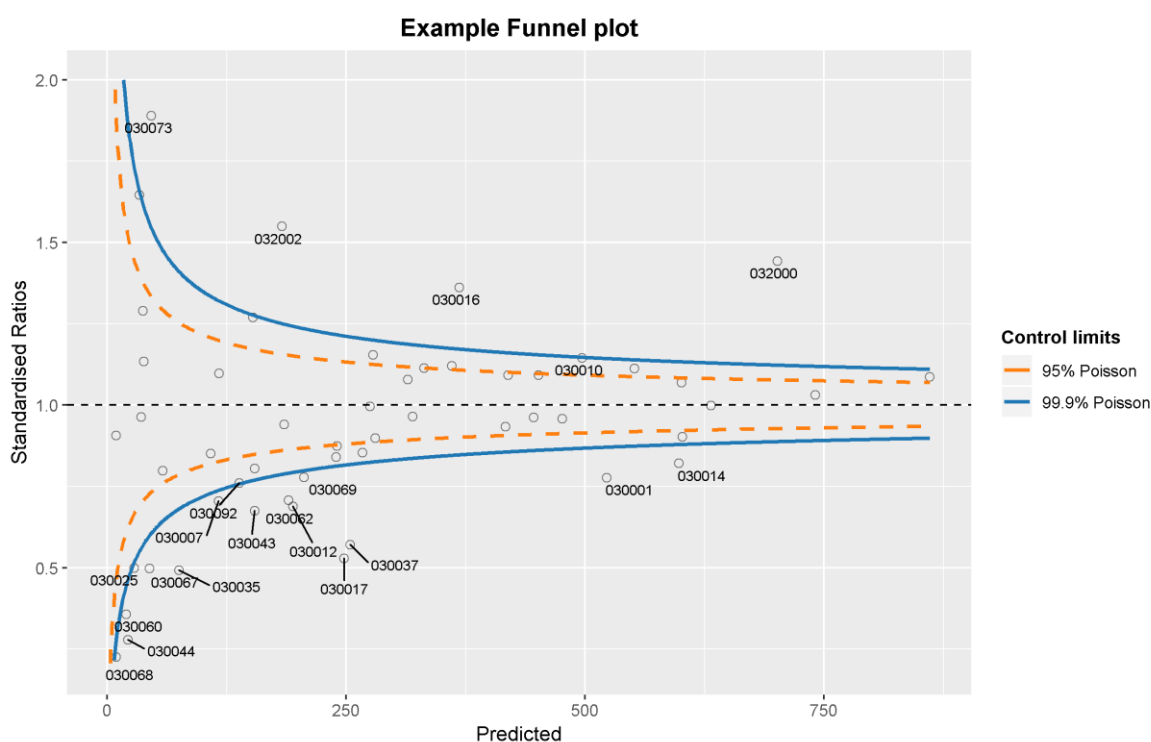


Figure 8.1 Example funnel plot using the 'medpar' dataset

Data are reused from the `COUNT R` package (Hilbe, 2016), originally collected from Medicare data for Arizona in 1991, discussed in (Hilbe, 2014). A standardised ratio has been created (y-axis) as the sum of observed LOS / the sum of predicted LOS (model predicted probabilities), built from a Poisson regression model, with the x-axis showing predicted LOS, and data points representing organisations.

Funnel plot limits can be calculated in several different ways, as there are various methods for approximation and exact limits, including:

- Using the Poisson density function we can compute exact Poisson limits. Due to the discrete nature of the Poisson distribution, this is not defined between integers, and requires an interpolation procedure for x-axis values that are not integers (Spiegelhalter, 2005a).
- Using the relationship between the Poisson and Chi-squared distributions, a chi-squared density function can be used to calculate an exact Poisson limit that is defined between integers, and therefore does not require interpolation (Ulm, 1990).
- Using transformations such as square-root (Care Quality Commission (CQC), 2014a, Spiegelhalter et al., 2012c) or natural logarithm (Clinical Indicators Team, 2016b), we can calculate an appropriate standard error on the transformed scale and back transform to our original scale. This is the approach used for expanded funnel plot limits with an additional variance component (see section 8.3.3).

Funnel plots based on Poisson limits suffer from the same overdispersion issues discussed in previous chapters. This means that, in the presence of overdispersion, the control limits are too conservative and too many points are outliers, as seen in Figure 8.1. Spiegelhalter et al.'s (2005b) paper proposed a correction to this to account for the clustering using random-effects. This uses the same theory as the inclusion of a random-intercepts used in GLMMs in previous chapters but calculated using a different mechanism. This method is applied post-hoc and is a somewhat blunter approach than estimating random-effects within a model. This mechanism is described below, and the resulting τ^2 can be added to the standard error used to draw the control limits and expand them to account for overdispersion.

Funnel plots are a clear representation of standardised ratio data that factor size and uncertainty into the plot. Although they contain more information than a simple league table, funnel plots do still have a degree of ranking that may lead to stigma or misinterpretation (Lilford et al., 2004). This is not an argument to abandon comparison, but an argument for comparing organisation on multiple indicators rather than one in isolation (Keogh, 2013).

8.3.2 Transformation to z-scores

When comparing indicators or viewing many of them, it is helpful to consider the data types and the differing natural scales. It would be challenging to compare infection rates, mortality, length-of-stay, incident reports etc. when they differ in scale. It is desirable from the regulator's perspective to have a single unified framework to compare indicators. The form adopted by CQC was the 'z-score' (Spiegelhalter et al., 2012c, Care Quality Commission (CQC), 2014a). Z-scores are standard scores that follow from the normal distribution. In normally distributed

data, the mean is central, and the standard deviation is the average distance of points from the mean. Figure 8.2 shows a theoretical normal distribution with density on the y-axis. The distribution is centred on the mean at zero and the x-axis corresponds to the number of standard deviations (z-scores) from the mean. The properties of the normal distribution are such that approximately 95% of data lie between two standard deviations below and above the mean. Similarly, 99.8% lie between three standard deviations above and below the mean (Altman, 1990).

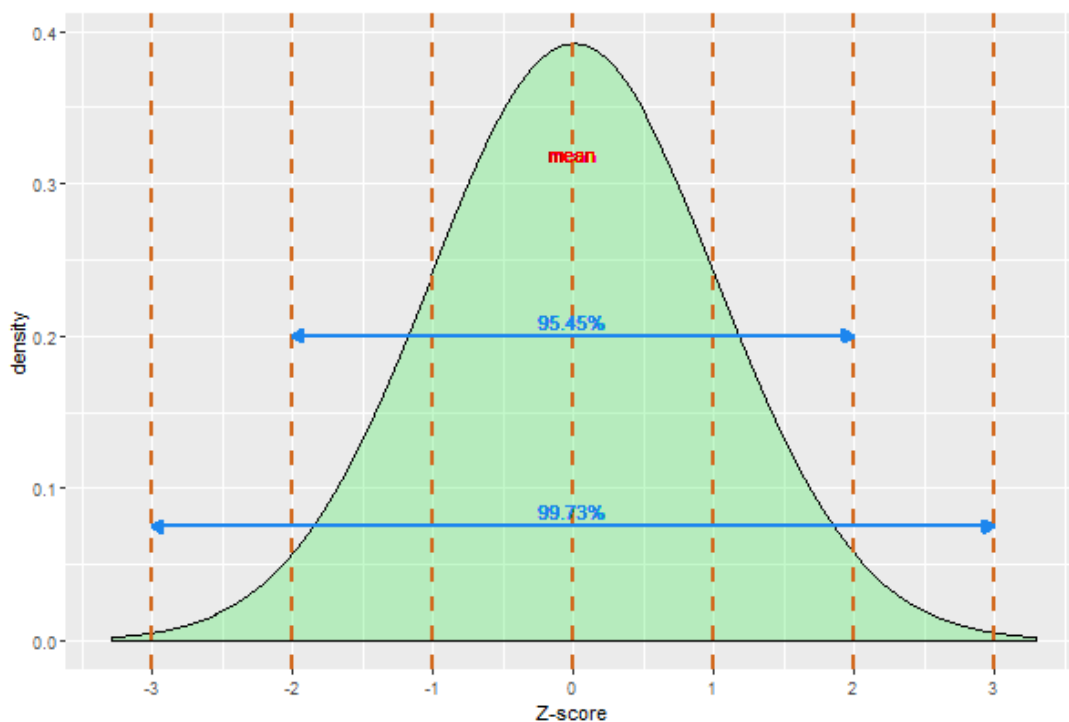


Figure 8.2 Theoretical normal distribution and z-scores

Y-axis represents density and x-axis the 'z-score', or number of standard deviations from the mean (0). Blue lines represent the central regions between -2 & 2, and -3 and 3 standard deviations and the percentage of values within this range.

It follows that, if we centre a normally distributed dataset on the mean and divide by the standard deviation, we are now scaled such that a value of 1 represents 1 standard deviation above the mean and -2 represents 2 standard deviations below the mean, and so on. This is the same process described as 'centring and scaling' for regression model input, discussed in Chapter 4. Therefore, all normally distributed indicators could be compared on this z-score scale, and statistical significance determined by the value of the z-scores.

The further complication is that our data are rarely normally distributed. Various transformation have been recommended (Spiegelhalter et al., 2012c), rendering them 'more

normal' before transforming to z-scores. This means applying a function of some kind to our data before we regard it as 'Y.'

The CQC guidance (Care Quality Commission (CQC), 2014b) for constructing z-scores uses three specific transformations depending on indicator type. Each transformation assumes a target value T, that is appropriate for that indicator type.

For Standardised Ratios (SRs), such as HSMR, SHMI and the proposed SIRR, an 'expected' value is calculated by indirect standardisation. For SRs, $T=1$ (observed = expected), and this corresponds to the idea of the national average. SRs are then square-root transformed before z-scoring:

$$Y = \sqrt{\frac{\text{observed}}{\text{expected}}} \quad \text{with standard deviation: } s = \frac{1}{2\sqrt{\text{expected}}}$$

The transformed z-score is therefore:

$$z = \frac{Y - 1}{s} = 2(\sqrt{O} - \sqrt{E})$$

The standard error in these calculations is then interpreted as the within-trust standard deviation. The SIRR methods based on the regression models will all be represented as standardised ratios. Although the CQC methods recommend a square-root transformation, log-transformation has also been suggested and adopted in the national mortality indicator SHMI (NHS Digital, 2017f). This was described as too severe an adjustment in many cases (Spiegelhalter, 2005b), but appears to aid normality in the SIRR models (see section 8.4.2).

8.3.3 Estimation of overdispersion using an additive model

Overdispersion has been discussed at many points for these models, but the approach of Spiegelhalter et al. (2012b) is to assume an additive overdispersion model, rather than multiplicative one such as the quasi-likelihood or negative binomial models applied in earlier studies (Marshall et al., 2004), and in earlier chapters of this thesis. This is conceptually identical to the random-intercepts estimated in the GLMM models, but its estimation takes a different form. It is estimated directly from the data, adjusting for the cluster sizes, but does not affect the expected values directly. Firstly, the dispersion ratio ϕ is estimated. The dispersion ratio is a measure of the ratio of the model error over the degrees of freedom and is approximately χ^2 distributed on the degrees of freedom. Its estimation is not, however, performed on the full dataset to avoid the undue influence of outliers. Z-scores are first 'Winsorised' before estimating ϕ .

Winsorising is the process, named after statistician Charles P Winsor, of scaling the extreme values in a distribution on the basis that they may not be representative. The mean and standard deviation of a distribution are strongly affected by outliers, and Winsorisation attempts to reduce this. In the context of CQC and Spiegelhalter's approach, it is used to scale z-scores removing $q\%$ from each end of the distribution (where q is commonly 10%), by:

- Ordering the z-score values from lowest to highest.
- Finding the q th, and the $(100-q)$ th percentile z-score.
- Setting z-scores lower than the q th percentile, equal to the q th percentile, and values higher than the $(100-q)$ th, equal to the $(100-q)$ th. In other words, z-scores <10% set to 10% value, and values >90% set to 90% value.

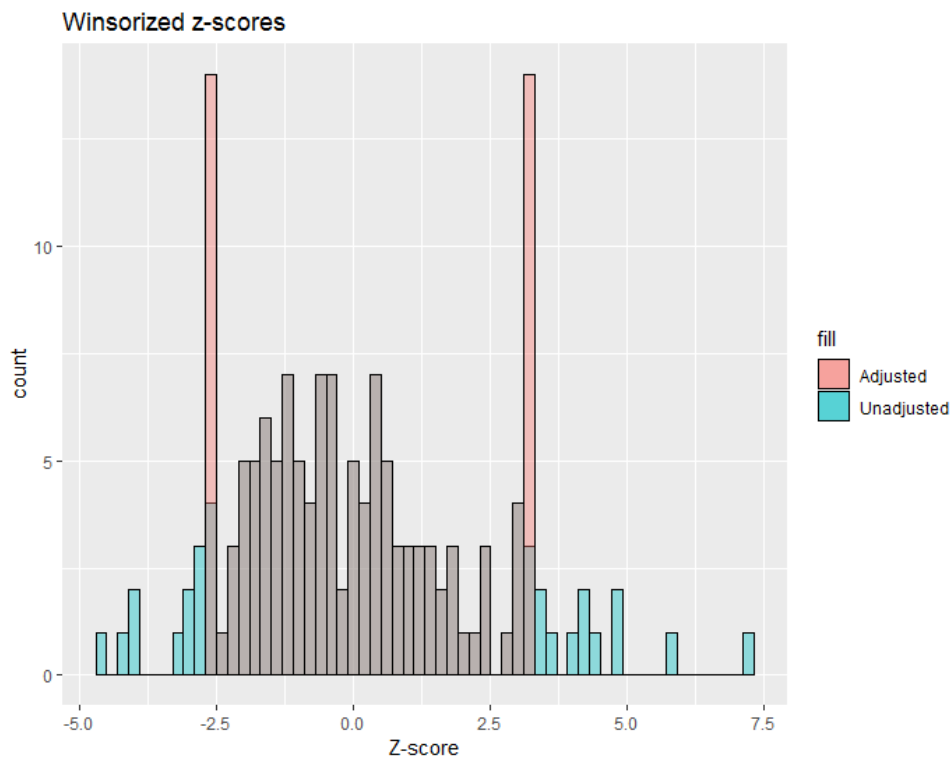


Figure 8.3 Example of Winsorisation of a distribution of z-scores

The red plot represents the Winsorised version of the unadjusted z-scores (blue), with a 10% Winsorisation at both ends of the distribution.

Figure 8.3 shows the effects of Winsorisation on the distribution which is used in place of the original z-scores. In the context of our z-scores, the sum of the squared Winsorised z-scores (Z_i) is divided by the number of trusts to obtain ϕ :

$$\phi = \frac{\sum_{i=1}^n Z_i^2}{n}$$

A logic step is then applied, and if $n * \phi < (n - 1)$ we assume no evidence of overdispersion. If there is no evidence of overdispersion, unadjusted z-scores can be use.

If we detect overdispersion, we then estimate the between-trust standard deviation τ^2 , and add this to s^2 when calculating our z-scores or associated funnel plot limits. The τ^2 can be calculated as follows:

$$\tau^2 = \frac{(n\phi - (n - 1))}{\sum_{i=1}^n w_i - (\sum_{i=1}^n w_i^2 / \sum_{i=1}^n w_i)}$$

Where $w_i = 1/S_i^2$ and i represents a particular trust, with S_i the within trust standard deviation. The adjusted z-score is then calculated as:

$$Z_{adj} = \frac{(Y - T)}{\sqrt{(S^2 + \tau^2)}}$$

The adjusted z-scores can be used for comparison between organisations, but a more desirable approach might be to substitute the τ^2 with the estimated variance from the random-effect models directly, rather than as a post-hoc calculation.

The Summary Hospital-level Mortality Indicator (SHMI) (Campbell et al., 2012) published by NHS Digital uses this additive overdispersion approach to expand the limits of the SHMI funnel plot (Clinical Indicators Team, 2016b). Although based on the Spiegelhalter et al. techniques, the SHMI limit calculation varies in two ways:

- SHMI uses a log transformation rather than a square root transformation for calculating the z-score. Spiegelhalter et al. discuss this in the appendix of their RSS paper, deeming it too severe a correction to the z-scores leaving a long negative tail (Spiegelhalter et al., 2012b).
- SHMI process does not use Winsorisation per se, but rather truncates the distribution. On enquiry with NHS Digital, they referenced a single paragraph in the appendix of (Spiegelhalter, 2005b) that suggest truncation as a possible alternative. This may be easier to calculate in NHSD's database systems, and as they do not present z-scores, it is not necessarily an issue to remove organisations.

8.3.4 Time series monitoring using 'cusum' charts.

Changes in mortality ratios, and other indicators may occur over time, but it is challenging to distinguish from fluctuations in case mix and changes in the number of patients seen. The CQC/Spiegelhalter methods approach this by using a time-based control chart called a risk-adjusted cumulative summary (CUSUM) plot (Care Quality Commission (CQC), 2014b, Grigg et al., 2003, Spiegelhalter et al., 2012c).

CUSUMs are sequential hypothesis tests, testing evidence for observations occurring at a reference rate (Null hypothesis, H_0), against evidence for a change in rate. They are commonly use log-likelihood ratios to form weights that are then cumulatively summed, of the form:

$$C_0 = 0$$
$$C_t = \max\{C_{t-1} + w_t, 0\},$$

Where C = is the cusum value, starting from 0, C_t the cusum value at observation/timepoint t , and w_t the cusum weight (log-likelihood ratio) for observation/timepoint t .

The cusum, used in this way, is usually started from zero and is restricted to be positive, to prevent 'inertia' (Woodall and Mahmoud, 2005) or building up 'credit' for a run of good performance (Grigg and Spiegelhalter, 2008). The plot can, however be calibrated to detect increases in the rate by setting the δ_2 parameter, discussed below.

A trigger value (h) is set for the plot, in a manner similar to the control limits discussed above. The trigger value is set for an acceptable limit to identify special-cause variation. In this case the trigger is usually pre-selected for a given false-positive rate. If a cusum plot reaches its trigger value, monitoring is expected to stop and remedial action taken (Grigg et al., 2003). In reality, the monitoring usually continues whilst remedial action is taken but the chart is reset, usually to zero, to avoid continual triggers.

The cusum is, however, known to trigger over time even with no signal. An important metric used to measure this is average-run-length (ARL), or the amount of data points before a false trigger. This in-control ARL can be used to calculate the time to alarm. The cusum has been shown to have the fastest time to alert, amongst comparable charting techniques, when a rate is out-of-control (Moustakides, 1986). The false positive rate of such charts will vary between groups/hospitals when using a fixed trigger value (Tian et al., 2015). Simulation of many cusums under the same conditions allow calculation of an observed false alarm rate and is a valid method for setting appropriate triggers (Bottle and Aylin, 2011). This method is computationally intensive and has led to the use of an approximation, demonstrated in the

paper, that is used in preference for Imperial College's mortality outlier monitoring (confirmed by personal communication with Dr Bottle, November 2018).

Grigg and Spiegelhalter, showed how to convert the ARL into a probability of false alarm that can be applied to general normally distributed cusum plots, and an approximation that can be used to set a desired false alarm rate (Grigg and Spiegelhalter, 2008). This was described as a marginal model over normally distributed z-scores in the context of mortality indicators and can be interpreted as the average false alarm rate across all units in monitoring. This approach is more feasible from a national monitoring perspective, given the use of z-scoring described above. It can also be adapted to factor cusums across different organisations over time.

Lucas and Crozier proposed an alternative set-up to the cusum chart, with a 'Fast Initial Response' (FIR) cusum (Lucas and Crosier, 1982). Under the FIR schema, the start value is half of the trigger value, and resets to half the trigger value rather than to zero. Lucas and Crozier suggest that this representation does not affect in-control cusums, as they will reduce to zero, but out-of-control charts will trigger more quickly. This approach has not been applied to NRLS models, as it makes the assumption that the process is out-of-control at the start of the run, and this is as unknown for NRLS models. The reset to half has, however, been adopted in Imperial College's mortality monitoring scheme that sends data to the CQC (Bottle and Aylin, 2008, Bottle and Aylin, 2011), and they have described it as putting an organisation 'on probation' after an alert.

The CQC's published method for cusums uses a similar structure to the cross-sectional approach detailed above, with two sources of variation. In the context of the z-scores and funnel plots above, σ represents the within organisation standard deviation and τ the between organisation standard deviation. Cusums are not cross-sectional, but longitudinal, occurring overtime. They are therefore modelled in a similar manner, where σ is the within period standard deviation and τ is the between period standard deviation.

The cusum is formulated from hypothesis tests where the null hypothesis is that the local mean (θ_k) of a z-score is in the upper part of its probability distribution (Care Quality Commission (CQC), 2014b):

$$H_o: \theta_k = \gamma_1 \tau,$$

Where γ_1 is a tolerance factor for the mean, that is commonly set to 0.5, and τ is the same standard deviation calculated in the additive model described above. It is therefore allowing

half of the ‘between period’ standard deviation around zero as a ‘null range’ for the local mean. The test for the alternative hypothesis is:

$$H_1: \theta_k = \gamma_1\tau + \gamma_2\sigma,$$

Where γ_2 is the difference in log-likelihood ratios deemed ‘out of control.’ σ is the within-organisation standard deviation, we are thus testing for a γ_2 multiple of the local standard deviation, given a tolerance of half the between-organisation standard deviation.

The z-scores are then further transformed to normal deviated under the null hypothesis using:

$$z_{kt}^* = \frac{z_{kt} - \gamma_1\tau}{\sigma}$$

With a corresponding change in the hypothesis test to:

$$H_0: \theta_k^* = 0$$

$$H_1: \theta_k^* = \delta$$

Where $\delta = \gamma_2$ and δ is commonly set to 2. This corresponds to a z-score of two, a doubling in the odds of event in the period, and an approximated 95% confidence interval around the national mean (or, indeed, the 2σ limit of a funnel plot).

8.4 CQC-style techniques applied to NRLS data models.

CQC techniques were applied to standardised incident reporting ratios (SIRRs), calculated using the outputs of incident reporting and DS models, as:

$$SIRR = \frac{\text{observed incidents}}{\text{predicted incident reports}}$$

8.4.1 Marginal vs. conditional

Models developed from GLMM & GAMs can be used to predict expected numbers of incident reports as either:

- **Marginal:** Predicting without the random-effects estimates allows the model to predict the global ‘average’ values based on the predictor variables.
- **Conditional:** Predicting expected values including the random-effects estimates for cluster (model is ‘conditional’ on the random-effects) adds additional adjustments for the clustered units.

Both marginal and conditional values have been extracted from models, but the z-scoring and CQC processes have been applied to marginal predictions. The rationale for this is two-fold:

- Random-intercept models were used to account for correlations due to clusters when estimating the effects of predators, aiming to remove average cluster effects (local differences in intercept) from the model. Marginal model predictions should better represent the global model effects, without allowance for the clusters, and are a 'less biased' prediction based on the exposure factors. In the NRLS modes we can predict the number of incident reports as expected for the average hospital, based on casemix. Conditional estimates would allow all organisations to deviate from this national model and absorb some of the variation between trusts.
- The additive random-effects model described above calculates a random-effect estimate for the between-organisation variation. Applying this to conditional estimates is nonsensical, as they are both attempting to adjust for the same effect, therefore counting the random-effect twice.

Conditional estimates in models provide additional information for within-hospital monitoring, answering the question: 'Are any data points high or low despite adjustments for the average effects of an organisation's culture?' These estimates are not presented as funnel plots, but conditional SRRs can be compared with marginal SRRs in scatter plots with an example in Chapter 10/Appendix B.

Poisson versus NB2 models in a conditional setting present slightly differently. The Poisson model showed residual overdispersion despite estimating the random-intercept. This represents the residual deviance in the model that is not explained by clustering, and may be related to aggregation, missing covariates etc. It presents as a wider spread within a funnel plot that, following the additive method above, would give a $\phi > 1$ and require adjustment. This is nonsensical, as the effects that the additive model are detecting have already been adjusted. This is not the case with the NB2 model, as the residual variance is scaled out of the model. Poisson and NB2 models were therefore both examined in case the NB2 model added obvious bias when visualising with a funnel plot.

Random Forest models do not include random-effects and so, in this sense, conditional and marginal models equate to the same thing. Model predictions were therefore presented as a single output.

8.4.2 NRLS results

8.4.2.1 Z-score results for 2015/16 GLMM, GAM and RF models

Transformations for indicators prior to z-scoring, described above, were applied to the data and comparison plots of the CQC method (square-root and Winsorisation) and SHMI (natural logarithm and truncation) were plotted and can be seen in Figure 8.4. These plots suggest a pattern that was conserved across all the models fitted. It appears that both transformations to z-scores are similar in shape with differences most notable in the Winsorisation/truncation stages (excluding 28 values each time for the SHMI method due to the truncation). The final adjusted z-scores were similar in shape for both transformations, but the distribution of the SHMI/log-transformed z-scores showed more data in the lower tail of the distribution. When combined with the higher number of outliers generated by the CQC method, it may suggest that the CQC method is under-adjusting the final z-scores, as many are larger than the expected range (see Appendix C.6 for tables). Spiegelhalter et al (2012c) suggested the log-transformation may be a 'bit extreme,' but the evidence for NRLS models suggests it may be less extreme in a sense (Figure 8.4) and a more suitable transformation than the square-root/Winsorisation method. The log-transformation also appears to have some precedent in the literature (Talbot et al., 2011, Hosmer and Lemeshow, 1995), and the SHMI method has therefore been adopted for funnel plot reporting on these models. The difference, however, is likely to be due to denominators. The SHMI method is calculating the ϕ based on only 80% of the organisations that the CQC method is (despite the extreme values being Winsorised), and the average squared z-scores in both techniques will bring about different effects. The CQC method appears to estimate a larger τ^2 , but the constraint on the z-scores during estimation appears to bias the calculation when applied to the full dataset. There also appears to be a unit-size dynamic, when compared with figure 8.5. SHMI limits are wider than CQC limits for DS incidents, but narrower for all incident models.

The Winsorisation, 10% at both ends of the range, is arbitrary. This threshold appears to be in general use for the SHMI truncation and CQC Winsorisation, but this threshold could also be altered to allow more suitable estimation. It is, however, a difficult balance between characterising heterogeneity for adjustment, and censoring it if it is too extreme.

According to CQC guidance, the adjusted z-scores should be calculated based on inflating the Winsorized values. This is somewhat nonsensical, as it standardises all organisations in the upper and lower Winsorisation band. This very likely to be a typo in the CQC guidance, as it is followed by a correct calculation using the raw z-score, and the academic publication from Spiegelhalter et al.(2012c) uses the raw z-score. SHMI guidance does discuss the calculation of τ^2 , but does not go as far as calculating adjusted z-scores. This has been done manually on the

log-transformed unadjusted z-scores, as using the truncated scores drops 20% of the data and, again, is nonsensical. It is the most extreme values we are trying to identify, so dropping them is counterintuitive.

A full list of z-score tables is included in Appendix C.6 for reference, generating too many outliers to list here when using thresholds of ± 3 . The use of z-scores here is less definite than the SPC principles of funnel plots below. Z-scores are continuous, and starting at the most extreme results, we can always choose the next successive score to investigate. In a sense, this contradicts the principles espoused by Goldstein and Spiegelhalter (1996) and Lilford et al. (2004), but it is also a pragmatic approach that starts with the most extreme reporting behaviour after casemix-adjustment. Care should be taken, however, not to present them as league tables that encourage comparisons between organisation, rather than against the national average. In this sense, funnel plots may be preferred, but could be considered to yield less information than z-scores.

At the request of NHS Improvement, the electronic version of this thesis censors the names of Trusts to guard against misinterpretation of experimental statistics.

Organisation names are obscured with black boxes :



Please contact the author, or UCL library, to view an uncensored copy.

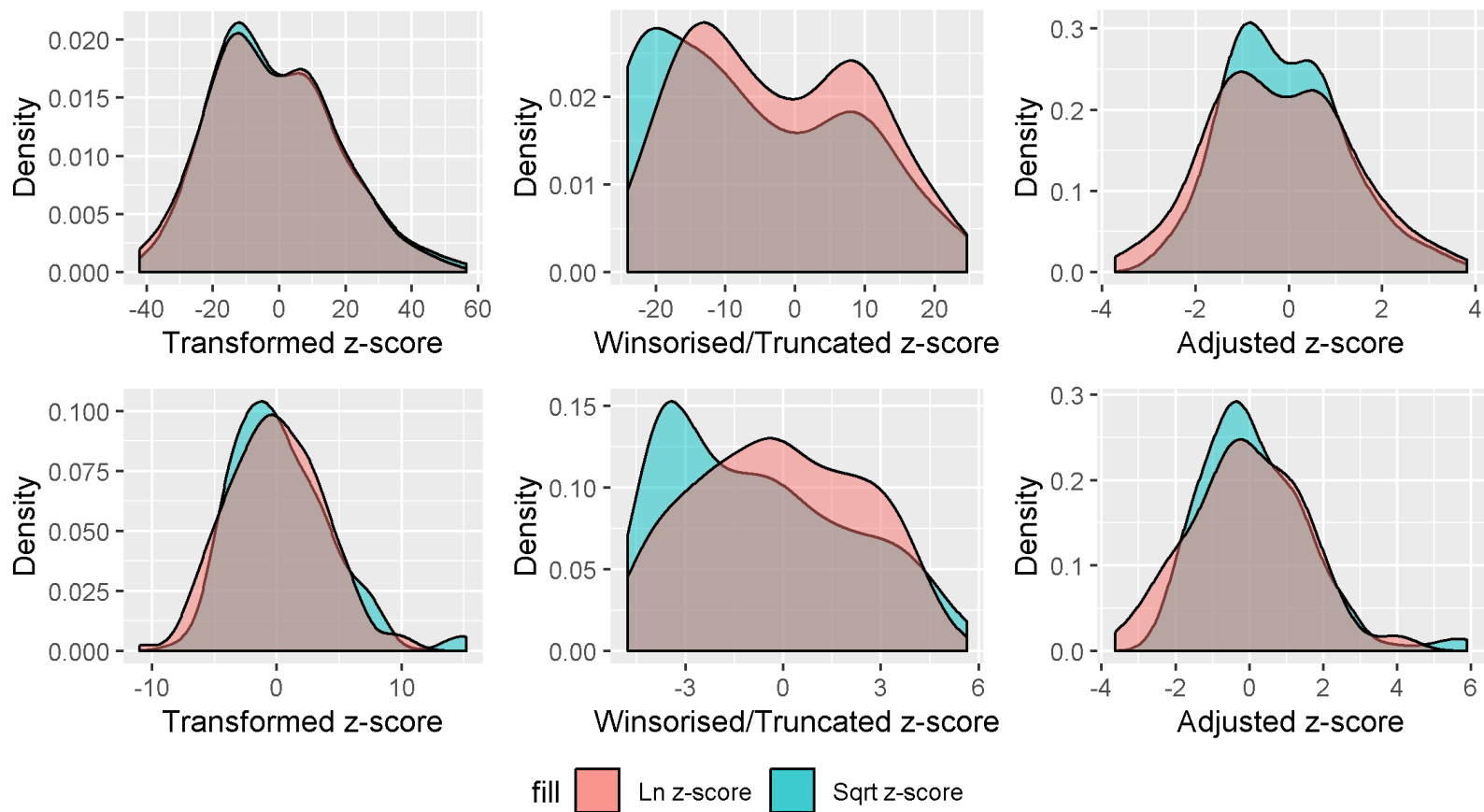


Figure 8.4 Distribution of z-score transformation and adjustment methods

CQC/Spiegelhalter technique (blue) and SHMI technique (red).

Top row relates to all incident Poisson models and bottom row related to Poisson DS incident model.

Left column is the transformed z-score, middle column is the Winsorised/truncated transformed z-score, and right column is the final adjusted z-score.

8.4.2.2 Funnel Plots

Funnel plot were calculated following the methods described above. Overdispersion was present in all models, as known from the model fitting process, but also confirmed via the ϕ calculation. The distribution of z-scores discussed suggested using the 'SHMI' method of log transformation and truncation for calculating the τ^2 value for inflating the funnel limits, but both techniques were examined and can be summarized on the Poisson GLMM data (for all incident reports) in figure8.5.

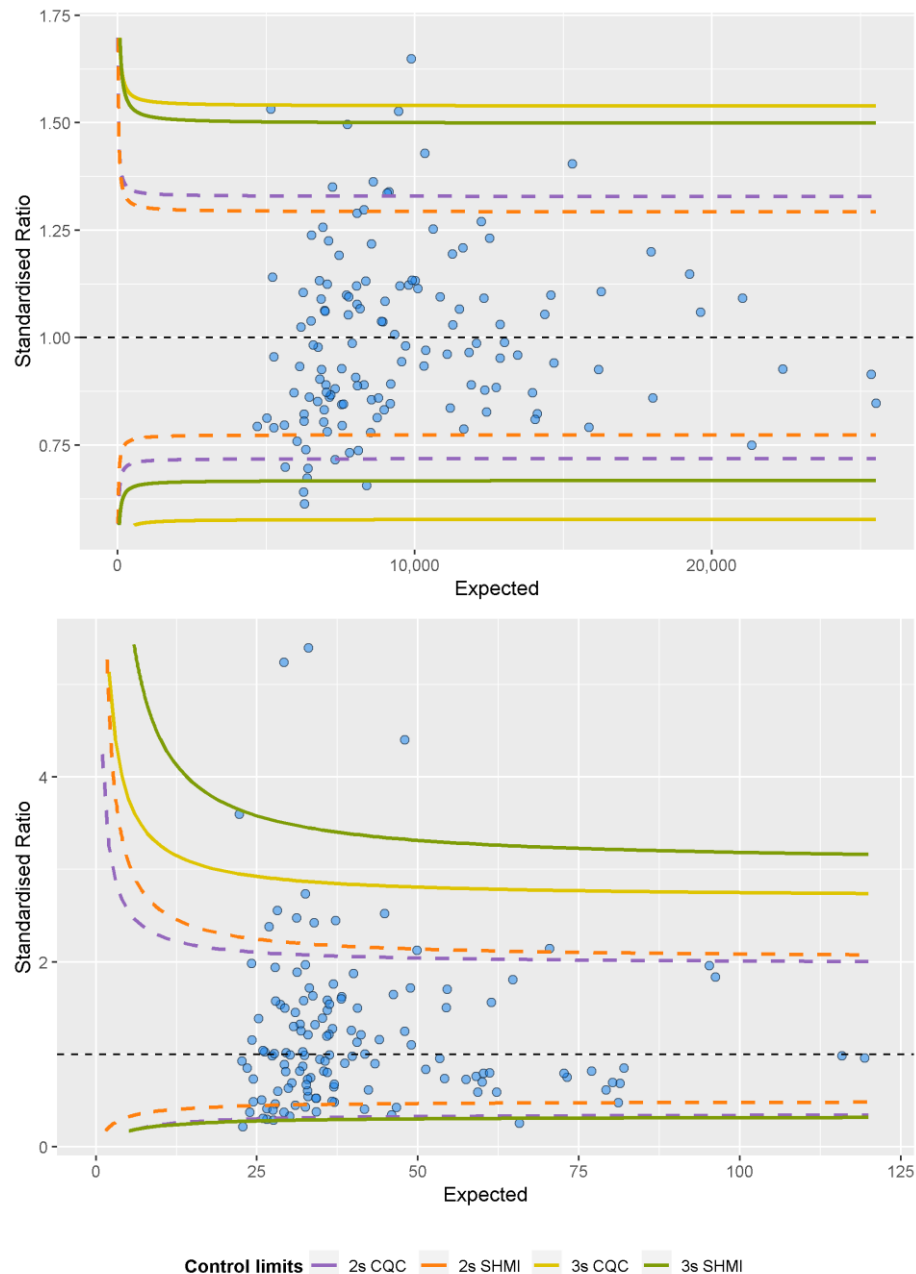


Figure 8.5 Comparison of overdispersion adjusted funnel plot methods

CQC/Spiegelhalter methods (square-root and Winsorisation) versus SHMI method (log and truncated). Upper plot represents all incidents and lower represents DS incident models.

Funnels calculated using the SHMI methodology appear to be more conservative than those based on the CQC/Spiegelhalter technique, on the all incident data, but the opposite is true for DS models. Given that we do not know what the 'true' answer is, it is appear heuristically appropriate to select the log transformation and truncation approach, as it is slightly more conservative for larger counts, but more permissive for counts based on very small numbers (DS models) where we would expect chance to play a larger role. An alternative method for future development of funnel plot might be to consider the false discovery rate, as suggested by Jones et al. (2008).

The SHMI-derived method was then applied to the other models for total incidents (figure 8.6 and 8.7) and DS incidents (figure 8.8). Figures 8.6 and 8.7 suggest a degree of consistency between the Poisson GLMM, NB2 GLMM, NB2 GAM and NB1 GAMs. The Poisson GAM appeared to have accentuated outliers, particularly for organizations: [REDACTED]. Observed incident reports numbers (numerator) are identical across models so differences in positions of points is due to differences the predicted values ('expected') between models (denominator). In the case of [REDACTED], the Poisson GAM appears to underestimate the expected number of incidents compared with other models, leading to a much higher standardised ratio.

NB1 and Random Forest models for all incidents (Figure 8.7) showed different patterns to other plots. NB1 suggested a larger τ^2 value and inflated the control limits. Random Forest methods showed the opposite, with points clustered more tightly and a lower τ^2 , leading to tighter control limits. Random Forests are not specifically modelling the clustering/random-effects elements, but merely fitting on the fixed effects. The decorrelating properties of random forest will likely reduce these effects, but the extent of this is unknown, and models may be biased.

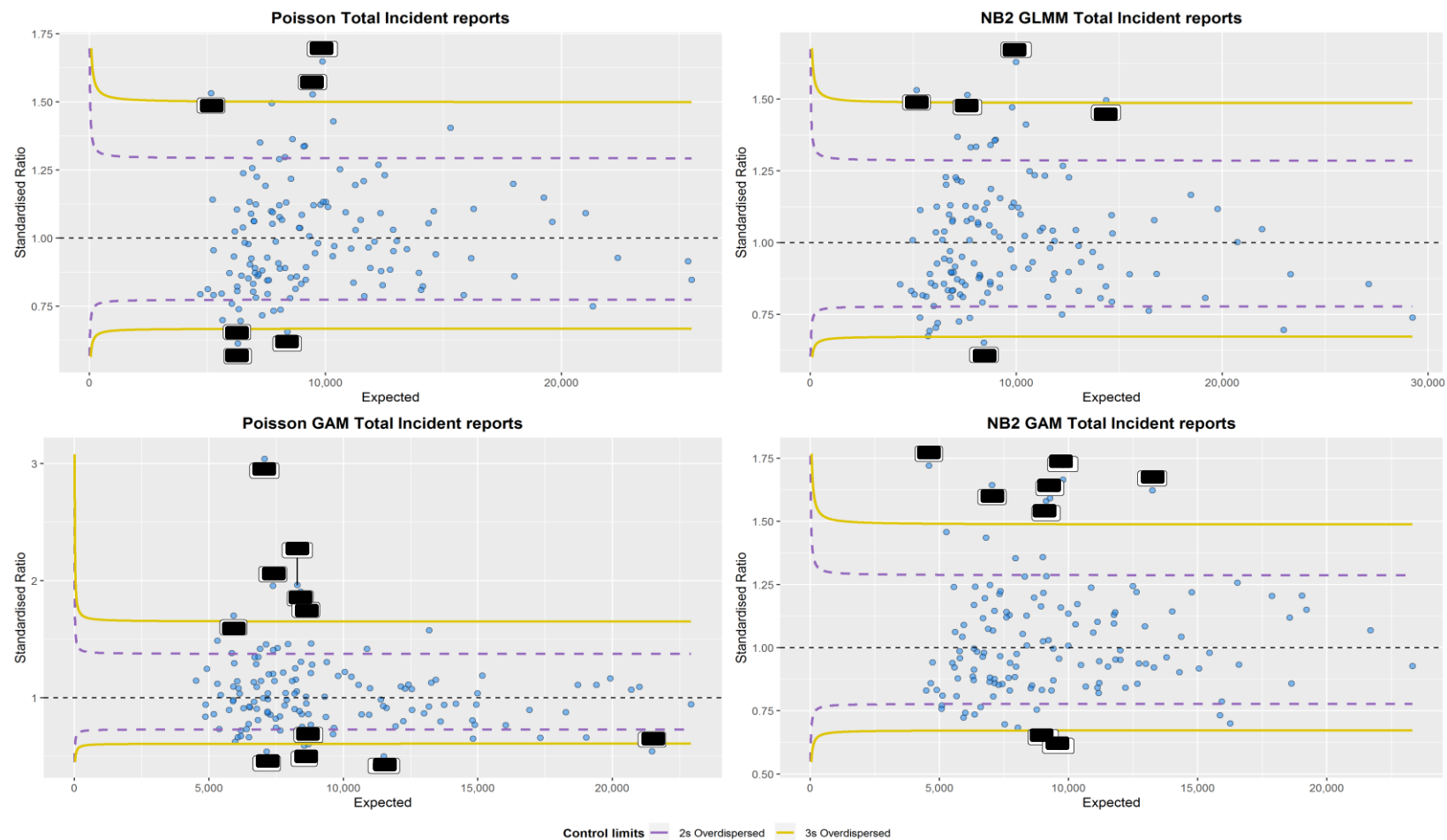


Figure 8.6 Standardised Incident Reporting Ratio funnel plots for total incident report models

Each plot represents a different modelling technique or distribution, denoted by the plot title. Control limits are 2σ (purple) and 3σ (yellow), with control limits expanded to adjusted for overdispersion, as per the methods described in (Campbell et al., 2012, Clinical Indicators Team, 2016b)

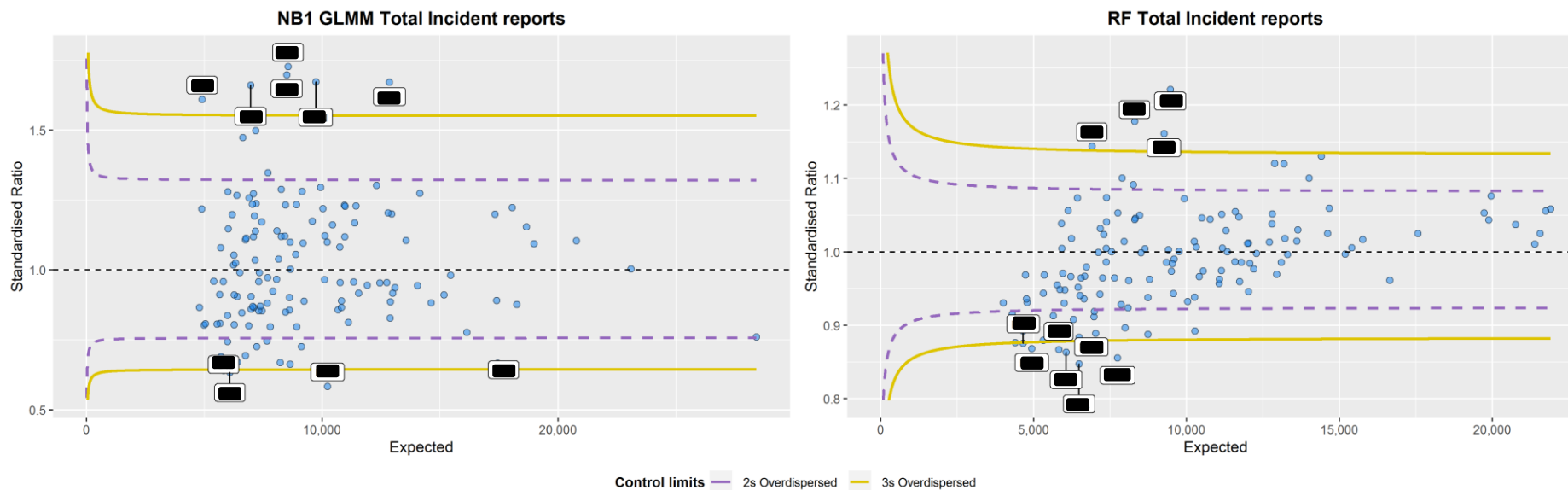


Figure 8.7 Standardised Incident Reporting Ratio funnel plots for total incident report models (2)

Each plot represents a different modelling technique or distribution, denoted by the plot title. Control limits are 2σ (purple) and 3σ (yellow), with control limits expanded to adjusted for overdispersion, as per the methods described in (Campbell et al., 2012, Clinical Indicators Team, 2016b)

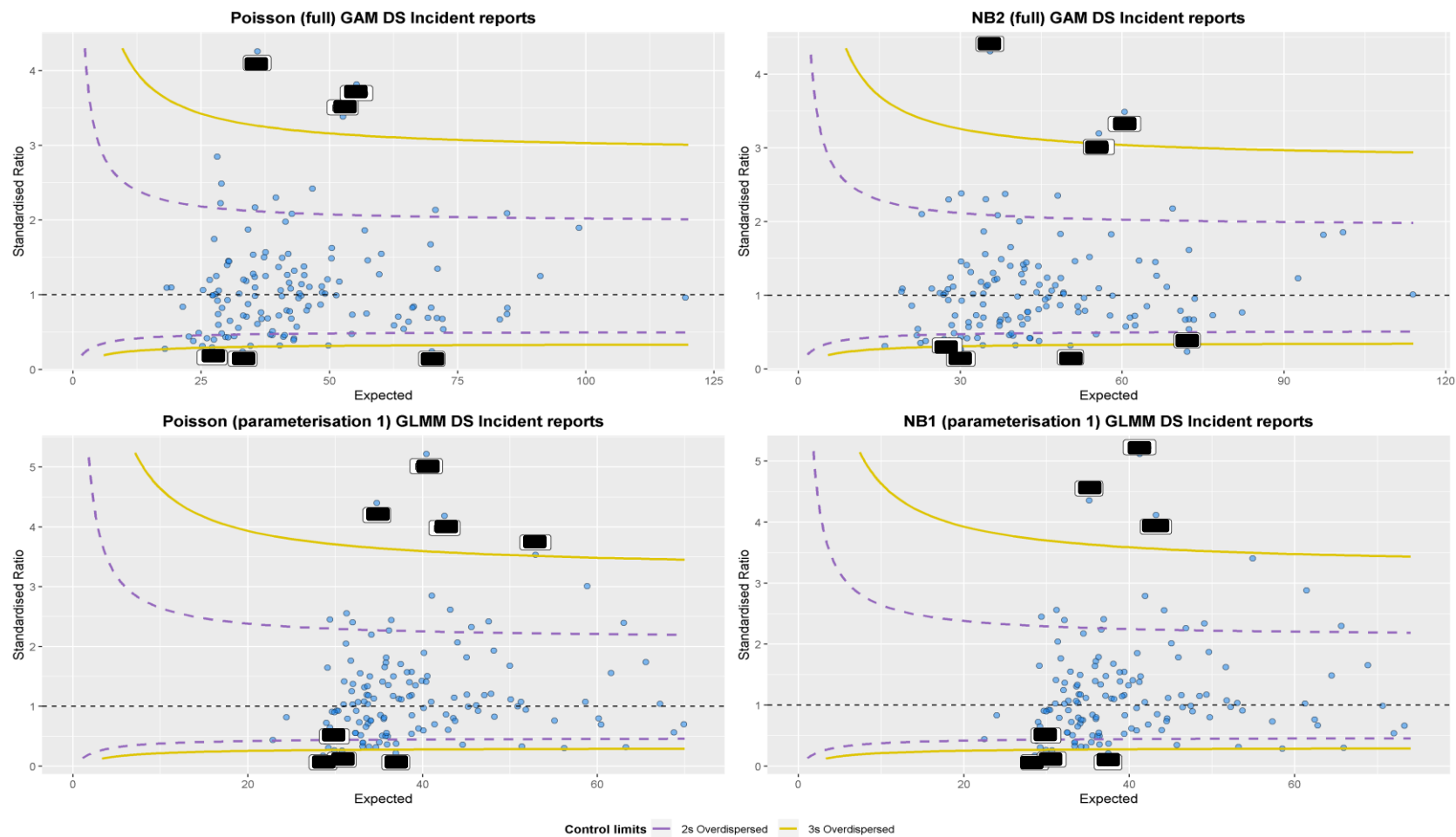


Figure 8.8 Standardised Incident Reporting Ratio funnel plots for death or severe harm NRLS incident report models

Each plot represents a different modelling technique or distribution, denoted by the plot title. Control limits are 2σ (purple) and 3σ (yellow), with control limits expanded to adjusted for overdispersion, as per the methods described in (Campbell et al., 2012, Clinical Indicators Team, 2016b)

8.4.2.2 Outlier organisations

A consistent group of outlier organisations was seen amongst most of the models, for both 'all incidents' or 'DS incidents' (table 8.1). These organisations are displaying the most variation in incident reporting, given their casemix/exposure. These organisations would be the first 'port of call' for a regulator to investigate. The organisations that are outliers in some models need a degree of validation and discussion with NHSI, and at the time of writing, have been sent to NHSI for comments. The degree of support for these organisations being labelled as outliers, supported by NHSI's other work streams, could act as further validation for model choice, and reduce down to single models.

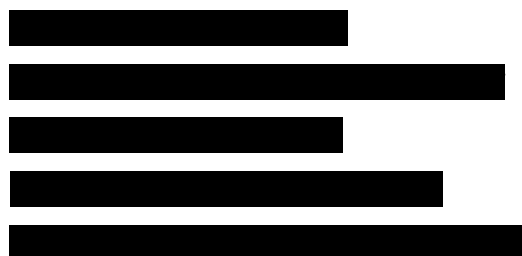
Trust Name	All incident models							DS incident models					
	Poisson GLMM	NB1 GLMM	NB2 GLMM	Poisson GAM	NB2 GAM	Random Forest	Total	Full Poisson DS	Full NB2 DS	Poisson GLMM (Param 1)	NB1 GLMM (Param 1)	NB1 GLMM (Param 2)	Total
	1						1						0
				1			1						0
		1					1			1	1	1	2
							0	1	1	1	1	1	5
		1					1						0
		1	1		1	1	4						0
						1	1						0
	1	1		1			3						0
		1			1	1	3						0
	1	1	1	1	1		5						0
							0	1	1	1	1		4
						1	1						0
							0						0
							0	1	1	1	1	1	5
				1			1						0
				1		1	2	1	1				2
							0	1	1	1	1	1	5
	1	1	1	1	1	1	6			1	1		2
						1	1						0
	1	1		1	1		4						0
	1		1			1	3						0
				1		1	2						0
				1	1		2					1	0
		1	1	1	1		4						0
				1			1						0
							0	1	1	1	1	1	5
						1	1						0
						1	1						0
									1				1
										1			1
		1		1	1		3						0
Total	6	10	5	12	8	11	-	6	7	8	7	6	-

Table 8.1 Funnel plot outlier organisations for 'All incidents' and death or sever harm (DS) incident models

Outliers are based on 3σ limits using adjustments for overdispersion based on, (Campbell et al., 2012, Clinical Indicators Team, 2016b)

In cases where organisations are consistent outliers across several models, we can assume that they are clear outliers, less affected by small shifts in the control limits between different modelling techniques. They include:

- Total Incident report models:



- Death or Severe Harm Incident reports:



Organisations that are single outliers in a given model may represent quirks of particular techniques/assumptions and could be used with other external data for validating model choice. This validation, again, requires input from regulators to assess. The number of outliers identified by models does not appear to be wholly consistent with the use of GLMMs, GAMS or RFs, nor specific quirks of distributions such as Poisson or NB2. E.g. NB2 GLMMs and NB2 GAMS do not share all outliers.

The manner in which these funnel plots are interpreted is of importance. Incident reporting rate is not an orthogonal outcome. We do not know if an organisation with a high reporting rates has a comparatively high incidence of incidents, or is simply more aware/mature in its reporting culture and better at recording and learning from error. The funnel plots, therefore, do not show 'good' or 'bad' organisations, but show organisations with systematically different incident reporting cultures that are not explained by natural variation or expected effects of exposure. Being an outlier organisation should trigger examination of the data submitted to NRLS, local recording of practices, and whether there is anything to be learnt from an organisation's incident reports. They may have issues to address locally, or show behaviours that might benefit the wider system.

8.4.2.3 CUSUMs outputs

Cusum outputs are presented in full in Appendix C.7. Both conditional and marginal cusums were assessed for inclusion, but marginal plots were chosen as they better fit the concept of a national exposure effects for constructing the additive random-effects model described above. Conditional predictions gave more alerts for all organisations, due to the reduced variance from the conditional means in this case. This works against the estimation of τ in the cusum calculations detailed in section 8.3.4, as variance assumptions are different between points. It is also somewhat contradictory, as they are both predicting values using the modelled random-intercept, but also attempting to estimate a further random-intercept in τ , i.e. they are estimating the same effect. A point for future research would be to examine whether conditional models using unadjusted z-scores based on conditional predictions, without estimating τ , give similar results. This may provide some measure as to whether the estimation of τ fits with the model estimated random-intercepts well.

All organisations showed some Cusum triggers, depending on the model used. This suggests that the variation from one month to the next is more extreme than in many other indicators used by CQC. A possible solution to this is to increase the threshold delta value to 3 instead of 2, corresponding to a tripling in the odds of death rather than a doubling. To do so would also require shifting of the Cusum trigger value, according to the methods described by Grigg and Spiegelhalter (Grigg and Spiegelhalter, 2008), to maintain a false discovery rate of 0.1%. This is an area for future development with these models. It has not been possible within the timeframe of this thesis but is suggested as the next stage of this work, as it may make SIRR cusums more useful to organisations and regulators, only focusing on extremely different reporting rates or extreme changes.

Figure 8.9 illustrates a set of Cusum plots for the organisation with provider code '██████████', using the NB2 GAM total incidents model. This organisation was selected as it demonstrated high rates at the beginning of the monitoring period, resulting into two early triggers for the 'increasing' Cusum. It then returned to an 'in control' state, before showing a significant decrease in reporting, triggering for a decrease in month 11. The example of ██████████ is a useful illustration of the different uses for funnel plots and Cusums. ██████████ is not an outlier organisation on the funnel plots, as their high and low rates average out over the year and suggest they are in control. As a cross-sectional cut of the data, this is true, and the use of a funnel plot is therefore to identify which organisations are systematically different compared with expected behaviour. The Cusum plots show the change in reporting rates from higher to lower than expected. This

might be most useful for an organisation monitoring its own behaviour, or for regulators to identify shifts in reporting rates.

CUSUM plots for all incident SRR

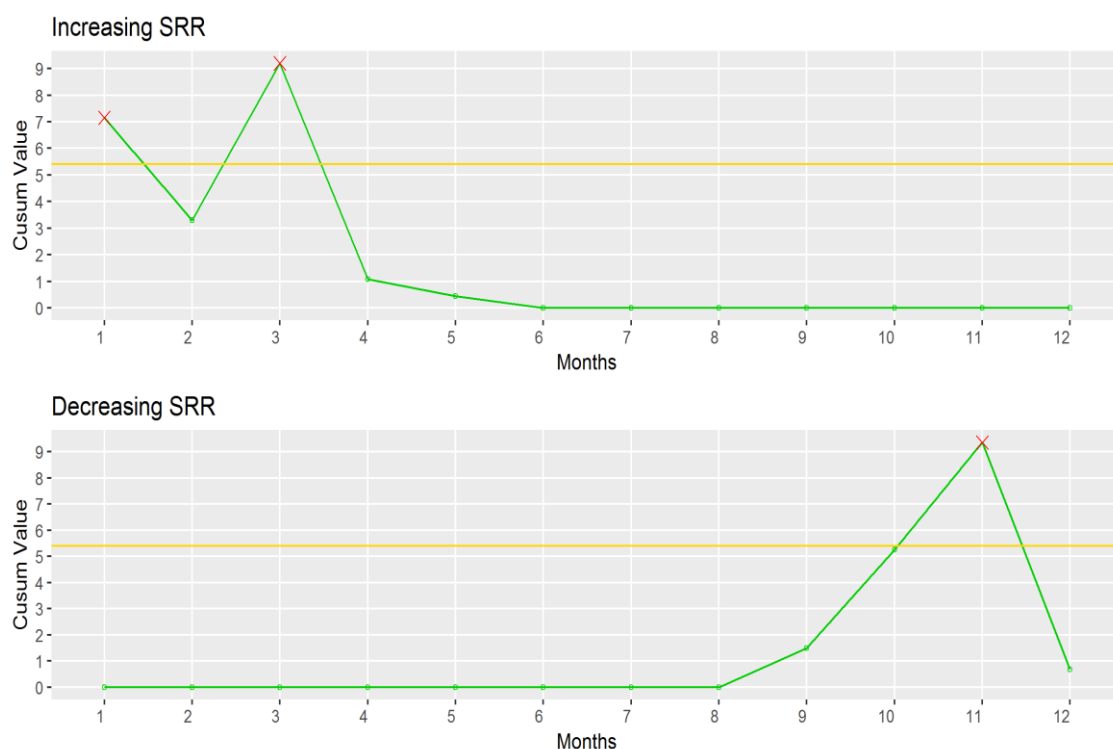


Figure 8.9 Illustrative CUSUM charts for Standardised Incident Reporting Ratios (SIRRs)

Plots are calculated using predicted values from 'all incident' Poisson GLMM models, for '██████████' (green). Y-axis is the cumulative log-likelihood ratio (Cusum value) that is reset after a trigger (red), dictated by present trigger values (yellow).

Organisations with notably different reporting behaviour triggered regularly using the CUSUM techniques presented above. Organisations that triggered at every month in the monitoring period were:





This constant triggering effect illustrates consistent reporting behaviour above or below the national reporting rates, rather than changes in behaviour. It also serves to highlight the problem with using cusums in this way: we are assuming the national average is 'correct.' The 'correct' value for a reporting rate is unknown, and it may not be possible to frame reporting rates in such a way. For the organisations above, such techniques will do little to help their monitoring, nor will they allow regulators to notice variation at these organisations. The cusums are 'saturated' in a sense and no longer sensitive to changes.

Death and severe harm incidents are, thankfully, rare events. This means that their prediction is challenging and does not generalise to the normal distribution particularly well. Given the low numbers, an approach based on cusums of z-scores is less helpful (Neuburger et al., 2017). The CQC guidance has also suggested an alternative parameterisation of the cusum based on Poisson or Negative Binomial distributions (Care Quality Commission (CQC), 2014b), but their technique has been proposed for monitoring of raw counts and is not specifically targeted at standardised ratios. This approach may hold potential for future work, as it is currently used by CQC for monitoring 'Never Events,' but a simulation study would be required to set appropriate limits based on false discover rates.

Another commonly proposed control chart for rare events is the g-chart (Morton et al., 2013, Neuburger et al., 2017). This focusses on the time between events, rather than the event themselves. This is not suitable for reporting at SIRR-level as our aggregation is over months, with small counts in most months. It may, however, be a useful tool for monitoring within trust level.

A further alternative option is to consider the proportion of incident reports that are death or severe. This could be rendered as a p-chart, or other common control charts for local monitoring. This technique was examined for a number of hospitals, and due to the relative

stability of the incidents figure versus the very low count of the death/severe count, it amounted to little more than a scale change. Data points remained in similar positions, with the same effects of small numbers in this case. For an in-depth examination of control charting and analysis techniques for incident reporting, see the thesis by Deng (2013).

Morton et al.(2013) suggest an alternative arrangement where charts can be plotted with smoothers based on GAMs (see Chapter 5). The confidence interval around the smooth therefore represent the predicted mean even if this changes over time, and can be used in a manner similar to a control chart to identify aberrant data points, but confidence intervals for small events in the presence of overdispersion are likely to be very wide.

8.5 Conclusions

The task of monitoring NHS organisations requires intelligence and information on which to judge or measure them. Methods have developed in this area to allow the comparison of organisations across multiple indicators, using both cross-sectional techniques and time-series methods. NHSI and the CQC currently produce counts and simple reporting rate indicators that are not casemix-adjusted. This chapter has answered aim 4 of this project by demonstrating that the models developed in preceding chapters can be used to derive a casemix-adjusted standardised incident reporting ratio (SIRR). These techniques can then be used appropriately by casting them in the same manner that CQC monitor many indicators: transforming to z-scores, using an additive random-effects adjustment, plotting indicators using statistical process control techniques for cross-sectional comparisons (funnel plots) and monitoring timeseries (CUSUM).

Incident reporting models have several major problems that complicate this application: they are highly overdispersed for total incident models, DS models are based on very small numbers incident reports and the CQC's default transformation for standardised ratios did not perform as well as an alternative method.

Adjusted Z-scoring, and assessment of overdispersion based on this was susceptible to the choice of transformation used, and whether the distributions were truncated or winsorized prior to assessment of ϕ . A method using a log transformation and truncation of the distribution before assessment of ϕ was adopted, as it appeared better suited to the dataset. Adjusted z-scores concurred with the funnel plots for the most extreme organisations but provide further detail that could allow a less drastic dichotomization into 'in control' and 'out of control' if required.

Funnel plots are a useful visual method that is suitable to summarise the modelling period and identify systematic variation. Given the current NRLS publication schedules, it would be appropriate to include funnel plots, an SIRR, and expected incident reports data with their current OPSIR.

These methods of adjustment could, of course, be replaced with an assessment of ϕ directly from the models. This could be substituted in to the construction of z-scores, or the variance components of multilevel model could be used in place of τ^2 . A further alternative would be, rather than using the techniques described above, organisations could be described in terms of their estimated random-intercept with a profiled confidence interval.

Whilst CUSUM methods have been applied, and can be used to monitor changes in incident reporting rates, they are calibrated to a doubling in the odds of incident and a halving of the odds. This threshold may be too low to be practically helpful given the residual overdispersion and the volatility of the indicator. Higher thresholds may be more appropriate, but this may be an inherent problem with this indicator. It may also suggest that this type of indicator, as it is not orthogonal, is a poor fit for this type of monitoring method, as 'in control' may not be a definable state for organisations with notably different reporting culture. A CUSUM that triggers every month is a blunt tool in this setting, and it is of little use.

These indicators function as expected in development and testing, but the real utility of these indicators can only be known in practice for the regulator. At the time of writing, the outputs of this work have been shared with NHSI, and received enthusiastically, but further work is yet to be agreed. The next appropriate step would be to validate these indicators to see if outlier organisations do show any appreciable differences to others and whether there are learning themes in this process. This validation process could include:

1. **Quantitative comparisons:** An initial comparison to other indicators including staff survey results, pressure ulcers, safety thermometer data, staff sickness, measures of mortality, readmission, length-of-stay or other clinically coded incidents/misadventure, but particularly CQC inspection ratings. Some of these indicators have been examined against incident reporting before, but not against more extensively standardised incident reporting indicators as developed in this project (Howell et al., 2015, Hutchinson et al., 2009).
2. **Qualitative review:** Interview or survey-based methods could be used to provide stronger evidence for differences in reporting culture and validate apparent culture differences seen in models. This could be used to compare culture at outlier organisations against others and identify learning from/for these organisations.

These methods must be found useful by regulators and NHS organisation, through this validation process, or they will add to the burden of NHS bureaucracy without facilitating learning. The aim of these metrics is to avoid future patient harm through learning and defensive changes to NHS systems. Any validation studies, or operational use of these indicators must resist any drive to create performance measures, as this will create perverse incentives for reporting, encourage gaming, or become blunt instruments used to penalize organisations (Lilford et al., 2004).

Chapter 9 Text mining models

9.1 Introduction

NRLS research, discussed in Chapter 2, has often suggested that the primary signal in the data set is buried within the free-text incident descriptions. This requires manual reading to extract meaning. Models built in previous chapters have focused on predicting incident reporting rates, but quantitative techniques may also be used to analyse free-text. This chapter presents an additional analysis route, that does not build directly on models in chapters 4 – 7, to answer aim 3 of the thesis. It examines simple analysis options for free-text in NRLS, presents visualisations of common words and their association with harm categories, and presents an application to predict levels of harm from free-text descriptions of incidents.

9.2 Text mining techniques

Text mining approaches have seen significant advances in the last 20 years, spurred on by substantial increases in processing power and the increased availability of data in the internet age. Internet search data, online shopping, product reviews and social media posts have been major resources in this development (Stieglitz et al., 2018). Text mining, or Natural Language Processing (NLP), have many methods but they can usually be categorised as:

- Analysis of term frequency
- Sentiment analysis, often associating positive or negative sentiment within text articles, such as tweets or customer reviews of products
- Supervised or, more commonly, unsupervised learning techniques (latent variable models) for semantic analysis, such as topics modelling
- ‘Word embedding’, commonly performed with neural networks, fragments of text are rendered as word vectors in high-dimensional models, with distance metrics, and graph theory used to draw out relationships.

9.3 Previous work with NRLS text

Text-based modelling has been applied to NRLS, most notably as the topic of a PhD thesis by Bentham (2010). This work focused on understanding the data and transforming them into an analysable format. NRLS has been shown to have many abbreviations, medical jargon, colloquialisms (Bentham and Hand, 2012, Bentham and Hand, 2009). Spelling mistakes are

common, such as 371 different ways of spelling *clostridium difficile* identified in one publication (Mayer et al., 2017). In Bentham's analysis, median word length for free-text descriptions was 20 characters, with the minimum being a single character. This single character is almost meaningless, and may be a full stop or series of 'x' characters. Error messages from software systems were also present, some of which appeared to be application code from a reporting system, such as:

```
`: span class="Number" onmouseover="doHover(this);"
onMouseOut="doUnHover(this);" 2208595N :b Number /b /span'
```

This appears to be HTML code, and is likely to be from a broken online reporting form (Bentham and Hand, 2012). There was no consistent time or data format within free-text sections. The NRLS team apply cleaning rules to remove all identifiable information where they find it, but some may remain in the text, as no rule set anticipates everything. Bentham showed that confusing terms and acronyms not only vary by institution, but may also be confused with other words, such as 'NICE,' the National Institute for Health and Care Excellence, will appear the same as the word 'nice' to the model.

Bentham's work rendered the text as a high-dimensional numeric vector space model (similar to the word-embedding principle described above). This model did not consider word order or grammar but considered distances between vectors in the projected space. Term-weighting was applied using the TF-IDF method, described below.

The dimensionality of the model was then reduced using principle components analysis (PCA) and the PEAKER anomaly detection algorithm (Zhang and Hand, 2005) used to find clusters that were closer than expected in the feature space. Several methods of validation were used, including expert clinical review, and using samples from the data comparing classifications. This model appears to have validated well against expert review and identified additional groups that were also of clinical interest, and were taken for further analysis by the review group.

Different text-based techniques have been applied in exploratory analyses by the NIHR Imperial Patient Safety Translational Research Centre at Imperial College. Imperial have conducted a work programme, commissioned by NHS England/NHS Improvement, to conduct NRLS analyses and review. Text mining techniques have been used to predict the level of harm associated with incidents in the current system and for use in data entry of new incidents (Mayer et al., 2017). These techniques have also been examined by data science company Mastodon C, who have advised Imperial on scaling text mining approaches for live data input forms, as well as the use of Elastic Search (Mastodon C, 2015). Mastodon C have also been

involved in incident report analysis, and used LDA text mining models (see below) with an NHS Trust in London (Mastodon C, 2019) (details of this work provided through personal communication, with the citation referring to a case-study on their website).

A recently published paper on the Arxiv platform (Cornell University, 1991), a repository for pre-prints that is moderated but not peer-reviewed takes a complex, and extensive, approach to neural network-based text embedding techniques. It examines incident reporting data, based on paragraph vectors, and constructs a document similarity graph (Altuncu et al., 2018). They then apply Markov Stability community detection algorithm that is used to identify groups of records with consistent content at different resolutions. Authors extract topics and relevant word descriptions from the groups and compare them against hand-coded categories, finding good consistency and additional clinical detail.

9.4 Preparing text for modelling

Free-text/unstructured data are not usually subject to the same validation rules and processing that categorical/structured data often are. It may contain many quirks of collection methods, situation/context and deficiencies that make analysis inconsistent and difficult for machines to process in a uniform manner. Language also contains ambiguities. A sentence is composed of words, but the information conveyed by the sentence may be more or less than the words themselves. i.e. the ‘sum’ may be more than it’s ‘parts.’ Depending on the algorithms chosen for parsing text to a machine analysable form, we may receive different outputs (Manning and Schütze, 1999). Various approaches exist to do this, but they commonly follow grammatic/language rules, or algorithmically learn a data representation from a large corpus.

In order to prepare text in general, and specifically for the NRLS case, a number of preparation steps are commonly used (Silge and Robinson, 2017):

- **Tokenisation:** Text must be split into the units of analysis (‘tokens’) and represented in a structure. This is commonly a split by word, but may also be by sentences, paragraphs, letters or other constructed n-grams such as word groups of 2 or 3. Models exist for this in R with text stored as a ‘corpus.’ A corpus is a database representation usually structured as a set of documents, with each document forming a separate entry, and words/tokens as elements of the document (Feinerer et al., 2008). An alternative, the ‘tidy data’ format, is commonly advocated in R. It is more suitable for modern systems structured around ‘key-value’ pairs, where columns represent values and row represent items (Wickham, 2014). This takes the form of data tables that are ‘long’ rather than ‘wide,’ and are usually more suitable for

database applications and high volumes of data. A text mining framework based on tidy principles has been created in R (Silge and Robinson, 2016) and was used as the basis for models in this chapter. Conversion to corpus/matrix format has been used for topic models that require this structure, but data can be converted to either format as required. The tidy framework was adopted primarily for its consistency, interoperability and easy of manipulation between R packages and the SQL Server database backend hosting the data. In this tidy format, two alternative tokenizations were examined: uni-grams (words) and ‘skip-grams’ (windows of word groups allowing words to be skipped). The skip-grams were chosen to capture the association between medical terms in incident reports. E.g. ‘pressure ulcer’ carries more information than ‘pressure.’ Each document was then represented as a single line per token with ID and harm keys. Table 9.1 shows how the two tokenizations represented two example incident reports: 1: ‘**Patient fell**’ (Moderate Harm), and ‘2: **Missed antibiotic dose**’ (No Harm).

Token type	Report ID	Token value	Harm
Uni-grams	1	Patient	Moderate
	1	Fell	Moderate
	2	Missed	No Harm
	2	antibiotic	No Harm
	2	dose	No Harm
skip-gram	1	Patient fell	Moderate
	2	Missed antibiotic	No Harm
	2	Missed dose	No Harm
	2	antibiotic dose	No Harm

Table 9.1 Example of ‘tidy’ tokenization and database representation

- **Lemmatization:** For inflected languages, lemmatization is the process of finding the root form of words. English has a reasonably simple morphology and lemmatization is not necessarily needed (Dalianis, 2018). This was not applied to NRLS due to the risk of transforming meaningful words.
- **Removing ‘stop words’:** In English and other languages, some words are used frequently as part of sentence structure without lending any meaning to a sentence, such as ‘the’ and ‘a’. The high presence of these words may skew text-mining, and they are commonly removed (Wilbur and Sirotkin, 1992). The exact words used in standard stop lists may vary slightly, but the Snowball project is a common source of lists (snowball.org) and was used in this project. Additional stop words may be considered if they dominate without adding meaning.

- **Cleaning:** Text may be 'messy' with use of numbers, other non-alpha numeric characters, possessive endings ('s') etc. These values are used inconstantly and, if their frequency is high, may skew topic modelling or other analyses. Numbers, non-alpha numeric and possessive characters were removed from NRLS, but wordclouds were calculated at each stage and the effects can be visualised in figure 9.1.
- **Stemming:** Similar to the possessive endings mentioned above, words that share common routes ("stems") may mean similar or the same thing and could be considered the same for topic generation. Stemming can be considered a more radical approach than lemmatization, and may lead to stems that aren't really words, requiring additional interpretation (Dalianis, 2018). This is context specific, e.g. 'Nurse,' 'nurses,' and 'nursing' may be considered the same, but 'nursing' may also refer to community nursing home, breast feeding, or protecting/carrying an injury. Stemming can be helpful in noisy datasets, but stemming algorithms may blunt some of the context specific information. NRLS was stemmed using the standard porter algorithm (Porter, 1980) and results before and after are visible in the wordclouds in Figure 9.1.

The cleaned, stemmed, tokenized data, with stop words and variants of 'patient' removed were used for topic modelling later in this chapter. The first iteration of modelling did not remove single letter words, but second modelling run for words and skip-grams removed these words.

9.5 Word frequency and document frequency

The frequencies of words in documents is a common first measure in text modelling. The frequencies of words within NRLS reports was analysed with the most common 100 words selected arbitrarily for visualisation. In total, and in all harm categories, the word 'patient' and its abbreviations dominated. The 'wordclouds' in the figures 9.1 and 9.2 represent the top 100 words in their respective groups: Figure 9.1 at different stages of data cleaning, and Figure 9.2 at different harm levels after cleaning and stemming.

In the NRLS data in figure 9.1, the word 'patient' and its abbreviation 'pt' or 'pts' dominated. Term counts revealed that it was 16 times more prevalent than the next most common word. 'Patient' was therefore removed from the dataset before topic modelling, with 'pt' and 'pts' also mapped to 'patient,' as it is unlikely to add value to models. On removal, other terms came to the fore, often mentioning nursing and wards, with 'cardiac' and 'arrest' particularly visible for 'death' incidents. Words related to pressure ulcers were common in low and moderate harm incidents, and words associated with beds, staffing and transfer were common in most levels of harm.

Visualisation of wordclouds is a helpful first step, but there are no semantics or word association included in this representation of words. After removing variants of patient and single-letter words, the wordclouds were recalculated using skip-grams with a window of three words, skipping up to one word. Wordclouds were then rebuilt and presented in Figure 9.3. 'Pressure ulcers' was dominant in all levels of harm except death. This is an interesting finding, when summary statistics (for a much longer period) based on level 1 incident types in Chapter 3, showed only 24 incidents categorised as pressure ulcers. Blood pressure was common in lower harm incident classes, but dominant in death incidents, along with similar terms of blood products, low pressure and low blood. A possible interpretation of this is that lower harm grades were more common in everyday hospital settings, and so the term 'ward' etc. was more frequent, but death incidents are rarer and may be more associated with trauma and/critical care. The terms 'left leg' and 'neuro obs' were surprisingly common to most harm levels.

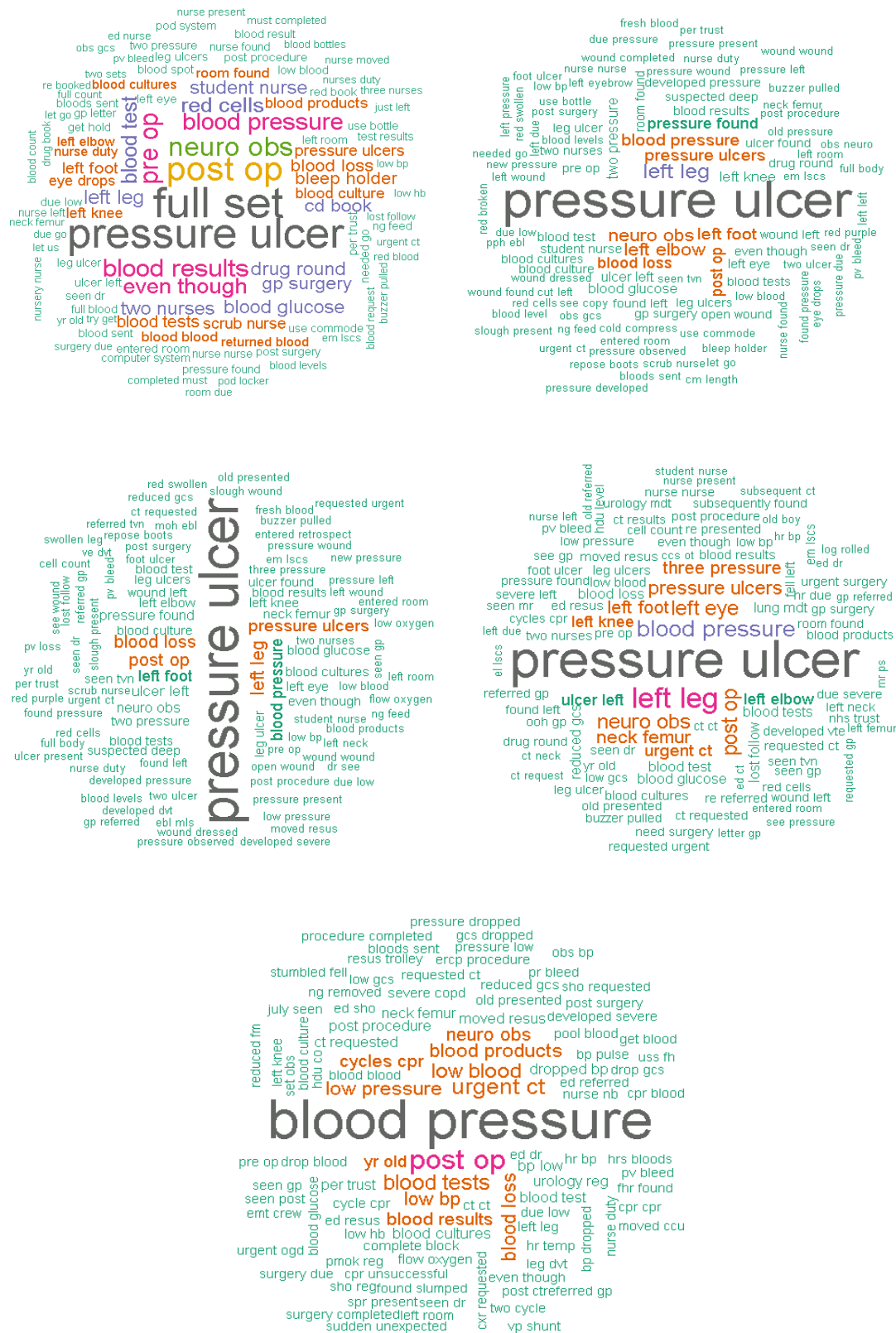


Figure 9.3 Wordclouds of top 100 bigrams, from ski-gram models, by harm level
Top row: left: 'No Harm' **right:** 'Low Harm'
Middle row: left: 'Moderate Harm', **right:** 'Severe Harm'
Bottom row: 'Death incidents'

A common next step in text mining, is to use word frequency in combination with document frequency to give more meaning to groupings. Per document, we can consider the term frequency (TF) as examined in the wordclouds above. Stop lists were discussed in section 9.4 as a way of dealing with high occurrences of low value words like 'the' and 'a', but stop word lists may not necessarily be the best approach (Saif et al., 2014). Stop words also have the draw back that, once removed, they cannot be used as part of search terms (Manning and Schütze, 1999). A more sophisticated approach might have considered that these terms are more important in some documents than in others. An alternative approach to this rigorous cleaning is to look at the frequency across documents, as well as within documents, and derive weights to adjust TF. Zipf's law suggest that we can compare the occurrence of words within documents with the frequency across the whole corpus (Manning and Schütze, 1999). The 'inverse document frequency' (IDF) (Spärck Jones, 1972) was the one of the first methods for this, that is still used in many applications (such as search engines) today. Silge and Robinson (2017) paraphrased this as: *"..the frequency that a word appears is inversely proportional to its rank"*. IDF decreases for commonly used words and increases for words that are rarely used across a set of documents. TF and IDF can be combined to weight the term frequency by multiplying them together (Salton and McGill, 1986). This is referred to as 'TF-IDF;' (or sometimes 'tfidf'), and can be used to infer how important a word is to a document in a collection (Silge and Robinson, 2017).

TF-IDF is best used in applications comparing relevance of words in particular documents, e.g. searching for documents related to a given word, e.g. spotting hospital acquired infections from important words in patient health records (Ehrentraut et al., 2018). Though commonly used for word relevance measures, TF-IDF in isolation can be spurious over large corpora where terms appear once in a document, as their low TF can be boosted by high IDF and make words seem more important than they may be (Salton and Buckley, 1988). Depending on an analyst's requirements, this may be unhelpful. TF-IDF also does not consider synonyms and miss-spellings that may limit its effectiveness, and use of a thesaurus may increase its effectiveness (Salton and Buckley, 1988). In the case of NRLS, if we define our document as the incident report and terms as words, a single instance of a word may be highly rated in TF-IDF calculations, and this does not necessarily lend any meaning to terms. This is particularly problematic when spelling mistakes and errors create rare variants of words, and may even repeat within a document, but be rare in the corpus. With a corpus as large as NRLS for a full year, it is impossible to inspect TDF-IF in enough depth, and it is better suited to situations with fewer documents, such as a review within a clinical service or within hospital. It may be a reasonable statistic for small-scale targeted review, or for use as a weighting for a larger

system. For NRLS analysis, it did not show potential for reducing data to topics of interest across the whole dataset.

9.6 Topic models

Although TF-IDF gives some meaning to terms across documents, it does little to reduce the information in a corpus to a manageable or interpretable amount (Blei et al., 2003). Topic modelling has often used a form of term frequency, but has been combined with the more formal statistical modelling approach of Latent Semantic Indexing (LSI) (Deerwester et al., 1990), that uses TF-IDF as a weighting. LSI was criticised as lacking a reason for its use of TF-IDF, when Bayesian methods and maximum likelihood could be used more directly on the text (Blei et al., 2003). LSI was, in turn, developed to 'probabilistic LSI,' where terms are viewed as multinomial random variables drawn from a mixture distribution within documents, and can be considered a latent class model (Hofmann, 1999). LSI was considered an important step in topic modelling, but its limitation was its lack of document-spanning structure.

9.7 Latent Dirichlet Allocation (LDA)

LDA (Blei et al., 2003) was proposed as a development from LSI and other earlier topic models, using the 'bag of words' paradigm, where word order and semantics were not considered. LDA is an unsupervised, generative probabilistic model, with a three-level structure of words, topics and documents (Cao et al., 2009). Words are distributed within topic, and topics are distributed across documents. The technique assumes that topics across documents, and terms within topics have a Dirichlet distribution, and word/token counts within documents are Poisson distributed (Wilson and Chew, 2010).

A more formal definition was given by Wilson and Chew (2010), as '*...The LDA algorithm models the D documents in a corpus as mixtures of K topics where each topic is in turn a distribution over W terms. Given ϑ , the matrix of mixing weights for topics within each document, and ϕ , the matrix of multinomial coefficients for each topic, we can use this formulation to describe a generative model for documents...*'. LDA can therefore be used to examine and associate the frequency of terms across documents and words to give meaning to topics.

LDA was further enhanced by the use of Gibbs sampling rather than expectation maximization (EM or VEM as originally proposed), at the cost of computation time but some increase in precision (Griffiths and Steyvers, 2004).

LDA does not use term weighing, so we return to the argument for removing stop words to prevent bias, as all word are weighted the same in LDA. Several alternatives have been suggested to further enhance LDA by include term weighting, including: TF-IDF (Truica et al., 2016) despite Blei et al's objections, and Point-wise mutual information (PMI) (Wilson and Chew, 2010). The driving reason for these enhancements is to reduce the reliance on arbitrary stop words lists, but none of these approaches are currently accepted as ideal or easily available in common software implementations of LDA.

The number of topics to derive from a given corpus is also an open question. Several suggestions have been made, including using expectation maximisation, minimisation techniques focussed on topic distance (Cao et al., 2009), Kullback–Leibler divergence (Arun et al., 2010), or Gibbs sampling techniques over the posterior distribution of topics (Griffiths and Steyvers, 2004). Topic selection appears to be a case of trial and error in analyses, guided by some of these approaches.

9.7.1 Using LDA to predict incident harm-level

The 'tidy' dataset used in the word frequency and data preparation stages was used as the basis for LDA, however, the common implementation in R's `topicmodels` package (Grün and Hornik, 2011) requires a 'Document-term' matrix as an input rather than tidy data structures. This is a matrix where rows represent each document, in this case 'document' was each incident, and the columns are counts of the tokens/words. The `tidytext` package (Silge and Robinson, 2016) was used to convert between formats, after initial processing in tidy format.

Number of topics, as described above, is somewhat arbitrary and topics numbers were examined between 5 and 200, with 4 main metrics provided using the `ldatuning` R package (Murzintcev, 2019). Gibbs sampling approach was used to fit models due to its greater accuracy, and VEM fitting exceeded available memory on the largest available server.

LDA models can be used to predict on two levels within a corpus:

- β : the matrix of per-topic-per-word probabilities. These can be used to find the most predictive words within topics.

- γ : The matrix of per-document-per-topic probabilities. These can be used to examine the probability a given topic represents a document.

LDA topic outputs (γ) were then considered as predictors for a multinomial classification problem to predict the harm level of incident reports. The main approaches for classification models were chosen as they are common machine learning methods for multinomial classification. All techniques could be progressed into finer tuning of models, but simple options were used for initial runs to assess whether modelling was possible. Modelling techniques used were:

- **Naïve Bayes Classifier (NBC):** A common first approach is using a Naïve Bayes classifier. Naïve Bayes is a family of simple classifiers that are ‘naïve’ in the sense that they assume all measurements are independent between classes, ignoring higher order interactions. ‘Bayes’ refers to deducing the class membership from the probability of belonging to each class using Bayes rule, based on the product of their univariate marginals (Hand and Yu, 2001). NBC scales well to large datasets, but can be accused of over-simplification, can show poor performance around decision boundaries, and down weight rare classes in multinomial problems (Rennie et al., 2003). The ‘standard’ implementation in R uses the `e1071` package (Dimitriadou et al., 2018) and this was used to fit NRLS LDA models.
- **Multinomial Logistic Regression:** Logistic regression modelled (GLMs) can be extended from binary classes to multinomial classes, where a reference level is set, and coefficients for each level of outcome represent difference from the reference level. There are several ways to estimate this kind of regression, but a common R implementation is using a neural network with a single hidden layer, made available in the `nnet` package (Venables and Ripley, 2002), that was fitted to topic predictions.
- **Least Absolute Shrinkage and Selection Operator (LASSO) regression:** L1-norm regularized regression that constrains the absolute sums of the regression coefficients to be less than a fixed value, is a common in high dimensional settings. This uses shrinkage to penalise model coefficients, as far as zero some in some cases, and therefore performs both shrinkage and parameter selection (Tibshirani, 1996). NRLS topic models were fitted and the lasso penalty, λ , was chosen using 10-fold cross-validation within Tibshirani’s own `glmnet` R package (Friedman et al., 2010).
- **Random Forest (RF):** as explored in Chapter 6, random forests are ensemble learning methods based on averaging many regression/classification trees with bootstrapped samples and bootstrapped predictors. RF models can be used for classification, where each model casts a ‘vote’ for a tree, rather than predicting a variable value as it does

for regression. RF is also tuned using its own out-of-bag error rate. The standard R implementation of RF using Breiman and Cutlers code struggles with larger datasets, so the H2O.ai implementation of RF was used in the `h2o` (H2O.ai Team, 2018) package. As mentioned in Chapter 7, H2O is a machine learning environment, using a java-virtual machine, that focusses on improving speed, scalability, memory usage and parallel processing elements or common algorithms (H2O.ai, 2018).

- **Gradient Boosting (GB):** as explored in Chapter 6, boosting uses regression trees (although it can be applied to many models types) and iterative model fitting, re-weighting on prediction error, and re-fitting the weighted data (Friedman, 2001). Boosting shows good predictive performance on many problems, but has various parameters that required tuning (Vezhnevets and Barinova, 2007). The H2O environment, as mentioned for RF model, was also used to fit GB models.

Both random forest and gradient boosting models were fitted using H2O default of 200 trees, but trees were increased to 1000 to test performance. Using more trees appeared to overfit the data and degrade predictive performance.

Another common approach in classification problems of this type is the ‘Support Vector Machine’ (SVM) (Cortes and Vapnik, 1995), but they were not applied in this case. SVMs attempt to find a hyperplane that maximises the separation between classes, sometimes using the ‘kernel trick’ to expand data into higher dimensions to allow separation (Hastie et al., 2009b). SVM are very effective at this, but they struggle with scale due to the computations required (Hsieh et al., 2014), and were impractical because of this during testing.

9.7.2 Model tuning, fitting and results

Various numbers of topics were tested and compared using the 4 metrics discussed above. Initial models were based on word tokens without removing single letter words, and Figure 4 was used as a ‘range-finder’ to identify numbers of topics. Three of the 4 metrics supported increasing numbers of topics, but one measure (Deveaud) decreasing from 40 topics. The Cao et al.(2009) metric also showed a levelling out at 40 topics, only to further decrease at > 50. Topics numbers greater than 100 showed some increase in the remaining three measures, the curve appears to have flattened, and applying Occam’s razer, we should opt for the simpler solution.

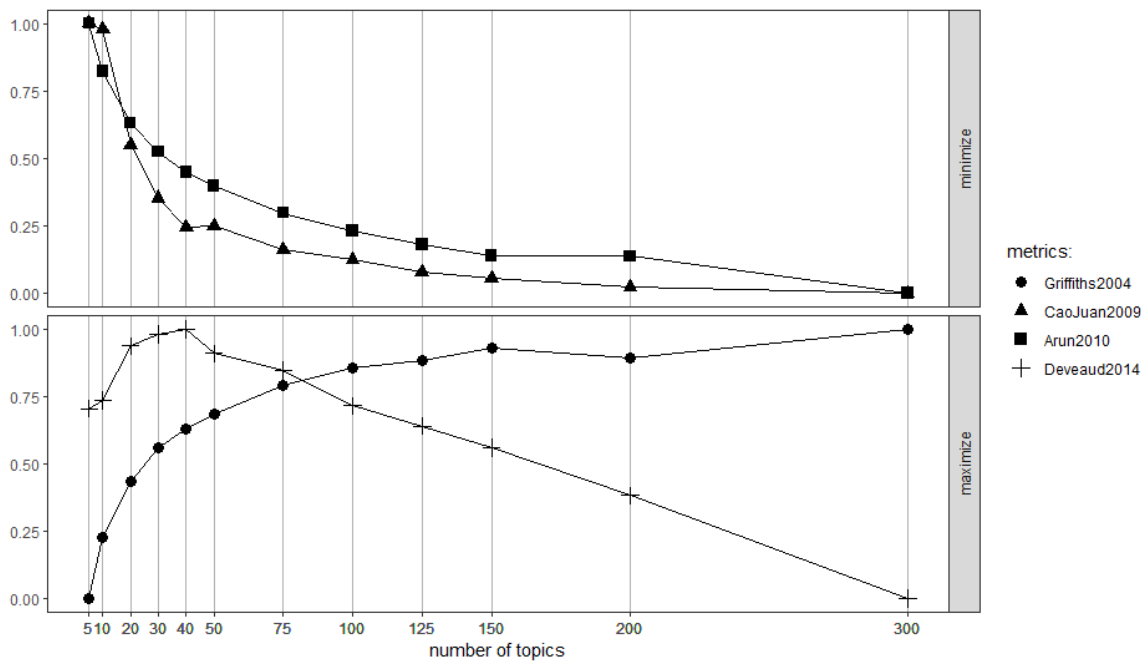


Figure 9.4 LDA metrics for word token models

Word-token based LDA models for NRLS free-text incident descriptions of incidents reporting for incidents occurring in 2015/16. Topic numbers ranges were initially tested between 5 and 300 topics, with metrics explained in section 9.7.1.

Three comparisons were then tested in multinomial models using 40, 100 and 150 topics, based on the four calibration metrics in Figure 9.4.

All techniques showed predictive ability for harm categorisation based on both 40 and 100 topics (Table 9.2). Naïve Bayes performed poorly compared to other tests on sensitivity and overall accuracy. LASSO and multinomial logistic regression showed near identical performance in most cases, as both were performing a multinomial logistic regression by different methods. Highest accuracy and sensitivity were observed in Random Forests models, with boosted trees performing slightly less well. Random Forest also showed the highest sensitivity values for minority classes.

Performance across NBC models was identical for 40 or 100 topics. Small increases in accuracy were observed for 100 topics using for multinomial regression and LASSO, but Random Forests and boosting showed better accuracy using 40 topics. This may suggest that, although 100 topics may be supported by some measures, a degree of overfitting is seen in Random Forest and boosting methods. It also lends support to using the Griffiths and Deveaud metrics.

Topics	Naïve Bayes			Multinomial Regression			Lasso			Random Forest			Gradient Boosting		
	40	100	150	40	100	150	40	100	150	40	100	150	40	100	150
Accuracy															
Total	55.34%	48.08%	44.11%	77.23%	77.52%	77.72%	77.23%	77.52%	77.71%	82.66%	81.40%	80.80%	81.43%	79.61%	79.65%
True Positive Rate (Sensitivity)															
No Harm	55.10%	47.06%	42.19%	95.92%	95.90%	95.93%	95.92%	95.90%	95.94%	96.20%	96.29%	96.91%	96.34%	94.25%	94.71%
Low Harm	62.37%	55.41%	53.66%	24.23%	25.65%	26.37%	24.21%	25.63%	26.34%	46.01%	41.43%	36.94%	41.40%	40.43%	38.84%
Moderate	14.22%	22.21%	24.01%	1.05%	1.04%	1.41%	1.06%	1.01%	1.39%	14.02%	4.74%	2.53%	4.79%	3.37%	4.41%
Severe	9.46%	16.88%	19.85%	0.03%	0.23%	0.25%	0.03%	0.20%	0.23%	25.17%	10.07%	5.24%	10.19%	9.41%	10.22%
Death	56.39%	72.38%	75.44%	1.36%	2.93%	3.33%	1.36%	2.24%	2.93%	39.46%	20.75%	17.21%	21.09%	23.67%	25.85%
True Negative Rate (Specificity)															
No Harm	77.55%	82.65%	84.73%	24.20%	25.46%	26.13%	24.18%	25.44%	82.16%	45.23%	39.81%	35.22%	39.79%	39.61%	38.23%
Low Harm	67.03%	69.89%	70.49%	95.23%	95.22%	95.28%	95.24%	95.22%	63.47%	95.33%	95.45%	96.15%	95.49%	93.32%	93.83%
Moderate	95.29%	93.01%	91.75%	99.91%	99.91%	99.88%	99.90%	99.91%	61.98%	100.00%	100.00%	100.00%	100.00%	99.94%	99.92%
Severe	97.28%	96.38%	95.09%	100.00%	100.00%	100.00%	100.00%	100.00%	83.92%	100.00%	100.00%	100.00%	100.00%	99.99%	99.99%
Death	94.06%	86.39%	83.93%	99.99%	99.98%	99.98%	99.99%	99.98%	88.58%	100.00%	100.00%	100.00%	100.00%	99.99%	100.00%

Table 9.2: Results of multi-class prediction models, using word tokens, for level of harm

Results based on LDA models of NRLS free-text description of incident reports, using for 40, 100 and 1550 topics.
Colours gradients are per-row of the table, with red indicating the lowest figures and blue indicating the highest figures

A wider problem in the use of LDA methods for predicting harm in NRLS is the severe class imbalance (Klement et al., 2011) within the data (table 9.3). This was also an issue in Chapter 7 in the context of sparse/low counts of death and severe incidents. This is common to all rare events/low counts in prediction settings, not just Poisson models (e.g. logistic regression (Harrell et al., 1996)). Algorithms commonly maximise the accuracy of a given classifier but, with severely imbalanced problems, this amounts to predicting the majority class (Drummond and Holte, 2005). This problem is particularly pronounced in these models, as the minority classes are so small compared to the majority ‘No Harm’ events. In this situation, the model is driven by classifying ‘No Harm’ and may perform poorly on the minority groups, yet still be globally accurate. In the extremes of this case, we could simply assume all incidents are ‘No Harm’ and be globally accurate. This accuracy is sometimes referred to as the ‘No Information Rate.’ The overall accuracy of our model predictions and the generalization to one class can be turned into a one-way binomial significance test. This test was significant at >99% suggesting that, despite the extreme class imbalance within the data, the model was significantly more accurate than classifying all cases as ‘No Harm.’ Table 9.3 shows that, classifying all incidents as no harm would lead to an accuracy rate of 75.02%, the proportion of the data represented by the majority class

Harm level	Incident reports in training set	
	Count	Percentage (%)
No Harm	833,678	75.02%
Low	240,244	21.62%
Moderate	31,920	2.87%
Severe	3,934	0.35%
Death	1,470	0.13%
Total	1,111,246	100.00%

Table 9.3 Distribution of incident reports in harm classes

Counts and percentages of NRLS incident reports at different harm levels for 2015/16.

Topic modelling was then repeated using skip-grams, constructed from with bi-grams only (i.e. fitting only two-word tokens). Topic numbers were selected from tuning plots (Figure 9.5) over a range of 10 – 80 topics. This more limited range was chosen due to the word token model performing best at 40 topics. An assumption was made that a latent structure of approximately 40 topics would be approximated in all LDA models. The skip-gram model tuning plot was less clear than the word-token plot, with the tuning metrics in Figure 9.5 showing less agreement. A range of topic numbers from 20 – 50 was therefore fitted to test

this. Models were refitted using just random forests and gradient boosted trees with results shown in table 9.4

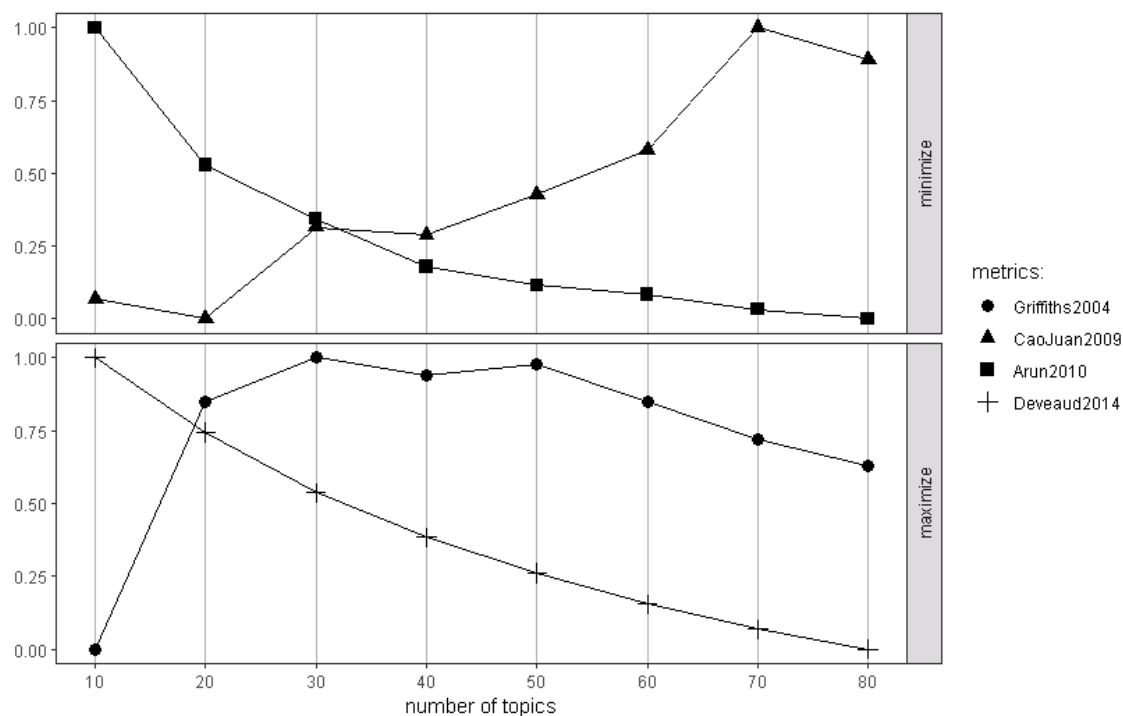


Figure 9.5 LDA metrics for skip-gram(bigram) token models

Skip-gram (bigram) token based LDA models for NRLS free-text incident descriptions of incidents reporting for incidents occurring in 2015/16. Topic numbers between 10 and 80 were tested, using the metrics explained in section 9.7.1.

Topics	Technique							
	Random Forest				Gradient Boosting			
	20	30	40	50	20	30	40	50
Accuracy								
Total	74.21%	74.04%	74.01%	73.97%	73.48%	73.49%	73.52%	73.54%
True Positive Rate (Sensitivity)								
No Harm	99.03%	99.20%	99.26%	99.41%	98.79%	98.94%	98.94%	99.04%
Low Harm	7.20%	6.08%	5.85%	5.21%	5.32%	4.88%	5.01%	4.75%
Moderate	3.12%	2.26%	1.87%	1.57%	0.23%	0.26%	0.36%	0.42%
Severe	5.11%	3.73%	3.26%	2.70%	1.12%	1.20%	1.12%	1.20%
Death	9.72%	7.15%	5.70%	4.58%	3.46%	3.13%	2.68%	3.24%
True Negative Rate (Specificity)								
No Harm	7.54%	6.33%	6.03%	5.35%	5.59%	5.13%	5.24%	5.00%
Low Harm	98.77%	98.97%	99.04%	99.21%	98.50%	98.68%	98.68%	98.79%
Moderate	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Severe	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Death	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 9.4 Results of multi-class prediction models, using skip-gram tokens, for level of harm

Results based on LDA models of NRLS for 20,30,40, and 50 topics. Colours gradients are per row of the table, with red indicating the lowest figures and blue indicating the highest figures

Random Forests again performed best, with an overall accuracy higher than boosting. Random Forest also showed better performance in sensitivity and specificity in minority classes. In this case, 20 topics was the best performing random forests. As topic number increase, both random forest and boosting methods appear to focus more on the majority class ('No Harm') with accuracy increasing, but performance decreasing in relation to minority classes. Performance for skip-gram/bi-gram models was considerably lower than for word-token based models, with overall accuracy approximately 10% lower for random forest and performance in minority classes worse. All models were significantly better than the no information rate, despite being poorer than the word-token model.

Models were then further enhanced by including categorical predictors from the record-level NRLS data. Categorical variables selected and included were:

- Location (level1) – with example values of 'Residence / home' and 'General / acute hospital'
- Location (level2) – with example values of 'Inpatient areas' and 'Outpatient department'
- Location (level 3) – with example values of 'Ward' or 'Radiology' (but may be NULL)
- Specialty (level 1) – with example values of 'Medical specialties' and 'Obstetrics and gynaecology'
- Specialty (level 2) – with example values of 'Cardiology' and 'Obstetrics'

The model matrix was expanded with variables coded as factors using dummy variables. Models were refitted as random forests and gradient boosted trees, using the word-token model with 40 topics. The `H2O.ai` package was used due to the number of variables and category levels. Model output is summarised in table 9.5.

The inclusion of categorical predictors did not substantially improve model performance. Overall accuracy for Random Forest improved slightly but dropped for boosted trees. Random forest performance improved slightly, for moderate and severe harm, in terms of sensitivity with specificity still 100%. It also improved notably for 'No Harm', the majority class. A small percentage change in this class represents many more cases than the other classes and is likely to be driving the overall accuracy. Boosted models, again, tended to focus on predicting the majority class and the minority class prediction suffered.

	Random Forest 40	Gradient Boosting 40
Topics		
Accuracy		
Total	83.68%	80.20%
True Positive Rate (Sensitivity)		
No Harm	97.46%	99.20%
Low Harm	46.00%	6.08%
Moderate	16.73%	2.26%
Severe	26.21%	3.73%
Death	37.69%	7.15%
True Negative Rate (Specificity)		
No Harm	7.54%	6.33%
Low Harm	98.77%	98.97%
Moderate	100.00%	100.00%
Severe	100.00%	100.00%
Death	100.00%	100.00%

Table 9.5 Results of multi-class prediction models, including NRLS categorical data, for level of harm

Results of LDA models with additional NRLS categorical predictors, for 40 topics.

9.8 Conclusions

This chapter has taken a different approach to aim 3 of this project, compared to chapters 4-8, by examining analysis methods for the free-text descriptions of incident reporting. Despite a few successful attempts (Altuncu et al., 2018, Bentham and Hand, 2012, Mayer et al., 2017), quantitative analysis of free-text in NRLS is not in common use. This chapter lays groundwork for simple approaches to term frequency measures and topic modelling that might be applied by NHS organisations or regulators without significant academic experience of text mining.

NRLS free-text data requires significant cleaning before use. Issues of spelling error, synonyms and medical jargon (other than mapping the word 'patient') have not been addressed. The techniques applied in this chapter are comparatively basic and not NHS/NRLS specific. The mapping of drug terms, such as official BNF names, and the development of specific term maps for clinical specialties would aid discrimination. The development of such resources, or pre-trained corpuses, would be a next useful step, and would also benefit other areas of health-related research and could be used with electronic patient records and other sources.

Term frequency at different harm levels, or TF-IDF within a definable subset, will be useful in clinical review of a specific setting, but the volume and variety of words is less useful for general analysis at national level. These tools would be useful for a subject specific, or organisation/setting specific review and indicate the important words driving particular sub-

groups of incidents. The use of these techniques should be examined by NHS organisations and regulators to aid their established patterns of review.

Word-token (unigram) models performed best when used for text modelling and, comparing the wordclouds in Figures 9.2 and 9.3, gave a greater variety of terms within each group (denoted by the colours gradients). Skip-grams appeared to focus on ‘pressure ulcers’ and ‘blood pressure’ that dominated plots. This is an interesting finding, given the lack of incidents classified as ‘pressure ulcer’ from the ‘incident type’ category (Chapter 3), and supports the view prevalent in the literature (Chapter 2), that categorisation is a poor description of many incidents.

LDA topic modelling proved useful for generating themes in the text despite a lack of specific NRLS mappings. These models, when based on word-tokens, performed surprisingly well for predicting harm levels. Skip-grams performed more poorly, but this may not be such a surprise given the dominant terms described in the last paragraph. This may have degraded the fit, losing nuance and making the skip-gram models focus on terms more predictive of the majority class. Other techniques for topic generation could be considered as next steps, or investigating different tokenisation schemes, as initial investigation of simple skip-grams did not improve models. Models could be extended to other years, either as a larger single corpus or modelling individually within years. These models do not necessarily have to build on year-long datasets, but a year served as a useful sample to develop these techniques.

The inclusion of extra NRLS categorical factors with LDA derived topics could be considered a better model if prediction of the minority classes (such as moderate, severe harm or death incidents) is of higher value than general accuracy. This will ultimately be dictated by the intended use of models. If their use is for basic analysis of all events, validation of the whole dataset, or to act as an input filter (where incident report forms might suggest a level of harm to the user from their text description, as suggested by other literature (Mastodon C, 2015, Mayer et al., 2017)), overall classification without the categorical predictors may be better. If the intended use is for the detection and potential reclassification of the minority classes, then the model with additional categorical data would be preferred. Simple categorical factors were included, due to complications with missing data etc., but it would be possible to include seasonality, time of day and other data types into the models in principle. The main hurdle with these data fields is the data quality, as described in Chapters 2 and 3, rather than analysis techniques.

The class imbalance in the data was a major hurdle for predicting harm in these models. This will be true for any prediction model with such an imbalance (Drummond and Holte, 2005).

Options to increase performance of models in these settings often use resampling techniques such as bootstrapping (already performed by the random forest model) or techniques such as up/down sampling or Synthetic Minority Over-sampling techniques (SMOTE) (Chawla et al., 2002, García et al., 2012). The major draw backs of techniques such as SMOTE is that they make more sense in binary classification settings and they may underestimate the features in the majority class if down-sampling is used. SMOTE techniques were attempted with the 40 topic word-token dataset, before reclassification using random forests and boosting, but significant accuracy was lost with only small gains in minority classes and therefore unsuitable for these models.

Text-based quantitative analysis is a viable option for NRLS data and has been demonstrated here and by others, as described in this chapter, further address aim 4 of this project. NHSI, and other investigators, would be advised to invest the minimal resource required to use these techniques. They are not restricted to harm prediction, and could be used to predict other variables, or in a generative manner. This may bear fruit in identifying incidents groups, patterns not detected elsewhere, or aid current review processes by providing targets for clinical review and associated resources.

Chapter 10 Development of a reporting tool

10.1 Introduction

The sponsors of this PhD project (my employers), stipulated the outputs should be used to develop tools to be incorporated into their online NHS benchmarking tool, Healthcare Evaluation Data (HED) (University Hospitals Birmingham NHS Foundation Trust, 2019). HED is primarily based on HES data (as described in Chapter 4). HED also incorporates, and links to, other datasets including the NRLS, with most data refreshed monthly. The system is used by approximately 60 NHS organisations at the time of writing. Access controls limit the levels of data access available to individual users. HED allows national, regional and local comparisons of indicators with Trusts able to identify patient records within their own organisation if required. Data are processed and presented in interactive, point-and-click 'modules' to allow users to analyse, visualise and extract data. This tool provides a means for users to access model outputs from chapters 4-8 and addresses aim 4 of the project.

10.2 Designing and building reporting processes

The following chapter describes the process developed from the methods developed in previous chapters. A major requirement for this process is to be documented and transparent for other members of staff to use, maintain and further develop. HES data are downloaded and processed, by several members of the UHB team, as part of HED's regular monthly update process. HES base data were used from this point, and NRLS data are downloaded and processed entirely for the purposes of this module.

10.2.1 Software architecture for designing a report

The structure of the HED system can be summarised as having two main analysis layers:

- **Dashboard layer:** Pre-processed data, hosted on an SQL Server instance, accessed and presented in tables or graphical form. These tables respond to the organisation of the user and present suitable overviews of indicators and trust position relative to the national distribution. This is comparable to the CQC's Intelligent Monitoring/Insight reports.
- **Module layers:** Interactive data reports including a range of visualisations, tables and text areas. These areas can be controlled by sets of filter panels (controlling the data

available within module) and selection boxes controlling the focus and aggregation levels of tables/visualisations.

HED already has modules for examining NRLS data as published by NHSI. The aggregation level of these data, including the bed-day calculations from KH-03, limit the ability to further examine it. To develop a standardised incident reporting model, with access to individual data, both a module and indicator for the dashboard layer are required. The following requirements were specified by the HED team and shaped the build process:

- The ability to load data either monthly or quarterly, in-line with HES publications.
- Loading and refresh scripts should be self-contained, with minimal interventions from analysts, allowing various team members to run the loading/update procedure.
- Statistical modelling should be explained in a manner that is accessible and available to the users, and:
- Modules should be developed in Spotfire, using HED standard filter panels, and point-and-click elements linking tables/visualisations.

At the time of writing, HED and NRLS are in the process of negotiating a new data access agreement. Current data access rules extend to the end of 2016/17, but 2017/18 and 2018/19 data are yet to be received. The module and dashboard indicators have been developed on 2015/16 and 2016/17 data, but data structures have been designed to be extended when new data are available.

Processes required to receive, load, model and deploy new data can be summarised as:

- Download monthly NRLS extracts from NHSI secured FTP site
- Resolved delimiting issues using SAS (as described in Chapter 3)
- Loading text files to database servers and constructing base tables
- Constructing a modelling dataset
- Running models and building CUSUMs
- Loading data and modelling output into modules

The requirements were then tackled in three main steps:

1. Building/coding and documentation for Microsoft SQL Server stored procedures
2. Building/documentation and an R package
3. Building a working module

10.2.2 Building and coding SQL Server stored procedures.

Stored procedures are saved T-SQL scripts that can be recalled and run by users or automated by servers. They can receive input parameters, perform any valid T-SQL actions, and can return tables and parameters. Stored procedures are commonly used for processes that are repeated or performed regularly. They allow the database server to cache query plans and statistics that can be used to optimise the procedure for the future.

The scripts created for these procedures are not included in this thesis, as they are unlikely to be of interest to the reader and contain significant internal server details for HED that pose security issues. A summary of each procedure and the relevant SQL constructs is detailed below, with a summary in table 10.1, and a full flow chart detailed in Appendix D:

1. Download monthly files in comma-delimited format from NRLS secured FTP server using Mozilla FileZilla. Data are processed by NRLS using SAS, the conventional .csv format is used with free-text surrounded by double quotes (i.e. “free-text”). As detailed in Chapter 3, this holds for text that does not contain double quotes or invalid characters. SAS is intermittent in its application of double quotes, and this prevents other systems loading files manually. A SAS macro, originally described in Chapter 3, was written to load, reformat and export as files in tab-delimited format. Tabs are invalid in NRLS free-text fields, so this was robust for import. A second phase of this development work will automate these actions from the command line using batch files, but it is currently triggered manually.

Stage	Technology Used	Script/File Name
1	FileZilla FTP SAS	Download & Reformat NRLS text file
2	MS SQL Server	Database server load of files and base-table build
3	MS SQL Server	NRLS aggregation, joining with HES data and collation of report tables
4	R	Load and format modelling data, model and predict
5	MS SQL Server	Upload predictions to database server and build CUSUMs
6	TIBCO Spotfire	Present data in interactive module

Table 10.1 Summary of stages to build HED NRLS reporting module

2. The first database processing section involves two distinct scripts:
 - a. [Update_NRLS_1_load_data] - The year and month of the datafile are supplied as a parameter and the parameters are used to construct dynamic

SQL queries to build table names containing the month and year. Data are then uploaded to this table using the 'BULK INSERT' syntax.

- b. [Update_NRLS_2_clean] - A parameter control is used to determine whether to load just the new data or reload all data sequentially to a single NRLS base table. A de-duplication procedure checks for Incident IDs that are already present in the table, copying duplicated records to a duplicated table, and retaining the most recent record in the main table. This base table is used for the following models but is also available for other applications within the HED system for future development.
3. The second database stage takes the base table, selecting just acute hospital data, then builds aggregates, summarises HES data and then creates the hybrid dataset:
 - a. [Update_NRLS_3_aggregate] - The procedure creates an aggregated incident reporting table by hospital, age, sex, year, month, and harm level.
 - b. [Update_NRLS_4_build_base_tables] - This procedure is the most extensive in the process. It further aggregates the NRLS by trust, year, month and harm level. HES inpatient data is joined to a calendar table and used to identify bed-days and their demographics as detailed in Chapter 5. Organisation naming is then resolved to the organisation at the end of the financial year, and the two tables joined to form a modelling table. HES outpatient and A&E attendances are summarised by ages, naming is resolved and then joined back to the modelling table. A&E waiting time percentiles are calculated and joined back to the main table.
 - c. [Update_NRLS_5_Final_Dataset] - The modelling table is reformatted, with relevant reporting measures added.

Tables produced are available for import into R for modelling. UHB's plans are to move R execution in the SQL Server environment, but this is reliant on the commissioning of new servers that are not in production at the time of writing.

10.2.3 Modelling procedures and creating an R-package

R has been used for building all models used in this thesis, and as an exploratory or analysis tool, it is sufficient to work in R-script format. This is not robust for regular model building, reproduction, fault tolerance, and deployment to other users. R code is commonly built into an R package (as discussed in Chapter 5) for this purpose. An R package contains several structures (Wickham, 2015):

- A NAMESPACE file that declares the package name and dependencies
- A DESCRIPTION file that details the function, author, source and licence
- R scripts formatted as functions. Functions are encapsulated code that can take input parameters, perform actions, and output parameters in a similar manner to the SQL stored procedures.
- Support and details files that are generated for each function, explaining expected inputs/outputs, what the functions do, and further details such as examples of use.

The R scripts must be manually built and tested, with parameters at the top of the script that are interpreted by the package building utilities ('devtools' (Wickham et al., 2018)) and turned into the support files. The easiest approach to package building that reduces development time and sources of error is to use the `roxygen2` package and RStudio that, together, automatically format and rebuild parts of the process. This approach was taken to create a package called `SIRRmodels`. This package contained two scripts:

- `run_models()` - a wrapper function for the modelling process. This function picks up the modelling table from the database server, formats some of the R-specific elements (such as factor encoding) that cannot be performed in SQL Server, parallelises model build by modelling year. GLMM, GAM and random forest models for all incidents and death/severe incident models are built, as detailed in previous chapters. Conditional and marginal model predictions, and model metrics such as MAE are exported to csv files.
- `funnel_plot()` - a wrapper for plot functions use in Chapter 8 to provide ad-hoc reporting facilities. This function allows the use of overdispersed or traditional Poisson control limits and the highlighting of outliers on plots.

Once this package was built and tested, a departmental decision was taken to replace it with a single departmental R package, `HEDfunctions`, based on `SIRRmodels` format but extended to other HED processes as well. This has been adopted as the main R-modelling package in use and contains a variety of other functions for statistical modelling, plotting, confident interval estimation and z-scoring methods. The package has been committed to source control software 'git,' and hosted internally on a source control server. UHB assert's its intellectual property rights regarding this code, and I have been unable to include it in the thesis or release it as open-source.

10.2.4 Final processing

Data are reloaded, after completion of the R modelling function, to the SQL Server Database location where it can be accessed for loading into a live server environment. A load script retrieves the model prediction outputs and loads them, performs a final cleaning routine consisting of formatting labels for data manipulation, and adds metadata for servers.

Cusums are then built in the SQL Server environment. These metrics are a technical challenge for processing. In the R environment, row-comparison operations (comparing the last row with the current one) are simple to construct, but this is not the case in SQL Server. SQL Server is highly optimized for join and column-level calculations, but several options exist for tackling this problem:

- An SQL ‘cursor,’ a type of iteration method that is designed for stepping through sequences and they are commonly used for administrative operations like shrinking tables or rebuilding indexes. Cursors can also be used to step through rows of tables, but the common belief is that this is slow and can have unintended consequences. Training material for SQL Server commonly treats this as a ‘last resort.’
- Loop functions can use variables to iterate through rows, saving values to variables for use in the subsequent iteration.
- A SQL construct known colloquially as the ‘quirky update.’ This loophole in the T-SQL syntax allows a user to both assign a value to row in a table and to a separate variable simultaneously without extra processing overhead, i.e. using: row value = variable name = calculated value.

All three options described above were examined and tested for efficiency, with cursors and loops similar in processing time, but the ‘quirky update’ taking less than 2% of the run time of the other techniques. The ‘quirky update’ syntax was therefore implemented for cusums built from model outputs.

10.2.5 Construction of analysis module

Modules are designed in a desktop tool that allows data exploration and structuring of visualisations and filters.

Two data tables were loaded containing the SIRR data and cusum data. The size of data tables was small, at less than 10Mb, and below the threshold that causes load time issues. The data were therefore embedded in the reports without issue.

10.2.5.1 HED design specification

HED modules have conserved formats and sets of design standards. These include:

- A structured cover page with an 'Overview', 'Usage instruction' and a detail 'Analyst summary' with extensive notes relating to the data processing, modelling and exclusions.
- HED logo branding, standard descriptions, and order of pages. (E.g. first analysis page contains a left aligned filter 'tab' that can be shown or hidden by users clicking on appropriate buttons).
- Filter panels contain standard, consistent filters including time-period in a hierarchy of fiscal year, fiscal quarter and month, organisation with hierarchies for regions and a mechanism for pre-set peer groupings.
- Pages of the module should proceed left to right from highest level of aggregation to lowest, with the final page reserved for record-level data extraction.
- Selection of data items on visualisation or tables should highlight an item and either restrict data in subsequent visualisations, or highlight it, depending on the context.
- Visualisations are rarely anchored to particular data items or groupings, and usually allow users to select from a list. In this case, this rule was broken, as the overdispersion requirements of the funnel plots meant restricting them to plot by organisation only.

10.2.5.2 Final spotfire module construction

The construction and flow of the module can be seen in a set of screen shots in Appendix B, with filter panels visible, and two organisations highlighted to demonstrate the marking/restriction of HED modules. The content of each page is as follows:

1. Cover page (not numbered in the module by convention)
 - a. A summary of the SIRR indicator and it's intended use
 - b. A set of instruction/directions for use of each subsequent page
 - c. An 'Analyst Summary,' giving an extensive explanation of the data used, restrictions, assumptions and modelling techniques. This is to allow analyst users to further interpret what figures may mean.
2. All incident models
 - a. Filter panels (controlled and hidden with 'action' buttons). Creating filter panels involves first creating data hierarchies and using normal Spotfire functions, and aliasing values, such as month numbers, with text values.

Hierarchies can then be added as filter items. Items in filters can be clicked by users to select or deselect them in various groupings.

- b. A summary table with aggregated bed-days, observed incidents, expected incidents and SIRR at trust level.
 - c. A funnel plot based on a scatter plot of $x = \text{Expected incidents}$, $y = 100 * (\text{observed/expected})$. The y axis is scaled to centre on 100, matching other standardised indicators available in HED, with a centre line drawn for reference and gridlines hidden to avoid confusion. Funnel limits are drawn in two ways, the 99% Poisson limit by a pre-calculated lookup table, and overdispersed limits by calculating the τ^2 value from the data and inflating a log transformed limit calculation based on the SHMI methodology (explained in chapter 8). An 'action' button triggers the calculation of τ^2 , due to complications of Spotfire's API (explained in detail below).
 - d. A cusum plot of values for organisations selected in the funnel plot or summary table. This entails a further scatter plot, with Spotfire set to join consecutive points with a line, rather than using a 'line' plot. This plotting method allows easier control of the trigger limit. Calculated values for triggers were used to control the point size and the point shape to highlight triggers with a larger pointer of a different shape and colour. This trigger value of 5.4 was included (see Chapter 8).
3. Death or Severe Harm models
 - a. A summary table, similar to the previous page, focussed on DS incidents rather than all incidents.
 - b. A funnel plot, similar to the previous page, focussed on DS incidents rather than all incidents. The calculation of τ^2 , for this plot was also linked to the action button on the previous page to align both plots simultaneously.
 4. Comparison of marginal and conditional models
 - a. A scatter plot of SIR values calculated, for all incidents, with both marginal (y-axis) and conditional (x-axis) predictions. This allows comparison of national-average SIRRs (marginal model predictions) and SIRRs with local adjustments for culture (conditional model predictions).
 - b. A second scatter plot, as per 4.a), but using DS incidents.
 5. Extraction of record level data.
 - a. For users with appropriate access, an export table linked to both their login ID and a checking clause to ensure they have selected their own trust. Data will only be displayed for a user's own organisation and is subject to filter

selections. Users can directly export these records to 'csv' format using 'right-click' options.

10.2.5.3 Calculating overdispersion

The calculation of τ^2 , presents a technical issue for Spotfire, as it needs to be related to the user-selected filter option in the filter panel. Users are encouraged to use 12-month periods but have the facility to change this to match specific analysis needs. Spotfire can automate recalculation of values based on filter panels, but this is comparatively slow and appears as a latency for users. An alternative solution is to link a recalculation script to an 'action' button that requires the user to click on it to recalculate the funnel plot limits. This approach required writing a script using the programming language 'Iron Python' (a .NET implementation of the Python programming language) to take 'snapshot' of the data and run the calculation described in Chapter 8. The τ^2 , is then used to expand the plot funnel limits.

10.3 Module release review and update

The module will be released to users, and debuted at the HED user group tentatively scheduled for 2020, pending the renewal of UHB's data sharing agreement with NHSI and NHSI's approval. At the time of writing, the module is not currently live. Modules are released in HED to a 'Pre-release' section, noting that they are new and requesting comment from users. After three months of publication, feedback from users will be sought and comments used to shape a review of the modules function and form.

The replacement for the NRLS is scheduled for roll-out during 2019. This may be an impediment to promoting a module based on older data, but until new data are made available and differences understood, this module will remain available in the HED system.

10.4 Summary

The models and methods demonstrated in Chapters 5 - 8 can be readily applied to live analytics platforms such as HED that allow users, whether NHS trust staff or regulators, to investigate their data. The strengths of this approach are the visualisation of large volumes of incident data, the ability to highlight areas of interest and view them in several visualisations, ultimately exporting data as required for further investigation.

This module also provides the HED system with a rational, evidence-based case mix adjusted indicator for comparison with crude reporting rates reported elsewhere. HED has no plans to

charge for accessing these data, and charges only a subscription fee to an organisation to cover the cost of software, hardware and resources for running the HED system. All current HED users (with sufficient information governance approvals) will be able to view the SIRR module. Data export options are only enabled for staff with approval of their NHS Trust Caldicott.

This chapter, although technical in nature, demonstrates the process of turning secondary care data research into practical use for hospital monitoring and learning from incident reporting. The majority of SIRR module development was conducted in isolation, as part of my PhD work, and coded in SQL, R, SAS and Iron Python by me. A small number of sections interface with other HED processes and dataflows, with some code reuse for HED procedures such as encryption. A full description of authorship of work is included at the start of this thesis.

Chapter 11 Discussion

Medical error and unsafe care have been globally recognised as leading causes of harm in healthcare (Jha et al., 2013). Measurement of error is a major issue, with incident reporting being recognised as a process to help organisations learn from error (Donaldson, 2002, Berwick, 2013). These systems also have their short-comings (Pham et al., 2013), but the UK's experience is now over 13 years old and has provided a wealth of information for the NHS. The National Reporting and Learning System (NRLS), the NHS repository for incident reports in England and Wales, has been examined in this thesis. This final chapter summarises the analyses conducted and their outcomes, and places them in the wider context of incident reporting. It also makes recommendations for both NRLS' structure/development and analytical processes. It addresses the aims of this thesis, stated in the introduction (Chapter 1), by:

1. Using a literature review to identify prior analysis of NRLS data, it's strength and limitations (Chapter 2).
2. Investigating the structure and possible parametrisations of the NRLS for quantitative analysis from the literature review, direct analysis of the NRLS and pre-modelling work on parametrisation and construction of an aggregated dataset (Chapters 2,3 & 5).
3. Identifying appropriate statistical modelling methods both theoretically and by applying and testing count-based statistical methods, free-text analysis methods, and assessing their strengths and limitations (Chapters 3-7 & 9).
4. Examining reporting and presentation methods that could be used for analysis by submitting organisation, regulators and researchers. This includes representation as standardised ratios and display in funnel plots, using cusums for time series monitoring, developing an online reporting tool, and by examining wordclouds, term-frequency measures, and prediction of harm categories in free-text (Chapters 8, 9 & 10).

11.1 Incident reporting in the context of patient safety

In assessing patient safety, we might pose the simple question: are health services becoming safer or not (Vincent et al., 2008)? Our definition of safety in healthcare is constantly evolving (Pedersen, 2016), with the idea of safety difficult to define, and sensitive to cultural factors and time periods (Vincent and Amalberti, 2015). Despite more patient safety interventions and programmes, trends in adverse events do not appear to be reducing, and it is unclear

whether this is due to increased incidence or improvements in methods of observation and detection (Shojania and Thomas, 2013). It is unclear whether or not many patient safety interventions and policies are effective and whether this can be measured or distinguished from larger secular trends (Benning et al., 2011). Whilst there appears to be some correlation between higher reporting rates and other markers of safety culture (Hutchinson et al., 2009), the claim that this is fact (Macrae, 2016), is not well supported with evidence. Depending on the measures of adverse events, the number of cases qualifying may be vastly different (Classen et al., 2011). Studies have suggested that, despite increases in overall reporting rates, reductions in preventable incidents have been elusive (Benning et al., 2011, Baines et al., 2013, Landrigan et al., 2010).

Under-reporting is a major issue for incident reporting systems (Pham et al., 2013), and this reduces our ability to make inference or estimate the true levels of adverse events (Noble and Pronovost, 2010), and they may not be suited to this task (Sari et al., 2007). Nonetheless, incident reporting systems have led to a greater understanding of patient safety and direct action in the NHS (Franklin et al., 2014, Panesar et al., 2009).

Analysis of adverse events/medical error in NHS hospitals has tended to focus on clinical case-note review techniques (Carson-Stevens et al., 2015, Hogan et al., 2012, Vincent et al., 2001), or has used these techniques to infer overall error rates through methods such as the Global Trigger Tool (Landrigan et al., 2010).

There is an important distinction to be made between 'hard' and 'soft' data related to incident reporting (Samuriwo et al., 2016). There is much depth, nuance and information to be gleaned from in case-note reviews, particularly related to 'soft' data (Martin et al., 2015). It appears common opinion that narrow reviews, in great depth, are of more value than shallower, wider-ranging techniques (Vincent, 2004). 'Hard data' and quantitative methods in this field have not been well developed, partly due to the scale of the task and the challenges with the quality and classification of the data (Howell et al., 2017, Carson-Stevens et al., 2018).

The methods presented in this thesis use the 'hard' data from NRLS, including treating the free-text as 'hard' data. They go further than those previously presented (Pham et al., 2010, Howell et al., 2015, NHS Improvement, 2017c, NHS Improvement, 2017b, Stuttford et al., 2018, Panesar et al., 2013b, Wahr et al., 2014) and develop a casemix-adjustment method for comparing observed reporting rates against expected reporting rates. These methods have been presented as indirectly-standardised ratios (referred to as a 'Standardised Incident Reporting Ratios' (SIRR)). This thesis examines how SIRRs can be used in-line with current

regulatory frameworks for cross-sectional comparison and monitoring, as well as text-mining methods for NRLS. Vincent (2007) lamented the lack of formal testing of methods related to incident reporting, and the methods presented here answer this need by providing tools for quantitative analysis of incidents, including statistical process control techniques previously advocated for learning from error (Battles and Stevens, 2009).

11.2 NRLS data set structure

The structure of the NRLS has been investigated primarily through literature review (Chapter 2) summary statistics and analysis (Chapter 3). This structure has shown significant weaknesses and incompatibilities with large-scale numerical analyses, mainly related to under-reporting and data quality. Major issues include:

- The definitions and perception of incidents is not clear. There are differing perceptions of whether certain events class as 'incidents' at all and whether staff considered reporting them, or even knew how to (Evans et al., 2006).
- Many barriers to staff reporting exist (Pinto et al., 2012), but fundamentally, staff are pressed for time and incident reporting is of lower value than direct patient care activities. Incidents may silently take their toll on staff members, without these effects being recognised as secondary incidents in themselves (Quillivan et al., 2016), further reducing staff inclination to report. Fear of blame, what Reason termed the 'blame trap' (Reason, 1990), and disciplinary action is a key barrier (Cooper et al., 2017, Radhakrishna, 2015), recently highlighted by the high profile case of Dr Hadiza Barwa-Garba (Cohen, 2017) whose private reflections of practice were indirectly provided to prosecutors. Barriers may be exacerbated by organisational cultures and differences between staff groups (Braithwaite et al., 2008), and generic reporting systems/forms that are overly long or ill-suited to particular clinical settings/specialties (Scott-Warren et al., 2012).
- The non-mandatory nature of reporting (Francis, 2013). Simply increasing reporting level is not an adequate method to adjust for under-reporting (Williams et al., 2016). Vincent has suggested that, with hindsight, a systematic data collection would have made more sense than voluntary reporting (Vincent, 2007).
- Missing data items are a huge problem for analyses, affecting case ascertainment (MacLennan and Smith, 2011) and the ability to analyse reports effectively (Hignett et al., 2013).

- Inflexible and misclassified categorical data were highlighted in many publications. Incidents often relate to multiple settings, specialties and the underlying processes may be multifaceted (Fowler, 2013). Classifying a drug error on a ward round as a 'medication error' might miss the fact that low staffing was an issue leading to the distraction of a nurse, or a dose-form/labelling issues may have led to the error. This is also a 'staffing' incident, and potentially an 'equipment' incident, but the current structures do not allow more than one classification. The studies of higher quality assumed this was the case and looked, for example, at anaesthetics and surgical reports when trying to find incidents related to anaesthesia (Thomas and McGrath, 2009). The text mining techniques presented in Chapter 9 offer methods to mediate some of these issues by generating latent structures from topic models with probabilities for each topic, rather than dichotomising.
- Anonymisation is a cornerstone of patient (and staff) data practices in the NHS for good reason. A just and open safety culture necessitates a degrees of protection for staff and anonymisation is a key principle in protecting patient identifiable data (NHS Digital, 2018b). These practices also make information governance and sharing of incident data easier. There are two key problems with anonymisation in NRLS reports: firstly, identification of repeated incidents for the same patient, such as choking (Guthrie et al., 2015), self-harm (James et al., 2012) or attempted suicide (Bowers and James, 2011), is impossible. Knowing the number of patients these incidents relate to is key to understanding and learning from them. Secondly, linking data at patient-level is impossible, and this prevents fully patient-centric case-note reviews. Using identifiers, with proper access controls and information governance arrangements, would substantially strengthen the data asset, allowing patient related incidents to be tracked through systems from primary care, hospitals, and electronic prescribing for instance.

NRLS-related literature suggested that the main 'signal' in NRLS data is contained within the free-text descriptions of incidents (Mayer et al., 2017, Howell et al., 2017, Evans et al., 2019). Making use of these data currently requires clinical review and qualitative techniques, taking time and considerable clinical expertise that could arguably be better spent delivering care. If NRLS classification structures were reviewed, and better suited to multiple categorisation, they could save time and help target review and analysis activities.

Despite the limitations of the NRLS in its raw form, the strategy advanced in this thesis is to add value to the NRLS dataset by augmenting it with data from other sources. When discussing the frustrations of incident reporting systems, Shojania (2008) lamented the lack of denominators when dealing with incident reports. This thesis has advanced this area by providing casemix adjusted denominators. Techniques are advanced by creating an aggregate 'panel' dataset with incident counts per organisation per month, combined with similarly aggregated casemix variables from Hospital Episode Statistics (HES). Chapters 5 and 10 detail the creation and processing cycles for the dataset, and they could be applied by other analysts/organisations by following the methods described. These approaches for combining datasets could be readily applied to other datasets that might benefit from augmentation for casemix-adjustment.

11.3 Statistical models build, and overdispersion

Models were based on count data using the Poisson regression model as a basic framework (Chapter 5). The major limitation with Poisson regression models is overdispersion.

Overdispersion was considerable in the dataset, and appeared to arise from:

- **Clustering/repeated measurements at trusts** - demonstrated by the success of random-intercept models, allowing clusters to vary from the global intercept (national average).
- **The aggregation processes** - supported by the success of the negative binomial models, that weighted variance at small and large organisations differently. A single incident is a larger proportion of the outcome for a small organisation than it is for a large one.
- **Inadequate specification of predictors** - Predictors were not necessarily characterised in the best form, and are likely to be proxies for unmeasurable factors. E.g. age and comorbidity score may be a proxy for 'unwellness.' Some of the shared, unmeasured characteristics will be absorbed into the intercept, and random-intercept terms, but better specification may yield better predictions.
- **'Noise' / random variation in the dataset**

Poisson models, without random-intercepts, were improved upon by using quasi-likelihood and negative binomial models. These models gave better estimates of the variance within the data at the cost of increased bias. The inclusion of a random-intercept greatly improved the models and allowed the scaling in the negative binomial models to focus on the effects of

aggregation. The structure of a Standardised Incident Reporting Ratio has some precedent, with other studies also using bed-days as a proxy for size and random-effects for clustering (Landrigan et al., 2010, Baines et al., 2013).

Generalized Additive Models (Chapter 6) improved upon the regressions in Chapter 5 by allowing smoothed predictors (more representative of the overall trend / less 'noisy') to be fitted instead. When combined with the random-intercept, and negative binomial distributions, these models appeared to give the best performance of all models (for the total incident reports model). Using multidimensional smooth terms to model related factors degraded model predictions due to overfitting, and the experience of these models suggests that simpler, independent, smooth terms were preferable. Overdispersion affected the estimation of how smooth models should be, and the introduction of extra penalty terms improved fit on testing data.

Several machine learning techniques, regression trees, boosted/bagged trees, random forest and artificial neural networks (ANN) were also fitted to the data. Random forests are a promising technique, but are comparatively rare in medical research compared to traditional regression techniques. Although random forest have recently been applied to similar problems (Cafri et al., 2018). Random forests performed well, due to their resampling and decorrelating properties, but gave notably different output predictions to the other maximum likelihood-based GLM(M) and GAM models, when compared at trust level (Chapter 8). The lack of distribution assumptions, that is sometimes a strength of algorithmic learning methods, may have been the downfall in this application, as the strength of the Poisson assumptions helped regression models. GLMM and GAM techniques outperformed these approaches in general, likely due to the properties of count data and the distributional assumptions. ANNs did not show any strengths over GLMM/GAM approaches in this setting. ANNs are complicated non-linear models that hold the potential for better performance, but they may have been limited by the modestly sized training data set, and inexperience in their tuning.

Model assessment was performed using Mean Absolute Error (MAE) on a testing dataset (2016/17 fiscal year). This was the preferred approach available to check that model generalised underlying relationships to new data (Chai and Draxler, 2014), but MAE may still be deficient as a loss function. It is possible that data from 2016/17, used as the testing set, is

spurious in some way and could give misleading output. This is supported by the drop in incident reports, due to reporting cut-offs described in Chapter 3. Models could be further tested using data from other periods and greater use of cross-validation or bootstrapping approaches. Although this was a pragmatic approach to choosing the best predictive model, its output is relative, allowing us to choose the best model from a set of candidates. A comparison against error based entirely on incidents per bed-day (similar to that used in CQC's key lines of enquiry), with no casemix-adjustment, showed casemix-adjusted models to be substantially better in terms of prediction error. We do not, however, know whether a model is 'acceptable' or 'good.' This can only be judged through external validation and comparison to other indicators, which is the next logical progression of this work.

Once developed on total incident reporting, model structures were retrained on death or severe harm incidents only (Chapter 7). These rare events created a sparsity problem, and fitting approaches demonstrated very similar fits between complex models and substantially reduced models. The model's reliance on the seasonality, measures of organisational size and random-intercepts suggested that that these incidents were not well predicted by case mix variables, but the effects of overdispersion were not as strong as those in the total incident models. The reduced overdispersion could, however, be an artefact of the sparsity. Lilford et al.(2010) suggested it is difficult to use data based on small numbers for evaluating policy, or applying causal inference in healthcare and, in this case, it may be difficult to judge whether any intervention related to these incidents is successful in bringing down rates. They apply this rationale to the Harvard Medical Practice study (Brennan et al., 1991), suggesting that the changes in rates of death even when halving the number of adverse events, is small. It also suggests that undue focus on these events, rather than on specialties where harm is rare but incidents are common, may overlook important issues affecting many more patients (Shojania, 2012). Does this mean we should abandon these methods for death or severe incidents? The results here suggest not, but caution should remain around them, and SIRR methods may require further development before their strengths and limitation are understood. The small numbers of these events mean that it remains practical to clinically review all reports, even given resource constraints. Quantitative analysis techniques could be used to bring additional value to monitoring at trust level, identifying targets for clinical review, or deriving new learning from events using the text mining techniques.

Chapter 8 examined how these methods might interface with current regulatory indicators and monitoring processes, focussing on NHS Improvement and the Care Quality Commission's techniques. The chapter demonstrated how a casemix-adjusted reporting indicator is treated under this framework, where this fits NRLS data and where it does not. The model predictions were used to generate casemix-adjusted marginal predictions that corresponded to the 'expected number of incident reports' per trust, per month. Predictions were then included as part of a ratio of observed to expected, where a value of 1 represented the same numbers of observed and expected, and is referred to as the 'Standardised Incident Reporting Ratio' (SIRR). Techniques used to examine these indicators included z-scoring and funnel plots, both of which used an additive overdispersion model calculated in a post-hoc fashion. Both techniques showed promise for identifying variation greater than expected, but care may be required around the presentation and use of such indicators to avoid the appearance of league tables. The valid interpretation of these techniques is assessing how far a data point is from the expected range (0 for z-score, 1 for funnel plot), not comparing organisation with each other. Which technique to use is a choice for regulators, as both carry useful information, but the aim of such indicators is to be used in a process control fashion, identifying organisations with significantly different reporting behaviour. Methods published by CQC suggested a square-root transformation to SIRRs before calculating the overdispersion elements. A log-transformed version of these techniques described elsewhere yielded better results due to more correction of the lower tail of the transformed distribution. This suggests that CQC may wish to examine whether their transformation of other indicators is also similarly affected.

Chapter 8 also explored the use of Cusum techniques for monitoring changes in the SIRR over time. These techniques adjust for clustering within organisations over time, but their compatibility with SIRRs is questionable. Calibration of these plots was set, as per published CQC practice, to a doubling of the odds compared to the reference rate (national rate), but this may be better set to a tripling. This would be in-line theoretically with 3σ limits used in the funnel plots, and may be a more appropriate standard. A change such as this would require adjustments to trigger values, and may require simulation studies, particularly in the case of DS incidents. The utility of cusum techniques is not clear from initial work in this thesis. Despite conforming to the frameworks described in regulatory information guidance, other control chart approaches or monitoring techniques may be more appropriate for these data. Cusum techniques require more work on calibration and external validation before they could be recommended for use in monitoring NRLS incident reports.

The text mining approaches examined in Chapter 9 demonstrate the ease with which such techniques can be applied, without NRLS-specific developments. Topic generation using LDA was combined with random forests for prediction of harm levels, achieving high accuracy, but driven significantly by the class imbalances, as death and severe harm are rare compared to 'no harm.' Text mining is recognised as the approach with the most potential for large-scale analysis of incident data (Howell et al., 2017). If not adopted for primary analysis, it could be used as a useful secondary source/validation for assessment of harm. Studies have shown disagreement between the recorded level of harm and the level of harm assessed from clinical review or incidents (Thomas et al., 2002), and these techniques may aid consistency.

The next steps with such approaches are to continue using the bag-of-words models with NRLS-specific dictionaries and mapping of terms. This could be achieved in-part through the use of nearest neighbour techniques on projected features, or using word2vec/text embedding approaches, to predict the 'correct' terms for abbreviations or mis-spellings. Mapping of specific medical terms could be achieved by using external data sources such as SNOMED (Ruch et al., 2008), or drug names, trade names and abbreviations using the British National Formulary (BNF) (Joint Formulary Committee, 2019). E.g. mapping 'Ventolin' to 'Salbutamol.' Such mappings will greatly increase the coherency and consistency of topics, as well as aiding analyses that use TF-IDF approaches for understanding key terms in the description of subsets of incidents.

Chapter 10 describes the processes associated with turning research into usable tools to allow NHS organisations to benefit from these analyses. The initial interactive module is planned for launch using the Healthcare Evaluation Data (HED) benchmarking tool, pending approval from NHS Improvement. This approach directly moves research in this thesis into practice and makes it accessible to NHS trusts.

Much of the work in this thesis would benefit from further external verification. At the time of writing, I am engaging with NHSI to present and demonstrate this work. Outlier organisations from funnel plots, z-scoring approaches and use of text mining could be validated by comparison with NHSI's other workstreams, clinical expertise and experience, as described in Chapters 8 and 9. These validation process would likely include:

- **Standardised ratios:** Quantitative comparison with other proxy indicators such as staff survey results, staff sickness, other clinical indicators, and CQC ratings. Qualitative

comparisons based on interviews or surveys examining reporting culture. These measures could help validate whether outlier organisation really do show differences in reporting culture when compared to others.

- **Text mining methods:** in the examples demonstrated in chapter 9, visualization methods can aid understanding of incident groups, but the major emphasis was prediction of harm. This was tested within the chapter on both the training and testing sets. Further planned validation of these methods should include examining reports from each of the projected 40 topics to identify what each topic represents. Further development work, such as identifying incident clusters should be examined by comparison with clinical review of the same reports.

In any validation studies, or operational use of these indicators, care should be taken to avoid gaming or unhelpful incentives or penalization of organisations (Lilford et al., 2004).

11.4 Applications in NHS organisations and barriers

To use the models described in this thesis in practise, they may be applied in either a regulatory/monitoring capacity, or further targeted to identify and learn from error.

The methods in this thesis, particularly the GAM and random-intercept models require a significant understanding of statistics and this may present barriers for their use in local NHS organisations. As a member of a team in an NHD Informatics services, I am fortunate to have had the investment in my PhD to learn these skills, but the majority of NHS informatics staff would likely require significant training to implement them locally. They would also struggle to do so without the use of HES data, which is not routinely available to NHS organisations. I attempted to address this lack of skill and resource by building the interactive 'module' in Chapter 10, as this provides access to the data in a controlled manner without the burden of applying the statistical methods to the raw data.

The text mining methods are an area of growth for NHS organisations at present. Some teams have, or are developing these skills, but it is not yet part of routine work. I have presented my work to NHS Improvement, who are keen to apply text mining methods, but I am prevented from publishing further technical details at present due to UHB asserting their intellectual property. I hope to further this work by publishing academic papers. text mining methods hold the most potential for benefitting patient care, as they can be used to highlight reports for review, or find patterns, that are not currently seen.

11.5 Wider applications

This work has shown that it is feasible to use multilevel modelling approaches for hospital administrative datasets. Multilevel modelling appeared to be the right choice for NRLS models using aggregated data. A similar analysis of adverse events in hospitals in the Netherlands (Baines et al., 2013) used GLMMs to account for repeated measures, including a centring of predictors in a uniform manner for all years, conceptually similar to that performed on the training/testing sets in Chapters 5-8. Cultural differences at organisations appear to be obscured by the scale of the overdispersion but, when adjustments are made with multilevel models, patterns can emerge from the noise (Baines et al., 2015). The use of these techniques for directly estimating random-intercepts and clustering could be employed more widely in NHS national indicators such as SHMI or HSMR (Campbell et al., 2012, Jarman et al., 1999) and may lead to more robust models. The techniques currently used ignore clustering or adjust with the comparatively crude post-hoc fix (see Chapter 8). These approaches could be considered for other indicators, such as the SHMI which is currently under review (Clinical Indicators Team, 2019).

A similar case could be made for the use of GAMs, that have the ability to fit ‘noisy’ data with smoothed predictors. The ability to select and compare the smoothness of these models is a major attraction, and they allow additional penalties to impose further smoothing if necessary. These methods are currently being applied in my work with HED, modelling emergency readmissions to hospitals. They are unlikely to be published due to commercial interests but have shown better performance than logistic regression in this setting.

The development of text mining techniques using bag-of-words methods should prompt the creation of medical dictionaries/lookup tables that would be directly relevant to the analysis of text. Text data are present in other areas of healthcare (and more widely), including patient notes, patient feedback or other electronic health record elements. The incentive to develop these resources may have been lacking until recent years, given the perceived complexity. Some of this could be achieved through the methods described above but would require external validation from professional groups including clinicians, clinical coders and researchers. Such a dictionary/reference source would quickly be adopted and spur other research. The increasing use of electronic health records, social media and ‘app’ data, and advances in data management and analysis (through ‘big data’ techniques and the ‘AI’/‘Deep Learning’ boom) may drive the call for such a reference.

The data processing and construction of the hybrid dataset described in Chapter 5 also suggested several smaller points related to the use of HES. Firstly, the lack of a universal ‘bed-day’ calculation is a huge oversight for the NHS, given its need to manage capacity and quality. The current standard methods are either inequitable, ignore short stays and day cases, or add arbitrary biases. The calculation proposed here is equally biased, but better describes exposure than other alternatives. This will not be solved until times of admission and discharge become mandatory in HES. Secondly, A&E waiting times appear to be unvalidated, and it is surprising that so many hospitals clearly stop monitoring time (or sending it to SUS/HES at least) after 4 hours. This may be remedied by the planned replacement A&E dataset that is yet to be rolled out (the Emergency Care Dataset or ‘ECDS’), but this should be a known data quality warning for all users of national A&E HES.

11.6 Future NRLS analyses

This thesis suggests that data quality in NRLS, models based on aggregated NRLS data, and methods for handling clustered/overdispersed data, would all be improved by better characterisation of predictors (referred to as ‘feature engineering’ in machine learning literature). This could involve more in-depth exploratory analysis of sources such as HES, transformation/projection techniques to describe latent variable predictors or drawing on additional datasets. Additional data sources may also shed light on incident reporting at trusts. The most obvious missing data source in terms of exposure variables is the ‘critical care’ HES dataset. This forms a separate dataset from other HES sources and was not included for practical reasons around data sharing agreements and a lack of data warehousing at the start of the project. Critical care has been referenced in many of the articles discussed in Chapter 2, and the intensive intervention and high clinical risk in this setting may be a strong risk factor for incidents. This may not influence predicted values, but it may remove some of the uncertainty in models that is currently absorbed into the intercept term or the random-intercepts. It may, therefore, reduce the overdispersion. Other data sources such as litigation rates or staffing data from electronic staff records may also increase predictive ability in these models. Various sources of information on adverse events exist (Hogan et al., 2008), not just incident reports, and methods should be further expanded to consider other data.

A further alternative is to consider if the methods of comparison presented here, using predicted values, might be better conveyed by using other metrics such odds ratios, or by examining estimates of the random-effect for identification of outliers.

A major issue in future NRLS reporting work is the future of the NRLS itself. During the course of this project, and the transfer of NRLS legacy systems between different NHS bodies, a review process for NRLS has been held. A new system, the Development of the Patient Safety Incident Management System (DPSIMS) project (NHS Improvement, 2017a) is in-progress and building a replacement for the NRLS and the StEIS system (used for reporting 'never events'). It is being built using Agile development methodologies (Agile delivery community, 2016), a software development process that focusses on capturing 'user stories' to drive development requirements, prioritising fast development and deployment. It is currently in the 'beta' phase described by Agile techniques (NHS Improvement, 2018), where a working system is being built, and then tested at scale. This phase was scheduled to end in February 2019, but there have been no further releases of project details, or technical specifications, at the time of writing. At this stage, it is not clear how these methods will fit with the new reporting system, but it is likely they could be adapted.

The further development of text-based models is a major target. As demonstrated elsewhere (Mayer et al., 2017) text mining models are being taken seriously for aiding data entry and classification of incidents. These techniques could provide a welcome boost to data quality and save clinical teams valuable time. They also risk undermining the same processes by making default answers too easily accepted and may encourage a 'path of least resistance' for incident reporting. It will be important to test these text mining-based approaches over time to ensure they are still representative of incident descriptions, rather than being dominated by default values. Care must also be taken around variables dominated by large class imbalances, such as the no harm incidents discussed in Chapter 9. Further development of these models using under or over sampling, with greater use of specific mapping tools, or even larger corpora (potentially using neural networks with text embedding methods) may increase the sensitivity to minority class groups.

11.7 Recommendations

The following list summarises recommendation for NHS Improvement and DPSIMS for incident reporting structures and analysis techniques:

1. Reduce the reporting routes for incidents, all of which should be mapped to the NRLS database to make a single central repository. This will allow central analysis of

incident data from an authoritative source (with DPSIMS already aiming to include never events as well).

2. Make incident reporting mandatory at all level of harm, from all NHS organisations, integrated care, and private health and care organisations. This is to say that, where they are already reported locally, they should be mapped to the national system, and where systems are absent, they should be implemented.
3. Implement more rigorous data validation rules that reject reports with missing or invalid fields, with the onus on providers to resubmit. This will reduce uncertainty in modelling and analysis.
4. NHSI should encourage software providers, and local organisations, to consider the burden of incident reporting on staff with regards to their time. Electronic systems that allow easy reporting, including text-based prediction of categorical variables, should be implemented.

Two recommendation for all standardised ratio indicators where clustering and overdispersion is likely, including the current NHSD SHMI review (Clinical Indicators Team, 2019):

5. Examine multilevel modelling approaches including random-intercepts for clusters. Relying on post-hoc techniques to assess overdispersion is an indirect and imprecise approach, particularly when relying on transformations and asymptotics. Marginal models may still be used for presenting funnel plots in the currently accepted fashion.
6. CQC would be advised to consider whether square-root transformation is appropriate for all the standardised ratios they monitor. The log-transformation techniques associated with SHMI performed better for SRRs, and may suggest squared-transformed indicators are not strict enough.

A final recommendation for NHS Digital, regarding SUS and HES, is:

7. Admission and Discharge times should become mandatory for all providers to solve the deficient bed-day calculations currently used.

11.8 Final comments

The NRLS and its DPSIMS replacement are vast and underused resources for patient safety learning. A reluctance to tackle the systematic issues around classification and missing data have led to the view that it cannot be analysed quantitatively. I believe my work has demonstrated this to be false. The methods presented here are not avant-garde, but

systematic application of common statistical processes. If DPSIMS is constructed effectively, quantitative techniques may advance significantly using its data. This could free up clinical time and help direct analysis to areas that may be missed through sampling methods for audit alone.

The task of monitoring patient safety for teams at NHSI, and CQC, is an unenviable one. They are expected to monitor and identify 'bad' organisations, spot errant practitioners, spot deterioration and prevent poor care. They are expected to do this amid criticism from all sides, and they are underfunded given those expectations. The systems and constraints they work under may be old and in need of investment, and DPSIMS is a welcome development. We must, however, consider how data are used, as well as how they are entered. Building infrastructure to extract maximum benefit from DPSIMS would include more clinical review teams, but also investment in analytical and data science functions. The implementation of new techniques may be hampered by a lack of technical skill or awareness in some areas, or a lack of will to implement them. To improve the use of incident reporting that aids patient safety, analytical resource, not just clinical resource, should be considered.

Text-analysis may be the most exciting element of this project, with substantial promise to generalise to much of the electronic patient record agenda and increase benefits for patients and organisations. The NHS and regulators should invest in these skill sets to make use of data, rather than relying on external consultants, and build a sustainable set of expertise to tackle these developments. The increase of digital connectivity, and electronic health data only makes this need more pressing.

References

- ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G. & ISARD, M. Tensorflow: a system for large-scale machine learning. *OSDI*, 2016. 265-283.
- ADERY, C. A. H. 1968. A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30, 582-598.
- AGILE DELIVERY COMMUNITY. 2016. *Agile and government services: an introduction* [Online]. GOV.UK: Government Digital Services. [Accessed 11/03/2019 2019].
- AKAIKE, H. 1998. Information Theory and an Extension of the Maximum Likelihood Principle. In: PARZEN, E., TANABE, K. & KITAGAWA, G. (eds.) *Selected Papers of Hirotugu Akaike*. New York, NY: Springer New York.
- AKRAM, A. R. & HARTUNG, T. K. 2009. Intercostal chest drains: A wake-up call from the National Patient Safety Agency rapid response report. *Journal of the Royal College of Physicians of Edinburgh*, 39, 117-120.
- ALLAIRE, J. & CHOLLET, F. 2018. keras: R Interface to 'Keras'. 2.2.4 ed. CRAN: CRAN.
- ALTMAN, D. G. 1990. *Practical Statistics for Medical Research*, Taylor & Francis.
- ALTUNCU, M. T., MAYER, E., YALIRAKI, S. N. & BARAHONA, M. 2018. From Free Text to Clusters of Content in Health Records: An Unsupervised Graph Partitioning Approach. *arXiv preprint arXiv:1811.05711*.
- ANDERSON, D. R. & BURNHAM, K. P. 2006. AIC Myths and Misunderstandings. Website: Colorado State University.
- ARNOT-SMITH, J. & SMITH, A. F. 2010. Patient safety incidents involving neuromuscular blockade: analysis of the UK National Reporting and Learning System data from 2006 to 2008. *Anaesthesia*, 65, 1106-13.
- ARUN, R., SURESH, V., VENI MADHAVAN, C. E. & NARASIMHA MURTHY, M. N. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In: ZAKI, M. J., YU, J. X., RAVINDRAN, B. & PUDI, V., eds. *Advances in Knowledge Discovery and Data Mining*, 2010// 2010 Berlin, Heidelberg. Springer Berlin Heidelberg, 391-402.
- AUGUIE, B. 2016. gridExtra: Miscellaneous Functions for "Grid" Graphics. 2.2.1 ed.: CRAN.
- BAE, S., FAMOYE, F., WULU, J. T., BARTOLUCCI, A. A. & SINGH, K. P. 2005. A rich family of generalized Poisson regression models with applications. *Mathematics and Computers in Simulation*, 69, 4-11.
- BAINES, R., LANGELAAN, M., DE BRUIJNE, M., SPREEUWENBERG, P. & WAGNER, C. 2015. How effective are patient safety initiatives? A retrospective patient record review study of changes to patient safety over time. *BMJ Quality & Safety*, 24, 561-571.
- BAINES, R. J., LANGELAAN, M., DE BRUIJNE, M. C., ASSCHEMAN, H., SPREEUWENBERG, P., VAN DE STEEG, L., SIEMERINK, K. M., VAN ROSSE, F., BROEKENS, M. & WAGNER, C. 2013. Changes in adverse event rates in hospitals over time: a longitudinal retrospective patient record review study. *BMJ Quality & Safety*, 22, 290-298.
- BAIRD, M. & SMITH, A. 2009. Accuracy of reporters' assignment of patient harm in anaesthetic critical incidents from the UK National Reporting and Learning Scheme. *European Journal of Anaesthesiology*, 26, 204-205.
- BARACH, P. & SMALL, S. D. 2000. Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. *BMJ*, 320, 759-763.
- BARAI, I., HOWELL, A. M., BURNS, E. & DARZI, A. 2014. Are we maximising learning from reported surgical patient safety incidents? An assessment of how accurately the national reporting and learning system classifies surgical error. *International Journal of Surgery*, 12, S80.

- BARDSLEY, M., GEORGHIOU, T., SPENCE, R. & BILLINGS, J. 2016. Factors associated with variation in hospital use at the end of life in England. *BMJ Supportive & Palliative Care*, Online First, 1-8.
- BARDSLEY, M., SPIEGELHALTER, D. J., BLUNT, I., CHITNIS, X., ROBERTS, A. & BHARANIA, S. 2009. Using routine intelligence to target inspection of healthcare providers in England. *Quality and Safety in Health Care*, 18, 189-194.
- BATES, D., MÄCHLER, M., BOLKER, B. & WALKER, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67, 48.
- BATTLES, J. B. & STEVENS, D. P. 2009. Adverse event reporting systems and safer healthcare. *Quality and Safety in Health Care*, 18, 2-2.
- BENNEYAN, J. C., LLOYD, R. C. & PLSEK, P. E. 2003. Statistical process control as a tool for research and healthcare improvement. *Quality and Safety in Health Care*, 12, 458-464.
- BENNING, A., DIXON-WOODS, M., NWULU, U., GHALEB, M., DAWSON, J., BARBER, N., FRANKLIN, B. D., GIRLING, A., HEMMING, K., CARMALT, M., RUDGE, G., NAICKER, T., KOTACHA, A., DERRINGTON, M. C. & LILFORD, R. 2011. Multiple component patient safety intervention in English hospitals: controlled evaluation of second phase. *BMJ*, 342, d199.
- BENTHAM, J. 2010. *Discovering New Kinds of Patient Safety Incidents*. Doctor of Philosophy, Imperial College London.
- BENTHAM, J. & HAND, D. J. 2009. Detecting New Kinds of Patient Safety Incidents. In: GAMA, J., COSTA, V. S., JORGE, A. M. & BRAZDIL, P. B. (eds.) *Discovery Science, Proceedings*.
- BENTHAM, J. & HAND, D. J. 2012. Data mining from a patient safety database: the lessons learned. *Data Mining and Knowledge Discovery*, 24, 195-217.
- BERWICK, D. 2013. A promise to learn – a commitment to act: improving the safety of patients in England. In: HEALTH, D. O. (ed.). London.
- BLACK, N. 2010. Assessing the quality of hospitals. *BMJ*, 340, c2066.
- BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993-1022.
- BOLKER, B. 2017. *lme4 convergence warnings: troubleshooting* [Online]. Rpubs. Available: https://rstudio-pubs-static.s3.amazonaws.com/33653_57fc7b8e5d484c909b615d8633c01d51.html [Accessed 28/10/2017 2017].
- BOLKER, B. M. 2018. *GLMM FAQ* [Online]. 2018. Available: <http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html> [Accessed 11/04/2018 2018].
- BOLKER, B. M., BROOKS, M. E., CLARK, C. J., GEANGE, S. W., POULSEN, J. R., STEVENS, M. H. H. & WHITE, J.-S. S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24, 127-135.
- BOOTH, C. M., MOORE, C. E., EDDLESTON, J., SHARMAN, M., ATKINSON, D. & MOORE, J. A. 2011. Patient safety incidents associated with obesity: a review of reports to the National Patient Safety Agency and recommendations for hospital practice. *Postgrad Med J*, 87, 694-9.
- BOTTLE, A. & AYLIN, P. 2008. Intelligent Information: A National System for Monitoring Clinical Performance. *Health Services Research*, 43, 10-31.
- BOTTLE, A. & AYLIN, P. 2011. Predicting the false alarm rate in multi-institution mortality monitoring. *The Journal of the Operational Research Society*, 62, 1711-1718.
- BOWERS, L., DACK, C., GUL, N., THOMAS, B. & JAMES, K. 2011. Learning from prevented suicide in psychiatric inpatient care: an analysis of data from the National Patient Safety Agency. *Int J Nurs Stud*, 48, 1459-65.
- BOWERS, L. & JAMES, K. 2011. Learning from prevented suicide in psychiatric inpatient care: An analysis of data from the National Patient Safety Agency: Commentary on Bowers et al. (2011) response. *International Journal of Nursing Studies*, 48, 1587-1588.

- BRAITHWAITE, J., WESTBROOK, M. & TRAVAGLIA, J. 2008. Attitudes toward the large-scale implementation of an incident reporting system. *Int J Qual Health Care*, 20, 184-91.
- BREIMAN, L. 1996a. Bagging predictors. *Machine Learning*, 24, 123-140.
- BREIMAN, L. 1996b. Out-of-bag estimation. California: Berkley.
- BREIMAN, L. 1999. Prediction Games and Arcing Algorithms. *Neural Computation*, 11, 1493-1517.
- BREIMAN, L. 2001a. Random Forests. *Machine Learning*, 45, 5-32.
- BREIMAN, L. 2001b. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.*, 16, 199-231.
- BREIMAN, L., FRIEDMAN, J., STONE, C. J. & OLSHEN, R. A. 1984. *Classification and Regression Trees*, Taylor & Francis.
- BRENNAN, T. A., LEAPE, L. L., LAIRD, N. M., HEBERT, L., LOCALIO, A. R., LAWTHERS, A. G., NEWHOUSE, J. P., WEILER, P. C. & HIATT, H. H. 1991. Incidence of Adverse Events and Negligence in Hospitalized Patients. *New England Journal of Medicine*, 324, 370-376.
- BRESLOW, N. E. 1984. Extra-Poisson Variation in Log-Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33, 38-44.
- BRESLOW, N. E. & CLAYTON, D. G. 1993. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88, 9-25.
- BRESLOW, N. E. & DAY, N. E. 1980. *Statistical Methods in Cancer Research, vol1, The Analysis of Case-Control Studies*, Lyon, IARC ; Oxford University Press.
- BURNHAM, K. P. & ANDERSON, D. R. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33, 261-304.
- CAFRI, G., LI, L., PAXTON, E. W. & FAN, J. 2018. Predicting risk for adverse health events using random forest. *Journal of Applied Statistics*, 45, 2279-2294.
- CAMERON, A. C. & TRIVEDI, P. K. 1986. Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics*, 1, 29-53.
- CAMERON, A. C. & TRIVEDI, P. K. 2013a. Basic Count Regression. In: CAMERON, C. A. & TRIVEDI, P. K. (eds.) *Regression Analysis of Count Data*. 2 ed. Cambridge: Cambridge University Press.
- CAMERON, A. C. & TRIVEDI, P. K. 2013b. Generalized Count Regression. In: CAMERON, C. A. & TRIVEDI, P. K. (eds.) *Regression Analysis of Count Data*. 2 ed. Cambridge: Cambridge University Press.
- CAMERON, A. C. & TRIVEDI, P. K. 2013c. Measurement Errors. In: CAMERON, C. A. & TRIVEDI, P. K. (eds.) *Regression Analysis of Count Data*. 2 ed. Cambridge: Cambridge University Press.
- CAMERON, A. C. & TRIVEDI, P. K. 2013d. Model Specification and Estimation. In: CAMERON, C. A. & TRIVEDI, P. K. (eds.) *Regression Analysis of Count Data*. 2 ed. Cambridge: Cambridge University Press.
- CAMERON, A. C. & TRIVEDI, P. K. 2013e. *Regression Analysis of Count Data*, Cambridge University Press.
- CAMPBELL, M. J., JACQUES, R. M., FOTHERINGHAM, J., MAHESWARAN, R. & NICHOLL, J. 2012. Developing a summary hospital mortality index: retrospective analysis in English hospitals over five years. *BMJ*, 344, e1001.
- CAO, J., XIA, T., LI, J., ZHANG, Y. & TANG, S. 2009. A density-based method for adaptive LDA model selection. *Neurocomputing*, 72, 1775-1781.
- CARE QUALITY COMMISSION (CQC) 2014a. Intelligent Monitoring: NHS acute hospitals. Indicators and methodology Guidance to support the July 2014 Intelligent Monitoring update. July ed.: Care Quality Commission (CQC).
- CARE QUALITY COMMISSION (CQC) 2014b. NHS acute hospitals: Statistical Methodology. *Intelligent Monitoring*. Care Quality Commission (CQC).
- CARE QUALITY COMMISSION (CQC). 2016. *Regulation 18: Notification of other incidents | Care Quality Commission* [Online]. CQC. Available:

- <http://www.cqc.org.uk/content/regulation-18-notification-other-incidents> [Accessed 31/10/2016 2016].
- CARE QUALITY COMMISSION (CQC). 2017. *Monitoring NHS acute hospitals | Care Quality Commission* [Online]. CQC. Available: <http://www.cqc.org.uk/what-we-do/how-we-use-information/monitoring-nhs-acute-hospitals> [Accessed 13/11/2017 2017].
- CARSON-STEVENSON, A., DONALDSON, L. & SHEIKH, A. 2018. The Rise of Patient Safety-II: Should We Give Up Hope on Safety-I and Extracting Value From Patient Safety Incidents?; Comment on "False Dawns and New Horizons in Patient Safety Research and Practice". *International Journal of Health Policy and Management*, 7, 667-670.
- CARSON-STEVENSON, A., HIBBERT, P., AVERY, A., BUTLIN, A., CARTER, B., COOPER, A., EVANS, H. P., GIBSON, R., LUFF, D., MAKEHAM, M., MCENHILL, P., PANESAR, S. S., PARRY, G., REES, P., SHIELDS, E., SHEIKH, A., WARD, H. O., WILLIAMS, H., WOOD, F., DONALDSON, L. & EDWARDS, A. 2015. A cross-sectional mixed methods study protocol to generate learning from patient safety incidents reported from general practice. *BMJ Open*, 5, e009079.
- CARSON-STEVENSON, A. P. 2017. *Generating learning from patient safety incident reports from general practice*. PhD, Cardiff University.
- CARTER, A. W., MOSSIALOS, E. & DARZI, A. 2015. A national incident reporting and learning system in England and Wales, but at what cost? *Expert Review of Pharmacoeconomics and Outcomes Research*, 15, 365-368.
- CASSIDY, C. J., SMITH, A. & ARNOT-SMITH, J. 2011. Critical incident reports concerning anaesthetic equipment: analysis of the UK National Reporting and Learning System (NRLS) data from 2006-2008*. *Anaesthesia*, 66, 879-88.
- CATCHPOLE, K., BELL, M. D. & JOHNSON, S. 2008. Safety in anaesthesia: a study of 12,606 reported incidents from the UK National Reporting and Learning System. *Anaesthesia*, 63, 340-6.
- CATCHPOLE, K. & MCCULLOCH, P. 2009. Incidents in anaesthesia: past occurrence and future avoidance. *J Perioper Pract*, 19, 342-7.
- CHAI, T. & DRAXLER, R. R. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geoscientific model development*, 7, 1247-1250.
- CHARLSON, M. E., POMPEI, P., ALES, K. L. & MACKENZIE, C. R. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*, 40, 373-83.
- CHAWLA, N. V., BOWYER, K. W., HALL, L. O. & KEGELMEYER, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- CHOLLET, F. & ALLAIRE, J. 2018. *Deep Learning with R*, Manning Publications.
- CHOULDECHOVA, A. & HASTIE, T. 2015. Generalized Additive Model Selection. *arXiv.org*, 1506.03850, 23.
- CHRISTOU, M. & KONSTANTINIDOU, M. 2012. Safety of offshore oil and gas operations: Lessons from past accident analysis. In: TRANSPORT, I. F. E. A. (ed.). Luxembourg: European Union.
- CLASSEN, D. C., RESAR, R., GRIFFIN, F., FEDERICO, F., FRANKEL, T., KIMMEL, N., WHITTINGTON, J. C., FRANKEL, A., SEGER, A. & JAMES, B. C. 2011. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. *Health Aff (Millwood)*, 30, 581-9.
- CLEVELAND, W. S. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74, 829-836.
- CLEVELAND, W. S., GROSSE, E. & SHYU, W. M. 1992. Local regression models. In: CHAMBERS, J. M. & HASTIE, T. J. (eds.) *Statistical Models in S*. New York: Wadsworth & Brooks/Cole (Springer-Verlag).

- CLINICAL INDICATORS TEAM 2016a. 5.1 Patient safety incidents. *In: NHS DIGITAL* (ed.) Version: 1.3 ed. <https://indicators.hscic.gov.uk/webview/> NHS Digital.
- CLINICAL INDICATORS TEAM 2016b. Indicator Specification: Summary Hospital-level Mortality Indicator *In: DIGITAL, N.* (ed.). London: NHS Digital, 1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE, United Kingdom.
- CLINICAL INDICATORS TEAM 2019. Summary Hospital-level Mortality Indicator (SHMI) -Review Update January 2019. *In: DIGITAL, N.* (ed.). Leeds: NHS Digital.
- COHEN, D. 2017. Back to blame: the Bawa-Garba case and the patient safety agenda. *BMJ*, 359, j5534.
- COLLET, D. 1952. *Modelling Binary Data*, Florida, Chapman Hall/CRC.
- CONWAY, R. W. & MAXWELL, W. L. 1962. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12, 132-136.
- COOPER, J., EDWARDS, A., WILLIAMS, H., SHEIKH, A., PARRY, G., HIBBERT, P., BUTLIN, A., DONALDSON, L. & CARSON-STEVENSON, A. 2017. Nature of Blame in Patient Safety Incident Reports: Mixed Methods Analysis of a National Database. *Annals of family medicine*, 15, 455-461.
- CORNELL UNIVERSITY 1991. Arxiv.org. New York, USA: Cornell University.
- CORTES, C. & VAPNIK, V. 1995. Support-vector networks. *Machine Learning*, 20, 273-297.
- COUSINS, D. 2011. Insulin, hospitals and harm: a review of patient safety incidents reported to the National Patient Safety Agency.
- COUSINS, D., ROSARIO, C. & SCARPELLO, J. 2011. Insulin, hospitals and harm: a review of patient safety incidents reported to the National Patient Safety Agency. *Clin Med (Lond)*, 11, 28-30.
- COUSINS, D. H., GERRETT, D. & WARNER, B. 2012. A review of medication incidents reported to the National Reporting and Learning System in England and Wales over 6 years (2005-2010). *Br J Clin Pharmacol*, 74, 597-604.
- CRITICAL APPRAISAL SKILLS PROGRAMME. 2016. *CASP Checklists* [Online]. Available: <https://www.casp-uk.net/casp-tools-checklists/> [Accessed 2016].
- CUONG PHAM, J. & COLANTUONI, E. 2010. The harm susceptibility model : a method to prioritise risks identified in patient safety reporting systems.
- DALIANIS, H. 2018. *Clinical Text Mining: Secondary Use of Electronic Patient Records*, Cham, Springer.
- DEAN, C. B. & BALSHAW, R. 1997. Efficiency Lost by Analyzing Counts Rather than Event Times in Poisson and Overdispersed Poisson Regression Models. *Journal of the American Statistical Association*, 92, 1387-1398.
- DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. & HARSHMAN, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.
- DENG, L. 2013. *Analysis of Patient-Safety related data using Statistical Modeling*. Lancaster University.
- DEPARTMENT OF HEALTH. 2013a. *Payment by Results in the NHS: a simple guide - GOV.UK* [Online]. www.gov.uk: Department of Health. Available: <https://www.gov.uk/government/publications/simple-guide-to-payment-by-results> [Accessed 10/10/2017 2017].
- DEPARTMENT OF HEALTH 2013b. Step-by-Step guide: Calculating the 2013-14 National Tariff. *In: HEALTH, D. O.* (ed.). Department of Health.
- DEPARTMENT OF HEALTH & SOCIAL CARE 2017. Draft Health Service Safety Investigations Bill. *In: CARE, D. O. H. S.* (ed.). London: UK Parliament.
- DERSIMONIAN, R. & LAIRD, N. 1986. Meta-analysis in clinical trials. *Control Clin Trials*, 7, 177-88.
- DIETTERICH, T. G. 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40, 139-157.

- DIMITRIADOU, E., HORNIK, K., LEISCH, F., MEYER, D., WEINGESSEL, A., CHANG, C.-C. & LIN, C.-C. 2018. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. . 1.7-0 ed. Vienna, Austria: CRAN.
- DONALDSON, L. 2000. An organisation with a memory. *In: DEPARTMENT OF HEALTH* (ed.). London: The Stationary Office.
- DONALDSON, L. 2002. An organisation with a memory. *Clin Med*, 2, 452-7.
- DONALDSON, L. J., PANESAR, S. S. & DARZI, A. 2014. Patient-safety-related hospital deaths in England: thematic analysis of incidents reported to a national database, 2010-2012. *PLoS Med*, 11, e1001667.
- DOWLE, M. & SRINIVASAN, A. 2017. data.table: Extension of `data.frame`. 1.10.4 ed. CRAN.
- DRUMMOND, C. & HOLTE, R. C. Severe Class Imbalance: Why Better Algorithms Aren't the Answer. 2005 Berlin, Heidelberg. Springer Berlin Heidelberg, 539-546.
- DUCHON, J. 1977. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. *Constructive theory of functions of several variables*. Springer.
- DUNN, P. K. & SMYTH, G. K. 2005. Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing*, 15, 267-280.
- EFRON, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7, 1-26.
- EFRON, B. & HASTIE, T. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge, Cambridge University Press.
- EGGER, M., SMITH, G. D., SCHNEIDER, M. & MINDER, C. 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315, 629-634.
- EHRENTAUT, C., EKHOLM, M., TANUSHI, H., TIEDEMANN, J. & DALIANIS, H. 2018. Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. *Health Informatics Journal*, 24, 24-42.
- EILERS, P., H. C. & MARX, B. D. 1996. Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11, 89-102.
- ELITH, J., LEATHWICK, J. R. & HASTIE, T. 2008. A working guide to boosted regression trees. *J Anim Ecol*, 77, 802-13.
- ELSTON, D. A., MOSS, R., BOULINIER, T., ARROWSMITH, C. & LAMBIN, X. 2001. Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology*, 122, 563-9.
- ELWYN, G. & CORRIGAN, J. M. 2005. The patient safety story. *BMJ*, 331, 302-4.
- EVANS, H. P., ANASTASIOU, A., EDWARDS, A., HIBBERT, P., MAKEHAM, M., LUZ, S., SHEIKH, A., DONALDSON, L. & CARSON-STEVENSON, A. 2019. Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ML) approaches. *Health Informatics Journal*, 0.
- EVANS, S. M., BERRY, J. G., SMITH, B. J., ESTERMAN, A., SELIM, P., O'SHAUGHNESSY, J. & DEWIT, M. 2006. Attitudes and barriers to incident reporting: a collaborative hospital study. *Quality and Safety in Health Care*, 15, 39-43.
- FAMOYE, F. 1993. Restricted generalized poisson regression model. *Communications in Statistics - Theory and Methods*, 22, 1335-1354.
- FEINERER, I., HORNIK, K. & MEYER, D. 2008. Text Mining Infrastructure in R. *Journal of Statistical Software; Vol 1, Issue 5 (2008)*.
- FETHERSTON, T. 2007. Risk management, adverse events and litigation in vitreoretinal surgery. *Clinical Risk*, 13, 7-11.
- FISHER, J. D., FREEMAN, K. & CLARKE, A. 2015. Patient safety in ambulance services : a scoping review. *Health Services and Delivery Research*, 3.
- FLOOD, C., GULL, N., THOMAS, B., GORDON, V. & CLEARY, K. 2014. Is knowledge and practice safer in England after the release of national guidance on the resuscitation of patients in mental health and learning disabilities? *J Psychiatr Ment Health Nurs*, 21, 806-13.

- FLOOD, C., MATTHEW, L., MARSH, R., PATEL, B., MANSARAY, M. & LAMONT, T. 2015. Reducing risk of overdose with midazolam injection in adults: an evaluation of change in clinical practice to improve patient safety in England. *J Eval Clin Pract*, 21, 57-66.
- FOWLER, A. J. 2013. A Review of Recent Advances in Perioperative Patient Safety. *Ann Med Surg (Lond)*, 2, 10-4.
- FOX, J. & WEISBERG, S. 2011. An R companion to applied regression, 2d ed. *Reference and Research Book News*, 26.
- FOX, J. & WEISBERG, S. 2012. Bootstrapping Regression Models in R, An Appendix to An R Companion to Applied Regression, Second Edition. *An R companion to applied regression, 2d ed.* 2nd ed. Portland: Ringgold Inc.
- FRANCIS, R. 2013. *Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry : volume 2 : analysis of evidence and lessons learned (part 2)*, London, The Stationery Office.
- FRANKLIN, B. D., PANESAR, S. S., VINCENT, C. & DONALDSON, L. J. 2014. Identifying systems failures in the pathway to a catastrophic event: an analysis of national incident report data relating to vinca alkaloids. *BMJ Qual Saf*, 23, 765-72.
- FREUND, Y. & SCHAPIRE, R. E. Experiments with a new boosting algorithm. Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, 07/03/1996 1996. Morgan Kaufmann Publishers Inc., 148-156.
- FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. 2000. Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, 28, 337-407.
- FRIEDMAN, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29, 1189-1232.
- FRIEDMAN, J. H. 2002. Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, 38, 367-378.
- FRIEDMAN, J. H., HASTIE, T. & TIBSHIRANI, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *2010*, 33, 22.
- GARCÍA, V., SÁNCHEZ, J. S., MARTÍN-FÉLEZ, R. & MOLLINEDA, R. A. 2012. Surrounding neighborhood-based SMOTE for learning from imbalanced data sets. *Progress in Artificial Intelligence*, 1, 347-362.
- GAWKRODGER, D. J. 2011. Risk management in dermatology: an analysis of data available from several British-based reporting systems. *Br J Dermatol*, 164, 537-43.
- GELMAN, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27, 2865-2873.
- GELMAN, A. & HILL, J. 2006a. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge, Cambridge University Press.
- GELMAN, A. & HILL, J. 2006b. Multilevel structures. In: GELMAN, A. & HILL, J. (eds.) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- GOLDSTEIN, H. 2010. *Multilevel Statistical Models*, John Wiley & Sons Inc.
- GOLDSTEIN, H. & SPIEGELHALTER, D. J. 1996. League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159, 385-443.
- GOODFELLOW, I., BENGIO, Y. & COURVILLE, A. 2018. *Deep Learning*, Cambridge, Massachusetts & London, England The MIT Press.
- GOVINDARAJULU, U. S., MALLOY, E. J., GANGULI, B., SPIEGELMAN, D. & EISEN, E. A. 2009. The Comparison of Alternative Smoothing Methods for Fitting Non-Linear Exposure-Response Relationships with Cox Models in a Simulation Study. *The International Journal of Biostatistics*, 5, 2.
- GREENWELL, B., BOEHMKE, B., CUNNINGHAM, J. & DEVELOPERS, G. 2018. gbm: Generalized Boosted Regression Models. R package version 2.1.4 ed. CRAN.
- GREGORY, A. T. & DENNISS, A. R. 2018. An Introduction to Writing Narrative and Systematic Reviews — Tasks, Tips and Traps for Aspiring Authors. *Heart, Lung and Circulation*, 27, 893-898.

- GREVEN, S. 2088. *Non-standard problems in inference for additive and linear mixed models*, Göttingen, Cuvillier Verlag.
- GREVEN, S. & KNEIB, T. 2010. On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, 97, 773-789.
- GRIFFITHS, T. L. & STEYVERS, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101, 5228-5235.
- GRIGG, O. & SPIEGELHALTER, D. 2008. The null steady-state distribution of the CUSUM statistic. *Technometrics*.
- GRIGG, O. A., FAREWELL, V. T. & SPIEGELHALTER, D. J. 2003. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat Methods Med Res*, 12, 147-70.
- GRÜN, B. & HORNIK, K. 2011. topicmodels: An R Package for Fitting Topic Models. 2011, 40, 30.
- GUTHRIE, S., LECKO, C. & RODDAM, H. 2015. Care staff perceptions of choking incidents: what details are reported? *J Appl Res Intellect Disabil*, 28, 121-32.
- H2O.AI. 2018. *H2O: Open Source, Distributed Machine Learning for Everyone* [Online]. Available: <https://www.h2o.ai/products/h2o/> [Accessed 10/12/2018 2018].
- H2O.AI TEAM 2018. h2o: R Interface for H2O. 3.20.0.8 ed.
- HAJJEM, A., BELLAVANCE, F. & LAROCQUE, D. 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84, 1313-1328.
- HAND, D. J. & YU, K. 2001. Idiot's Bayes: Not So Stupid after All? *International Statistical Review / Revue Internationale de Statistique*, 69, 385-398.
- HANDAYANI, D., NOTODIPUTRO, K., ANWAR, SADIK, K. & KURNIA, A. 2017. A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM). *AIP Conference Proceedings*, 1827, 020033.
- HARRELL, F. E., JR. 2001. *Regression Modeling Strategies*, New York, Springer-Verlag New York.
- HARRELL, F. E., JR., LEE, K. L. & MARK, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, 15, 361-87.
- HASTIE, T. 1992. Generalized Additive Models. In: CHAMBERS, J. M. & HASTIE, T. J. (eds.) *Statistical Models in S*. New York: CRC Press LLC.
- HASTIE, T. & TIBSHIRANI, R. 1986. Generalized Additive Models. *Statistical Science*, 1, 297-310.
- HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2009a. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, New York, NETHERLANDS, Springer.
- HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2009b. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. In: SCHEMATIC OF A SINGLE HIDDEN LAYER, F.-F. N. N. (ed.). New York, NETHERLANDS: Springer.
- HAUCK, W. W. & DONNER, A. 1977. Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*, 72, 851-853.
- HAWKES, C., CHAMBERS, S., SATHERLEY, P., SLOWTHER, A., PERKINS, G. & BOOTH, S. 2015. DNACPR incidents and complaints. *Resuscitation*, 96, 18.
- HEALEY, F., SCOBIE, S., OLIVER, D., PRYCE, A., THOMSON, R. & GLAMPSON, B. 2008. Falls in English and Welsh hospitals: a national observational study based on retrospective analysis of 12 months of patient safety incident reports. *Qual Saf Health Care*, 17, 424-30.
- HERBERT, A., WIJLAARS, L., ZYLBERSZTEJN, A., CROMWELL, D. & HARDELID, P. 2017. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *International Journal of Epidemiology*, 46, 1093-1093i.
- HIGGINS, J. P. T., THOMAS, J., CHANDLER, J., CUMPSTON, M., LI, T., PAGE, M. J. & WELCH, V. A. (eds.) 2019. *Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019)*: Cochrane.
- HIGNETT, S. 2013. Inpatient falls: what can we learn from incident reports?
- HIGNETT, S., SANDS, G. & GRIFFITHS, P. 2011. Exploring the contributory factors for un-witnessed in-patient falls from the National Reporting and Learning System database. *Age Ageing*, 40, 135-8.

- HIGNETT, S., SANDS, G. & GRIFFITHS, P. 2013. In-patient falls: what can we learn from incident reports? *Age Ageing*, 42, 527-31.
- HILBE, J. M. 2014. *Modeling Count Data*, Cambridge, Cambridge University Press.
- HILBE, J. M. 2016. COUNT: Functions, Data and Code for Count Data. 1.3.4 ed. CRAN.
- HOFMANN, T. 1999. Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. Berkeley, California, USA: ACM.
- HOGAN, H., HEALEY, F., NEALE, G., THOMSON, R., VINCENT, C. & BLACK, N. 2012. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. *BMJ Quality & Safety*.
- HOGAN, H., OLSEN, S., SCOBIE, S., CHAPMAN, E., SACHS, R., MCKEE, M., VINCENT, C. & THOMSON, R. 2008. What can we learn about patient safety from information sources within an acute hospital: a step on the ladder of integrated risk management? *Quality and Safety in Health Care*, 17, 209-215.
- HOSMER, D. W. & LEMESHOW, S. 1995. Confidence interval estimates of an index of quality performance based on logistic regression models. *Statistics in Medicine*, 14, 2161-2172.
- HOWELL, A.-M., BURNS, E. M., HULL, L., MAYER, E., SEVDALIS, N. & DARZI, A. 2017. Incident reporting: rare incidents may benefit from national problem solving. *BMJ Quality & Safety*, 26, 517-517.
- HOWELL, A. M., BURNS, E. M., BOURAS, G., DONALDSON, L. J., ATHANASIOU, T. & DARZI, A. 2015. Can Patient Safety Incident Reports Be Used to Compare Hospital Safety? Results from a Quantitative Analysis of the English National Reporting and Learning System Data. *PLoS One*, 10, e0144107.
- HSIEH, C.-J., SI, S. & DHILLON, I. S. 2014. A divide-and-conquer solver for kernel support vector machines. *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. Beijing, China: JMLR.org.
- HUBER, P. J. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 1967* 1967 Berkeley, Calif.: University of California Press, 221-233.
- HUNT, T. 2016. ModelMetrics: Rapid Calculation of Model Metrics. CRAN.
- HUTCHINSON, A., YOUNG, T. A., COOPER, K. L., MCINTOSH, A., KARNON, J. D., SCOBIE, S. & THOMSON, R. G. 2009. Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System. *Qual Saf Health Care*, 18, 5-10.
- IAEA 2010. *IRS Guidelines*, Vienna, INTERNATIONAL ATOMIC ENERGY AGENCY.
- IAEA 2015. *Operating Experience from Events Reported to the IAEA Incident Reporting System for Research Reactors*, Vienna, INTERNATIONAL ATOMIC ENERGY AGENCY.
- INNES, J. & CURTIS, D. 2013. Medication patient safety incidents linked to rapid tranquillisation: one year's data from the National Reporting and Learning System. *Journal of Psychiatric Intensive Care*, 11, 13-17.
- IOFFE, S. & SZEGEDY, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.
- JACKSON, C. H., RICHARDSON, S. & BEST, N. G. 2008. Studying place effects on health by synthesising individual and area-level outcomes. *Soc Sci Med*, 67, 1995-2006.
- JAMES, K., STEWART, D., WRIGHT, S. & BOWERS, L. 2012. Self harm in adult inpatient psychiatric care: a national study of incident reports in the UK. *Int J Nurs Stud*, 49, 1212-9.
- JARMAN, B., GAULT, S., ALVES, B., HIDER, A., DOLAN, S., COOK, A., HURWITZ, B. & IEZZONI, L. I. 1999. Explaining differences in English hospital death rates using routinely collected data. *Bmj*, 318, 1515-1520.

- JHA, A. K., LARIZGOITIA, I., AUDERA-LOPEZ, C., PRASOPA-PLAIZIER, N., WATERS, H. & BATES, D. W. 2013. The global burden of unsafe medical care: analytic modelling of observational studies. *BMJ Quality & Safety*, 22, 809-815.
- JOINT FORMULARY COMMITTEE. 2019. *British National Formulary* [Online]. Available: <http://www.medicinescomplete.com> [Accessed 11/03/2019 2019].
- JONES, H. E., OHLSEN, D. I. & SPIEGELHALTER, D. J. 2008. Use of the false discovery rate when comparing multiple health care providers. *Journal of Clinical Epidemiology*, 61, 232-240.e2.
- KELLY, S. P. & BARUA, A. 2011. A review of safety incidents in England and Wales for vascular endothelial growth factor inhibitor medications. *Eye*, 25, 710-716.
- KELLY, S. P. & JALIL, A. 2011. Wrong intraocular lens implant; learning from reported patient safety incidents. *Eye*, 25, 730-4.
- KEOGH, B. 2013. *Keogh review on hospital deaths published - NHSUK* [Online]. Department of Health. Available: <https://www.nhs.uk/NHSEngland/bruce-keogh-review/Documents/outcomes/keogh-review-final-report.pdf> [Accessed 13/11/2017 2017].
- KESKAR, N. S., MUDIGERE, D., NOCEDAL, J., SMELYANSKIY, M. & TANG, P. T. P. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.
- KINGMA, D. P. & BA, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- KLEMENT, W., WILK, S., MICHALOWSKI, W. & MATWIN, S. Classifying Severely Imbalanced Data. 2011 Berlin, Heidelberg. Springer Berlin Heidelberg, 258-264.
- KOHN, L. T., CORRIGAN, J. M. & DONALDSON, M. S. 2000. To Err is Human: Building a Safer Health System. In: KOHN, L. T., CORRIGAN, J. M. & DONALDSON, M. S. (eds.) *To Err is Human: Building a Safer Health System*. Washington (DC): National Academies Press (US), Institute of Medicine Committee on Quality of Health Care in America. Copyright 2000 by the National Academy of Sciences. All rights reserved.
- KUHN, M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28, 26.
- LAKER, M. F. 2009. Improving medication safety in the North-East of England: The potential enabling role of NRLS data. *Practice Development in Health Care*, 8, 54-57 4p.
- LANDRIGAN, C. P., PARRY, G. J., BONES, C. B., HACKBARTH, A. D., GOLDMANN, D. A. & SHAREK, P. J. 2010. Temporal Trends in Rates of Patient Harm Resulting from Medical Care. *New England Journal of Medicine*, 363, 2124-2134.
- LAPLACE, P. S. 1986. Memoir on the Probability of the Causes of Events. *Statist. Sci.*, 1, 364-378.
- LARSEN, K. 2018. GAM: The Predictive Modeling Silver Bullet | Stitch Fix Technology – Multithreaded. @stitchfix_algo.
- LECKO, C. 2010. Nutrition and patient safety a report from the National Patient Safety Agency (United Kingdom). *World Hosp Health Serv*, 46, 29-32.
- LECKO, C. & BEST, C. 2013. Hydration - the missing part of nutritional care. *Nurs Times*, 109, 12-14 3p.
- LEE, Y. & NELDER, J. A. 2004. Conditional and Marginal Models: Another View. *Statist. Sci.*, 19, 219-238.
- LIANG, K.-Y. & ZEGER, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- LIAW, A. & WIENER, M. 2002. Classification and Regression by randomForest. *R News*, 2, 18-22.
- LILFORD, R., MOHAMMED, M. A., SPIEGELHALTER, D. & THOMSON, R. 2004. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet*, 363, 1147-54.
- LILFORD, R. & PRONOVOST, P. 2010. Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. *BMJ*, 340, c2016.

- LILFORD, R. J., CHILTON, P. J., HEMMING, K., GIRLING, A. J., TAYLOR, C. A. & BARACH, P. 2010. Evaluating policy and service interventions: framework to guide selection and interpretation of study end points. *BMJ*, 341, c4413.
- LONG, P. M. & SERVEDIO, R. A. 2010. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78, 287-304.
- LUCAS, J. M. & CROSIER, R. B. 1982. Fast Initial Response for CUSUM Quality-Control Schemes: Give Your CUSUM A Head Start. *Technometrics*, 24, 199-205.
- MACFARLANE, A. J., GODDEN, S. & POLLOCK, A. M. 2005. Are we on track – can we monitor bed targets in the NHS plan for England? *Journal of Public Health*, 27, 263-269.
- MACLENNAN, A. I. & SMITH, A. F. 2011. An analysis of critical incidents relevant to pediatric anesthesia reported to the UK National Reporting and Learning System, 2006-2008. *Paediatr Anaesth*, 21, 841-7.
- MACRAE, C. 2016. The problem with incident reporting. *BMJ Quality & Safety*, 25, 71.
- MACRAE, C. & VINCENT, C. 2014. Learning from failure: the need for independent safety investigation in healthcare. *Journal of the Royal Society of Medicine*, 107, 439-443.
- MAGNUSSON, A., SKAUG, H. J., NIELSEN, A., BERG, C. W., KRISTENSEN, K., MAECHLER, M., VAN BENTHEM, K. J., BOLKER, B. M. & BROOKS, M. E. 2017. glmmTMB: Generalized Linear Mixed Models using Template Model Builder. Available: <https://cran.r-project.org/web/packages/glmmTMB/glmmTMB.pdf>.
- MAHAJAN, R., MATHEWS, L., RUSSELL, J. & GEMMELL, L. 2009. 'Wrong Drug' errors during anaesthesia as reported to the National Reporting and Learning System in the UK. *European Journal of Anaesthesiology*, 26, 205.
- MANNING, C. D. & SCHÜTZE, H. 1999. *Foundations of Statistical Natural Language Processing*, Cambridge, Massachusetts / London, England, MIT Press.
- MARRA, G. & WOOD, S. N. 2011. Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55, 2372-2387.
- MARSHALL, C., BEST, N., BOTTLE, A. & AYLIN, P. 2004. Statistical Issues in the Prospective Monitoring of Health Outcomes across Multiple Units. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 167, 541-559.
- MARTIN, G. P., MCKEE, L. & DIXON-WOODS, M. 2015. Beyond metrics? Utilizing 'soft intelligence' for healthcare quality and safety. *Social Science & Medicine*, 142, 19-26.
- MARTINEZ, E. A., SHORE, A., COLANTUONI, E., HERZER, K., THOMPSON, D. A., GURSES, A. P., MARSTELLER, J. A., BAUER, L., GOESCHEL, C. A., CLEARY, K., PRONOVOST, P. J. & PHAM, J. C. 2011. Cardiac surgery errors: results from the UK National Reporting and Learning System.[Erratum appears in Int J Qual Health Care. 2014 Aug;26(4):499]. *International Journal for Quality in Health Care*, 23, 151-8.
- MASTODON C 2015. NRLS: report on technical prototyping.
- MASTODON C. 2019. *patient safety text mining* [Online].
<https://www.mastodonc.com/casestudies/nhs/>: Mastodon C. Available:
<https://www.mastodonc.com/casestudies/nhs/> [Accessed 08/03/2019 2019].
- MAYER, E., FLOTT, K., CALLAHAN, R. & DARZI, A. 2017. National Reporting and Learning System Research and Development. London: Imperial College London.
- MCCULLAGH, P. & NELDER, J. A. 1983. *Generalized linear models*, London, Chapman & Hall.
- MCCULLAGH, P. & NELDER, J. A. 1989. *Generalized Linear Models, Second Edition*, Taylor & Francis.
- MCDONALD, C. J., WEINER, M. & HUI, S. L. 2000. Deaths due to medical errors are exaggerated in institute of medicine report. *JAMA*, 284, 93-95.
- MCGRATH, B. A. 2010. Patient safety incidents associated with tracheostomies occurring in hospital wards: a review of reports to the UK National Patient Safety Agency.
- MCGRATH, B. A., BATES, L., ATKINSON, D. & MOORE, J. A. 2013. Algorithm for management of tracheostomy emergencies on intensive care Reply. *Anaesthesia*, 68, 219-220.

- MCGRATH, B. A. & THOMAS, A. N. 2010. Patient safety incidents associated with tracheostomies occurring in hospital wards: a review of reports to the UK National Patient Safety Agency. *Postgrad Med J*, 86, 522-5.
- MCGRATH, B. A. & THOMAS, A. N. 2011. Patient safety incidents associated with tracheostomies: A comparison of levels of harm between critical care and ward environments. *British Journal of Anaesthesia*, 106 (3), 439.
- MEDICINES & PRESCRIBING TEAM, H. 2015. NICE Technology Appraisals in the NHS in England (Innovation Scorecard). In: (HSCIC), H. A. S. C. I. C. (ed.). HSCIC.
- MEDICINES AND HEALTHCARE REGULATORY AGENCY (MHRA). 2016. *Yellow Card Scheme - MHRA* [Online]. Medicines and Healthcare Regulatory Agency. Available: <https://yellowcard.mhra.gov.uk/the-yellow-card-scheme/> [Accessed 31/10/2016 2016].
- MEDICINES AND HELATHCARE REGULATORY AGENCY (MHRA) 2010. Serious Adverse Blood Reactions and Events (SABRE) In: MHRA (ed.). MHRA.
- MICROSOFT. 2018. *About Microsoft R Open: The Enhanced R Distribution*. MRAN [Online]. Available: <https://mran.microsoft.com/rro> [Accessed 13/08/2018 2018].
- MILLIGAN, F. 2012. Diabetes medication incidents in the care home setting. *Nurs Stand*, 26, 38-43.
- MILLIGAN, F. J., KRENTZ, A. J. & SINCLAIR, A. J. 2011. Diabetes medication patient safety incident reports to the National Reporting and Learning Service: the care home setting. *Diabet Med*, 28, 1537-40.
- MOHAMMED, M. A., CHENG, K. K., ROUSE, A. & MARSHALL, T. 2001. Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *The Lancet*, 357, 463-467.
- MOLENBERGHS, G. & VERBEKE, G. 2005. *Models for discrete longitudinal data*, New York ; London, New York ; London : Springer.
- MORTON, A., MENGENSEN, K. L., PLAYFORD, G. & WHITBY, M. 2013. Statistical Methods for Hospital Monitoring. *Wiley StatsRef: Statistics Reference Online*.
- MOUSTAKIDES, G. V. 1986. Optimal Stopping Times for Detecting Changes in Distributions. *The Annals of Statistics*, 14, 1379-1387.
- MUNN, Z., PETERS, M. D. J., STERN, C., TUFANARU, C., MCARTHUR, A. & AROMATARIS, E. 2018. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18, 143.
- MURZINTCEV, N. 2019. Idateuning: Tuning of the Latent Dirichlet Allocation Models Parameters. 1.0.0 ed. CRAN.
- NASA 2019. Aviation Safety Reporting System. ASRS program briefing. In: NASA (ed.).
- NATIONAL PATIENT SAFETY AGENCY 2004. Seven steps to patient safety: the full reference guide. London: National Patient Safety Agency.
- NELDER, J. A. & WEDDERBURN, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135, 370-384.
- NEUBURGER, J., WALKER, K., SHERLAW-JOHNSON, C., VAN DER MEULEN, J. & CROMWELL, D. A. 2017. Comparison of control charts for monitoring clinical performance using binary data. *BMJ Quality & Safety*, 26, 919-928.
- NHS DIGITAL. 2012. *NHS Safety Thermometer* [Online]. NHS Digital. Available: <http://content.digital.nhs.uk/thermometer> [Accessed 31/10/2016 2016].
- NHS DIGITAL. 2016. *Hospital Episode Statistics* [Online]. NHS Digital, 1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE, United Kingdom. Available: <http://content.digital.nhs.uk/hes> [Accessed 02/11/2016 2016].
- NHS DIGITAL. 2017a. *Clinical Classifications* [Online]. Available: <https://digital.nhs.uk/article/1117/Clinical-Classifications> [Accessed 10/10/2017 2017].
- NHS DIGITAL. 2017b. *Commissioning Data Sets (CDS)* [Online]. NHS Digital, 1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE, United Kingdom. Available: <http://content.digital.nhs.uk/commissioningdataset> [Accessed 10/10/2017 2017].

- NHS DIGITAL. 2017c. *Healthcare Resource Groups (HRG4)* [Online]. NHS Digital, 1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE, United Kingdom. Available: <http://content.digital.nhs.uk/hrg4> [Accessed 10/10/2017 2017].
- NHS DIGITAL. 2017d. *HES data dictionary* [Online]. NHS Digital. Available: <http://content.digital.nhs.uk/hesdatadictionary> [Accessed 14/11/2017 2017].
- NHS DIGITAL. 2017e. *Secondary Uses Service (SUS)* [Online]. NHS Digital, 1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE, United Kingdom; NHS Digital, 1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE, United Kingdom. Available: <http://content.digital.nhs.uk/sus> [Accessed 10/10/2017 2017].
- NHS DIGITAL. 2017f. *Summary Hospital-level Mortality Indicator* [Online]. Leeds: NHS Digital, 1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE, United Kingdom; NHS Digital, 1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE, United Kingdom. Available: <https://www.digital.nhs.uk/SHMI> [Accessed 13/11/2017 2017].
- NHS DIGITAL. 2018a. *Emergency Care Data Set (ECDS) - NHS Digital* [Online]. NHS Digital. Available: <https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/emergency-care-data-set-ecds> [Accessed 18/07/2018 2018].
- NHS DIGITAL. 2018b. *Understanding the health and care information we collect* [Online]. Available: <https://digital.nhs.uk/about-nhs-digital/our-work/keeping-patient-data-safe/how-we-look-after-your-health-and-care-information/understanding-the-health-and-care-information-we-collect> [Accessed 10/03/2018 2019].
- NHS ENGLAND 2015. Revised Never Events Policy and Framework. In: ENGLAND, N. (ed.). London: NHS England.
- NHS ENGLAND. 2017. *Bed Availability and Occupancy* [Online]. Available: <https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/> [Accessed 16/08/2017 2017].
- NHS ENGLAND. 2018. *NHS England » Winter resilience* [Online]. Available: <https://www.england.nhs.uk/winter/> [Accessed].
- NHS IMPROVEMENT. 2017a. *The future of the patient safety incident reporting: upgrading the NRLS* [Online]. NHS Improvement. Available: <https://improvement.nhs.uk/news-alerts/development-patient-safety-incident-management-system-dpsims/> [Accessed 11/03/2019 2019].
- NHS IMPROVEMENT. 2017b. *National patient safety incident reports* [Online]. NHS Improvement. Available: <https://improvement.nhs.uk/resources/national-quarterly-data-patient-safety-incident-reports/> [Accessed 10/03/2019 2019].
- NHS IMPROVEMENT. 2017c. *Organisation patient safety incident reports* [Online]. NHS Improvement. Available: <https://improvement.nhs.uk/resources/organisation-patient-safety-incident-reports-data/> [Accessed 10/03/2019 2019].
- NHS IMPROVEMENT. 2018. *Pilot project for the new national patient safety system* [Online]. Available: <https://improvement.nhs.uk/>: NHS Improvement. Available: <https://improvement.nhs.uk/resources/dpsims-project-pilot/> [Accessed 11/03/2019 2019].
- NOBLE, D. J. & PRONOVOST, P. J. 2010. Underreporting of Patient Safety Incidents Reduces Health Care's Ability to Quantify and Accurately Measure Harm Reduction. *Journal of Patient Safety*, 6, 247-250.
- NUCKOLS, T. K., BELL, D. S., LIU, H., PADDOCK, S. M. & HILBORNE, L. H. 2007. Rates and types of events reported to established incident reporting systems in two US hospitals. *Quality & safety in health care*, 16, 164-168.
- O'GRADY, I. & GERRETT, D. 2015. Minimising harm from missed drug doses. *Nurs Times*, 111, 12-5.
- OMAR, A., REES, P., EVANS, H. P., WILLIAMS, H., COOPER, A., BANERJEE, S., HIBBERT, P., MAKEHAM, M., PARRY, G., DONALDSON, L., EDWARDS, A. & CARSON-STEVENSON, A. 2015. Vulnerable Children and Their Care Quality Issues: A Descriptive Analysis of a National Database. *BMJ Quality & Safety*, 24, 732-733.

- PAN, J. & THOMPSON, R. 2003. Gauss-Hermite Quadrature Approximation for Estimation in Generalised Linear Mixed Models. *Computational Statistics*, 18, 57-78.
- PANESAR, S. S., CARSON-STEVENSON, A., MANN, B. S., BHANDARI, M. & MADHOK, R. 2012a. Mortality as an indicator of patient safety in orthopaedics: lessons from qualitative analysis of a database of medical errors. *BMC Musculoskeletal Disord*, 13, 93.
- PANESAR, S. S., CARSON-STEVENSON, A., SALVILLA, S. A., PATEL, B., MIRZA, S. B. & MANN, B. 2013a. Patient safety in orthopedic surgery: prioritizing key areas of iatrogenic harm through an analysis of 48,095 incidents reported to a national database of errors. *Drug Healthc Patient Saf*, 5, 57-65.
- PANESAR, S. S., CLEARY, K. & SHEIKH, A. 2009. Reflections on the National Patient Safety Agency's database of medical errors. *Journal of the Royal Society of Medicine*, 102, 256-258.
- PANESAR, S. S., IGNATOWICZ, A. M. & DONALDSON, L. J. 2014. Errors in the management of cardiac arrests: an observational study of patient safety incidents in England. *Resuscitation*, 85, 1759-63.
- PANESAR, S. S., NETUVELI, G., CARSON-STEVENSON, A., JAVAD, S., PATEL, B., PARRY, G., DONALDSON, L. J. & SHEIKH, A. 2013b. The orthopaedic error index: development and application of a novel national indicator for assessing the relative safety of hospital care using a cross-sectional approach. *BMJ Open*, 3, e003448.
- PANESAR, S. S., NOBLE, D. J., MIRZA, S. B., PATEL, B., MANN, B., EMERTON, M., CLEARY, K., SHEIKH, A. & BHANDARI, M. 2011. Can the surgical checklist reduce the risk of wrong site surgery in orthopaedics?--Can the checklist help? Supporting evidence from analysis of a national patient incident reporting system. *J Orthop Surg Res*, 6, 18.
- PANESAR, S. S., SIMUNOVIC, N. & BHANDARI, M. 2012b. When should we operate on elderly patients with a hip fracture? It's about time! *Surgeon*, 10, 185-8.
- PARLIAMENT, H. O. C. 2012. Health and Social Care Act 2012. In: PARLIAMNET, H. O. C. (ed.). London: The stationary Office.
- PEDERSEN, K. Z. 2016. Standardisation or resilience? The paradox of stability and change in patient safety. *Sociology of Health & Illness*, 38, 1180-1193.
- PHAM, J. C., COLANTUONI, E., DOMINICI, F., SHORE, A., MACRAE, C., SCOBIE, S., FLETCHER, M., CLEARY, K., GOESCHEL, C. A. & PRONOVOST, P. J. 2010. The harm susceptibility model: a method to prioritise risks identified in patient safety reporting systems. *Qual Saf Health Care*, 19, 440-5.
- PHAM, J. C., GIRARD, T. & PRONOVOST, P. J. 2013. What to do with healthcare incident reporting systems. *Journal of public health research*, 2, e27-e27.
- PINHEIRO, J. & BATES, D. 2000. *Mixed-Effects Models in S and S-PLUS*, Springer New York.
- PINTO, A., FAIZ, O. & VINCENT, C. 2012. Managing the after effects of serious patient safety incidents in the NHS: an online survey study. *BMJ Quality & Safety*, 21, 1001-1008.
- POISSON, S. D. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*, Bachelier.
- POLLACK, E. 2018. *Dynamic SQL: Applications, Performance, and Security in Microsoft SQL Server*, Apress.
- PORTER, M. F. 1980. *An Algorithm for Suffix Stripping*.
- POWELL, M. J. D. 2009. *The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives*. Cambridge University.
- PRONOVOST, P. J., THOMPSON, D. A., HOLZMUELLER, C. G., LUBOMSKI, L. H., DORMAN, T., DICKMAN, F., FAHEY, M., STEINWACHS, D. M., ENGINEER, L., SEXTON, J. B., WU, A. W. & MORLOCK, L. L. 2006. Toward learning from patient safety reporting systems. *Journal of Critical Care*, 21, 305-315.
- QUILLIVAN, R. R., BURLISON, J. D., BROWNE, E. K., SCOTT, S. D. & HOFFMAN, J. M. 2016. Patient Safety Culture and the Second Victim Phenomenon: Connecting Culture to Staff Distress in Nurses. *Jt Comm J Qual Patient Saf*, 42, 377-86.

- R CORE TEAM 2016. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- RABE-HESKETH, S. & SKRONDAL, A. 2012. Multilevel and Longitudinal Modeling Using Stata, Volumes I and II, Third Edition. 3rd ed.: Taylor & Francis.
- RADHAKRISHNA, S. 2015. Culture of blame in the National Health Service; consequences and solutions. *BJA: British Journal of Anaesthesia*, 115, 653-655.
- REASON, J. 1990. *Human Error*, Cambridge University Press.
- REES, P., CARSON-STEVENSON, A., WILLIAMS, H., PANESAR, S. & EDWARDS, A. 2014. Quality improvement informed by a reporting and learning system. *Arch Dis Child*, 99, 702-3.
- REES, P., EDWARDS, A., PANESAR, S., POWELL, C., CARTER, B., WILLIAMS, H., HIBBERT, P., LUFF, D., PARRY, G., MAYOR, S., AVERY, A., SHEIKH, A., DONALDSON, S. L. & CARSON-STEVENSON, A. 2015a. Safety incidents in the primary care office setting. *Pediatrics*, 135, 1027-35.
- REES, P., EDWARDS, A., POWELL, C., EVANS, H. P., CARTER, B., HIBBERT, P., MAKEHAM, M., SHEIKH, A., DONALDSON, L. & CARSON-STEVENSON, A. 2015b. Pediatric immunization-related safety incidents in primary care: A mixed methods analysis of a national database. *Vaccine*, 33, 3873-80.
- REES, P., EDWARDS, A., POWELL, C., WILLIAMS, H., HIBBERT, P., MAKEHAM, M., LUFF, D., PARRY, G., SHEIKH, A., DONALDSON, L. & CARSON-STEVENSON, A. 2015c. Identifying Priorities for Improved Child Healthcare: A Mixed Methods Analysis of Safety Incident Reports. *BMJ Quality & Safety*, 24, 730-731.
- REINSCH, C. H. 1967. Smoothing by spline functions. *Numerische mathematik*, 10, 177-183.
- RENNIE, J. D. M., SHIH, L., TEEVAN, J. & KARGER, D. R. 2003. Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. Washington, DC, USA: AAAI Press.
- REVOLUTION ANALYTICS & WESTON, S. 2015. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. . 1.0.10 ed. CRAN.
- RIBEIRO, M. T., SINGH, S. & GUESTIN, C. Why should i trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016. ACM, 1135-1144.
- RIPLEY, B. D. 2004. Selecting Amongst Large Classes of Models. Website: University of Oxford.
- ROBERTSON, J. A. & SMITH, A. F. 2010. Anaesthetic airway incident reporting in the National Reporting and Learning System. *Anaesthesia*, 65, 429-430.
- ROBINSON, D. 2017. broom: Convert Statistical Analysis Objects into Tidy Data Frames. 0.4.2 ed. CRAN.
- ROBINSON, P. M. & MUIR, L. T. 2009. Wrong-site surgery in orthopaedics. *J Bone Joint Surg Br*, 91, 1274-80.
- ROCOS, B. & DONALDSON, L. J. 2012. Alcohol skin preparation causes surgical fires. *Ann R Coll Surg Engl*, 94, 87-9.
- ROYAL COLLEGE OF OBSTETRICIANS AND GYNAECOLOGISTS 2012. Hospital Episode Statistics as a source of information on safety and quality in gynaecology to support revalidation. London, UK: Royal College of Obstetricians and Gynaecologists.
- RUCH, P., GOBEILL, J., LOVIS, C. & GEISSBÜHLER, A. 2008. Automatic medical encoding with SNOMED categories. *BMC medical informatics and decision making*, 8 Suppl 1, S6-S6.
- RUMELHART, D. E., HINTON, G. E. & WILLIAMS, R. J. 1986. Learning representations by back-propagating errors. *Nature*, 323, 533.
- RUNCIMAN, W. B., BAKER, G. R., MICHEL, P., DOVEY, S., LILFORD, R. J., JENSEN, N., FLIN, R., WEEKS, W. B., LEWALLE, P., LARIZGOITIA, I. & BATES, D. 2010. Tracing the foundations of a conceptual framework for a patient safety ontology. *Quality and Safety in Health Care*, 19, e56-e56.
- RUNCIMAN, W. B. & MOLLER, J. 2001. *Iatrogenic injury in Australia : a report*, Australian Patient Safety Foundation.

- RUTTER, P. D., PANESAR, S. S., DARZI, A. & DONALDSON, L. J. 2014. What is the risk of death or severe harm due to bone cement implantation syndrome among patients undergoing hip hemiarthroplasty for fractured neck of femur? A patient safety surveillance study. *BMJ Open*, 4, e004853.
- RYLANCE, P., FIELDING, C., HUTCHISON, A. & LIPKIN, G. 2015. Making Care of Haemodialysis Patients Safer: Outcomes of the Uk Renal Association Patient Safety Project, 2007-2015. *Nephrology Dialysis Transplantation*, 30, iii304-iii305.
- SAIF, H., FERNÁNDEZ, M., HE, Y. & ALANI, H. 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter.
- SALTON, G. & BUCKLEY, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24, 513-523.
- SALTON, G. & MCGILL, M. J. 1986. *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc.
- SAMURIWO, R., WILLIAMS, H., COOPER, J. & CARSON-STEVENSON, A. 2016. Improving skin care through data: a pitch for patient safety incident reporting. *Journal of Wound Care*, 25, 691-691.
- SANDALL, J., WATSON, K. & WISEMAN, O. 2012. Analysis of 772 reported 'death' and 'severe' obstetric and neonatal incidents from the United Kingdom NPSA national reporting and learning system (NRLS). *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 97, A73.1-A73.
- SANTELL, J. P., HICKS, R. W., MCMEEKIN, J. & COUSINS, D. D. 2003. Medication Errors: Experience of the United States Pharmacopeia (USP) MEDMARX Reporting System. *The Journal of Clinical Pharmacology*, 43, 760-767.
- SARI, A. B., SHELDON, T. A., CRACKNELL, A. & TURNBULL, A. 2007. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: retrospective patient case note review. *BMJ*, 334, 79.
- SAS FOUNDATION. 2018. *PROC GLIMMIX: PROC GLIMMIX Statement :: SAS/STAT(R) 9.22 User's Guide* [Online]. Available: https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_glimmix_a0000001411.htm#statug.glimmix.gmxoptmethodlaplace [Accessed 29/05/2018 2018].
- SCHAPIRE, R. E. 1990. The strength of weak learnability. *Machine Learning*, 5, 197-227.
- SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. & LEE, W. S. 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, 26, 1651-1686.
- SCHIELZETH, H. 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1, 103-113.
- SCHULZ, K. F., ALTMAN, D. G. & MOHER, D. 2010. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, c332.
- SCHWARZ, G. 1978. Estimating the Dimension of a Model. *Ann. Statist.*, 6, 461-464.
- SCOTT-WARREN, J., MCPHERSON, D., MAHAJAN, R. & SMITH, A. 2012. Anaesthesia incident reporting in the UK: Comparison of a generic national and a specialty-specific reporting system. *European Journal of Anaesthesiology*, 29, 219.
- SELLERS, K. F. & SHMUELI, G. 2010a. A flexible regression model for count data. *Ann. Appl. Stat.*, 4, 943-961.
- SELLERS, K. F. & SHMUELI, G. 2010b. Predicting Censored Count Data with COM-Poisson Regression. *Robert H. Smith School Research Paper*, No. RHS-06-129.
- SERIOUS HAZARDS OF TRANSFUSION. 2016. *Home - Serious Hazards of Transfusion* [Online]. Available: <http://www.shotuk.org/> [Accessed 31/10/2016 2016].
- SEVDALIS, N., JACKLIN, R., ARORA, S., VINCENT, C. A. & THOMSON, R. G. 2010. Diagnostic error in a national incident reporting system in the UK. *J Eval Clin Pract*, 16, 1276-81.
- SHAW, R., DREVER, F., HUGHES, H., OSBORN, S. & WILLIAMS, S. 2005. Adverse events and near miss reporting in the NHS. *Qual Saf Health Care*, 14, 279-83.

- SHEIKH, A. & HURWITZ, B. 1999. A national database of medical error. *Journal of the Royal Society of Medicine*, 92, 554-555.
- SHMUELI, G. 2010. To Explain or to Predict? *Statist. Sci.*, 25, 289-310.
- SHMUELI, G., MINKA, T. P., KADANE, J. B., BORLE, S. & BOATWRIGHT, P. 2005. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 127-142.
- SHOJANIA, K. G. 2008. The frustrating case of incident-reporting systems. *Quality and Safety in Health Care*, 17, 400-402.
- SHOJANIA, K. G. 2012. Deaths due to medical error: jumbo jets or just small propeller planes? *BMJ Quality & Safety*, 21, 709-712.
- SHOJANIA, K. G., DUNCAN, B. W., MCDONALD, K. M. & WACHTER, R. M. 2002. Safe but Sound Patient Safety Meets Evidence-Based Medicine. *JAMA*, 288, 508-513.
- SHOJANIA, K. G. & THOMAS, E. J. 2013. Trends in adverse events over time: why are we not improving? *BMJ Quality & Safety*, 22, 273-277.
- SHUN, Z. & MCCULLAGH, P. 1995. Laplace Approximation of High Dimensional Integrals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 749-760.
- SILGE, J. & ROBINSON, D. 2016. tidytext: Text Mining and Analysis Using Tidy Data Principles in R.
- SILGE, J. & ROBINSON, D. 2017. *Text Mining with R*, USA, O'Reilly.
- SPÄRCK JONES, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 11-21.
- SPIEGELHALTER, D., SHERLAW-JOHNSON, C., BARDSLEY, M., BLUNT, I., WOOD, C. & GRIGG, O. 2012a. Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175, 1-47.
- SPIEGELHALTER, D., SHERLAW-JOHNSON, C., BARDSLEY, M., BLUNT, I., WOOD, C. & GRIGG, O. 2012b. Statistical methods for healthcare regulation: rating, screening and surveillance [with discussion]. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 175, 1-47.
- SPIEGELHALTER, D., SHERLAW-JOHNSON, C., BARDSLEY, M., BLUNT, I., WOOD, C. & GRIGG, O. 2012c. Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175, 1-47.
- SPIEGELHALTER, D. J. 2005a. Funnel plots for comparing institutional performance. *Stat Med*, 24, 1185-202.
- SPIEGELHALTER, D. J. 2005b. Handling over-dispersion of performance indicators. *Quality and Safety in Health Care*, 14, 347-351.
- SPIEGELHALTER, D. J., AYLIN, P., BEST, N. G., EVANS, S. J. W. & MURRAY, G. D. 2002. Commissioned analysis of surgical performance using routine data: lessons from the Bristol inquiry. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165, 191-221.
- SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. & SALAKHUTDINOV, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15, 1929-1958.
- STIEGLITZ, S., MIRBABAIE, M., ROSS, B. & NEUBERGER, C. 2018. *Social media analytics – Challenges in topic discovery, data collection, and data preparation*.
- STUBBS, J., HAW, C. & DICKENS, G. 2008. Dose form modification - a common but potentially hazardous practice. A literature review and study of medication administration to older psychiatric inpatients. *Int Psychogeriatr*, 20, 616-27.
- STUTTAFORD, L., CHAKRABORTY, M., CARSON-STEVENSON, A. & POWELL, C. 2018. G190 Patient safety incidents in neonatology: a 10-year descriptive analysis of reports from nhs england and wales. *Archives of Disease in Childhood*, 103, A78-A78.

- SUN, G. W., SHOOK, T. L. & KAY, G. L. 1996. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol*, 49, 907-16.
- TALBOT, D., DUCHESNE, T., BRISSON, J. & VANDAL, N. 2011. Variance estimation and confidence intervals for the standardized mortality ratio with application to the assessment of a cancer screening program. *Statistics in Medicine*, 30, 3024-3037.
- TEMPLETON, R., WEBSTER, K. & MCGRATH, B. A. 2011. Patient safety incidents associated with displaced or obstructed tracheostomies: comparison of levels of harm between critical care and ward environments. *British Journal of Anaesthesia*, 107, 826-838.
- THE CHIRP CHARITABLE TRUST. 2020. *Aviation and Maritime Confidential Incident Reporting* [Online]. Available: <https://www.chirp.co.uk/> [Accessed 13/03/2020 2020].
- THE HEALTH SOCIAL CARE INFORMATION CENTRE, H. 2015. *Hospital Episode Statistics for England. Admitted Patient Care statistics, 2014-15* [Online]. The Health and Social Care Information Centre (HSCIC). Available: <http://content.digital.nhs.uk/searchcatalogue?productid=19420&q=title%3a%22Hospital+Episode+Statistics%2c+Admitted+patient+care+-+England%22&sort=Relevance&size=10&page=1#top> [Accessed 11/11/2016 2016].
- THERNEAU, T. & ATKINSON, B. 2018. rpart: Recursive Partitioning and Regression Trees. In: FOUNDATION, M. (ed.). CRAN.
- THOMAS, A. N. & GALVIN, I. 2008. Patient safety incidents associated with equipment in critical care: a review of reports to the UK National Patient Safety Agency. *Anaesthesia*, 63, 1193-7.
- THOMAS, A. N. & MCGRATH, B. A. 2009. Patient safety incidents associated with airway devices in critical care: a review of reports to the UK National Patient Safety Agency. *Anaesthesia*, 64, 358-65.
- THOMAS, A. N. & PANCHAGNULA, U. 2008. Medication-related patient safety incidents in critical care: a review of reports to the UK National Patient Safety Agency. *Anaesthesia*, 63, 726-33.
- THOMAS, A. N., PANCHAGNULA, U. & TAYLOR, R. J. 2009. Review of patient safety incidents submitted from Critical Care Units in England & Wales to the UK National Patient Safety Agency. *Anaesthesia*, 64, 1178-85.
- THOMAS, E. J., LIPSITZ, S. R., STUDDERT, D. M. & BRENNAN, T. A. 2002. The reliability of medical record review for estimating adverse event rates. *Ann Intern Med*, 136, 812-6.
- THUSU, S., PANESAR, S. & BEDI, R. 2012. Patient safety in dentistry - state of play as revealed by a national database of errors. *Br Dent J*, 213, E3.
- TIAN, W., SUN, H., ZHANG, X. & WOODALL, W. H. 2015. The impact of varying patient populations on the in-control performance of the risk-adjusted CUSUM chart. *Int J Qual Health Care*, 27, 31-6.
- TIBSHIRANI, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267-288.
- TRUICA, C., RADULESCU, F. & BOICEA, A. Comparing Different Term Weighting Schemas for Topic Modeling. 2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 24-27 Sept. 2016 2016. 307-310.
- ULM, K. 1990. Simple Method to Calculate the Confidence Interval of a Standardized Mortality Ratio (SMR). *American Journal of Epidemiology*, 131, 373-375.
- UNIVERSITY HOSPITALS BIRMINGHAM NHS FOUNDATION TRUST. 2019. *Healthcare Evaluation Data (HED)* [Online]. Birmingham, UK. Available: www.hed.nhs.uk [Accessed 08/03/2019 2019].
- VAN DEN BROEK, J. 1995. A score test for zero inflation in a Poisson distribution. *Biometrics*, 51, 738-43.
- VENABLES, W. N. & RIPLEY, B. D. 2002. *Modern Applied Statistics with S*, New York, Springer.
- VENABLES, W. N. & RIPLEY, B. D. 2013. *Modern Applied Statistics with S-Plus*, Springer New York.

- VER HOEF, J. M. & BOVENG, P. L. 2007. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88, 2766-72.
- VEZHNEVETS, A. & BARINOVA, O. Avoiding Boosting Overfitting by Removing Confusing Samples. In: KOK, J. N., KORONACKI, J., MANTARAS, R. L. D., MATWIN, S., MLADENIČ, D. & SKOWRON, A., eds. Machine Learning: ECML 2007, 2007// 2007 Berlin, Heidelberg. Springer Berlin Heidelberg, 430-441.
- VINCENT, C. 2007. Incident reporting and patient safety. *BMJ*, 334, 51-51.
- VINCENT, C. & AMALBERTI, R. 2015. Safety in healthcare is a moving target. *BMJ Quality & Safety*, 24, 539-540.
- VINCENT, C., AYLIN, P., FRANKLIN, B. D., HOLMES, A., ISKANDER, S., JACKLIN, A. & MOORTHY, K. 2008. Is health care getting safer? *BMJ*, 337, a2426.
- VINCENT, C., NEALE, G. & WOLOSHYNOWYCH, M. 2001. Adverse events in British hospitals: preliminary retrospective record review. *BMJ*, 322, 517-519.
- VINCENT, C. A. 2004. Analysis of clinical incidents: a window on the system not a search for root causes. *Quality and Safety in Health Care*, 13, 242-243.
- WAHR, J. A., SHORE, A. D., HARRIS, L. H., ROGERS, P., PANESAR, S., MATTHEW, L., PRONOVOST, P. J., CLEARY, K. & PHAM, J. C. 2014. Comparison of intensive care unit medication errors reported to the United States' MedMarx and the United Kingdom's National Reporting and Learning System: a cross-sectional study. *Am J Med Qual*, 29, 61-9.
- WALD, A. 1942. Asymptotically Shortest Confidence Intervals. *Ann. Math. Statist.*, 13, 127-137.
- WEDDERBURN, R. W. M. 1974. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, 61, 439-447.
- WHEWAY, J., AGBABIAKA, T. B. & ERNST, E. 2012. Patient safety incidents from acupuncture treatments: a review of reports to the National Patient Safety Agency. *Int J Risk Saf Med*, 24, 163-9.
- WHITE, H. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48, 817-838.
- WHITE, H. 1982. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50, 1-25.
- WICKHAM, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*, New York, Springer-Verlag.
- WICKHAM, H. 2014. Tidy Data. *Journal of Statistical Software; Vol 1, Issue 10 (2014)*.
- WICKHAM, H. 2015. *R Packages*, O'Reilly Media.
- WICKHAM, H., FRANCOIS, R., HENRY, L. & MÜLLER, K. 2017. dplyr: A Grammar of Data Manipulation. 0.7.2 ed. CRAN.
- WICKHAM, H. & HENRY, L. 2017. tidy: Easily Tidy Data with 'spread()' and 'gather()' Functions. 0.7.0 ed.
- WICKHAM, H., HESTER, J. & CHANG, W. 2018. devtools: Tools to Make Developing R Packages. 2.0.1 ed.
- WILBUR, W. J. & SIROTKIN, K. 1992. The automatic identification of stop words. *Journal of Information Science*, 18, 45-55.
- WILKS, S. S. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Statist.*, 9, 60-62.
- WILLIAMS, H., COOPER, A. & CARSON-STEVENSON, A. 2016. Opportunities for incident reporting. Response to: 'The problem with incident reporting' by Macrae et al. *BMJ Quality & Safety*, 25, 133-134.
- WILLIAMS, H., EDWARDS, A., HIBBERT, P., REES, P., PROSSER EVANS, H., PANESAR, S., CARTER, B., PARRY, G., MAKEHAM, M., JONES, A., AVERY, A., SHEIKH, A., DONALDSON, L. & CARSON-STEVENSON, A. 2015. Harms from discharge to primary care: mixed methods analysis of incident reports. *Br J Gen Pract*, 65, e829-37.
- WILSON, A. C., ROELOFS, R., STERN, M., SREBRO, N. & RECHT, B. The marginal value of adaptive gradient methods in machine learning. *Advances in Neural Information Processing Systems*, 2017. 4148-4158.

- WILSON, A. T. & CHEW, P. A. 2010. Term weighting schemes for Latent Dirichlet Allocation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics.
- WILSON, D. R. & MARTINEZ, T. R. 2003. The general inefficiency of batch training for gradient descent learning. *Neural Netw.*, 16, 1429-1451.
- WILSON, R. M., RUNCIMAN, W. B., GIBBERD, R. W., HARRISON, B. T., NEWBY, L. & HAMILTON, J. D. 1995. The Quality in Australian Health Care Study. *Medical Journal of Australia*, 163, 458-471.
- WITTES, J., CROWE, B., CHUANG-STEIN, C., GUETTNER, A., HALL, D., JIANG, Q., ODENHEIMER, D., XIA, H. A. & KRAMER, J. 2015. The FDA's Final Rule on Expedited Safety Reporting: Statistical Considerations. *Statistics in biopharmaceutical research*, 7, 174-190.
- WOOD, S. N. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 95-114.
- WOOD, S. N. 2010. *More advanced use of mgcv* [Online]. Available: <https://people.maths.bris.ac.uk/~sw15190/mgcv/tampere/mgcv-advanced.pdf> [Accessed].
- WOOD, S. N. 2015. *Core Statistics*, Cambridge, Cambridge University Press.
- WOOD, S. N. 2017a. GAM Theory. *Generalized Additive Models: An Introduction with R, Second Edition*. 2nd ed. Florida, USA: CRC Press.
- WOOD, S. N. 2017b. *Generalized Additive Models: An Introduction with R, Second Edition*, Florida, USA, CRC Press.
- WOOD, S. N. 2017c. Introducing GAMs. *Generalized Additive Models: An Introduction with R, Second Edition*. 2nd ed. Florida, USA: CRC Press.
- WOOD, S. N. 2017d. Smoothers. *Generalized Additive Models: An Introduction with R, Second Edition*. 2nd ed. Florida, USA: CRC Press.
- WOODALL, W. H. & MAHMOUD, M. A. 2005. The Inertial Properties of Quality Control Charts. *Technometrics*, 47, 425-436.
- WORLD HEALTH ORGANIZATION. 2017. *International Classification of Diseases* [Online]. World Health Organization. Available: <http://www.who.int/classifications/icd/en/> [Accessed 10/10/2017 2017].
- WORTH, A., PANESAR, S., HEALY, L. & SHEIKH, A. 2012. Improving anaphylaxis prevention and management in healthcare settings: lessons from analysing patient safety incident reports. *Allergy*, 67, 107-108.
- YADAV, N., YADAV, A. & KUMAR, M. 2015. *An Introduction to Neural Network Methods for Differential Equations*, Springer Netherlands.
- ZHANG, Z. & HAND, D. 2005. *Detecting Groups of Anomalously Similar Objects in Large Data Sets*, Madrid, Spain.

Appendices

Appendix A: Full Literature review search strategy

Medline & Embase

Both databases separately searched using the Ovid interface, using the same search terms with the exception of search 9.

All text string searches used the 'mp' setting: title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier. Final searches conducted on 23/03/2016.

1. NRLS
2. "reporting and learning system"
3. NPSA
4. "patient safety agency"
5. report*
6. 3 or 4
7. 5 and 6
8. ((incident? adj3 report*) and (database or NHS)).
9. exp Great Britain/ (Medline) exp United Kingdom/ (Embase)
10. 8 and 9
11. inciden*
12. 6 and 11
13. 1 or 2 or 7 or 12
14. limit 13 to yr="2001 -Current"
15. 13 not 14

Health Management Information Consortium (HMIC)

HMIC was queried using the Ovid search engine. Index terms were used in this search for geography and for NPSA publications, as a specific term for the National Patient Safety Agency (NPSA) exists. Text searches conducted using the 'mp' setting: title, other title, abstract, heading words. Final search conducted on 31/03/2016.

1. (NRLS or "reporting and learning system")
2. (NPSA or "Patient safety agency")
3. report*
4. inciden*
5. 2 and 3
6. 2 and 4
7. ((inciden* adj3 report*) and (NHS or database))
8. exp great britain/ or exp united kingdom/
9. exp National Patient Safety Agency/
10. 7 and 8
11. 1 or 5 or 6 or 9 or 10

CINAL Plus

CINAL Plus database was searched using the EBSCO host interface. Index terms were not used, except for geographical location (search 8).

All fields were searched. Final search conducted on 31/03/2016.

1. "NRLS" or "reporting and learning system"

2. NPSA or "patient safety agency"
3. inciden*
4. report*
5. 2 AND 4
6. 2 AND 3
7. (inciden* N3 report*) and (NHS or database)
8. (MH "United Kingdom+") OR (MH "Great Britain+")
9. 7 AND 8
10. 1 OR 5 OR 6 OR 9
11. 1 OR 5 OR 6 OR 9 - Limiters - Publication Year: 2001-2016

Cochrane library

Cochrane library was queried using the Cochrane collaboration website, using 'Advanced Search' tool. All text searches using the 'ti, ab, kw' setting: Title, abstract and keywords. The Cochrane library automatically searches for word variations. Final search conducted on 25/03/2016. No results were returned.

1. NRLS
2. "reporting and learning system"
3. NPSA
4. "patient safety agency"
5. inciden?
6. report*
7. 3 or 4
8. 7 and 5
9. 7 and 6
10. 1 or 2 or 8 or 9

National Institute for Health Research (NIHR): Centre for Reviews and Dissemination (CRD), Database of Abstracts of Reviews of Effects (DARE) & NHS Economic Evaluations (NHS EED)

These database, hosted by University of York, were access via their website. Final searches conducted on 10/04/2015.

1. NRLS
2. "reporting and learning system"
3. NPSA
4. "Patient safety agency"
5. 1 OR 2 OR 3 OR 4

The British Medical Journal (BMJ) Quality and Safety

At the time searches were conducted, this database was only partially indexed within MEDLINE, necessitating a separates search. The journal website was queries directly using the advanced search option using the 'Text | Abstract | Title' search box. Final search conducted 4/04/2016.

1. (("NRLS" OR "reporting and learning system") OR (("NPSA" OR "patient safety agency") and (inciden* OR report*))) - From: Jan 2001.

Web of Science

The web of science search engine was directly queried via its website using the 'Advanced Search' page. All test searches conducted using the TS setting: Titles, Abstracts, Keywords and Indexing fields such as Systematics, Taxonomic Terms and Descriptors. Additional subject terms were added due to web of science's interdisciplinary nature. Final search conducted 01/04/2016.

1. "Patient Safety" OR "Medical Errors" OR "Safety Management" OR "Risk Management" OR "Quality Assurance, Health Care" OR "Medication Errors" - Timespan=2001-2016
2. "NRLS" OR "Reporting and learning system" - Timespan=2001-2016
3. (("patient safety agency" OR "NPSA") and "report*") NOT TS=("prostate specific antigen") - Timespan=2001-2016
4. (("patient safety agency" OR "NPSA") and "report*") NOT TS=("prostate specific antigen") - Timespan=2001-2016
5. 3 OR 4 OR 2 - Timespan=2001-2016
6. 5 AND 1 - Timespan=2001-2016

ProQuest Dissertations and Theses

Searches were conducted using the 'Advanced search' option on the ProQuest search engine using the 'All- anywhere except full text' option. Final search conducted on 01/04/2016.

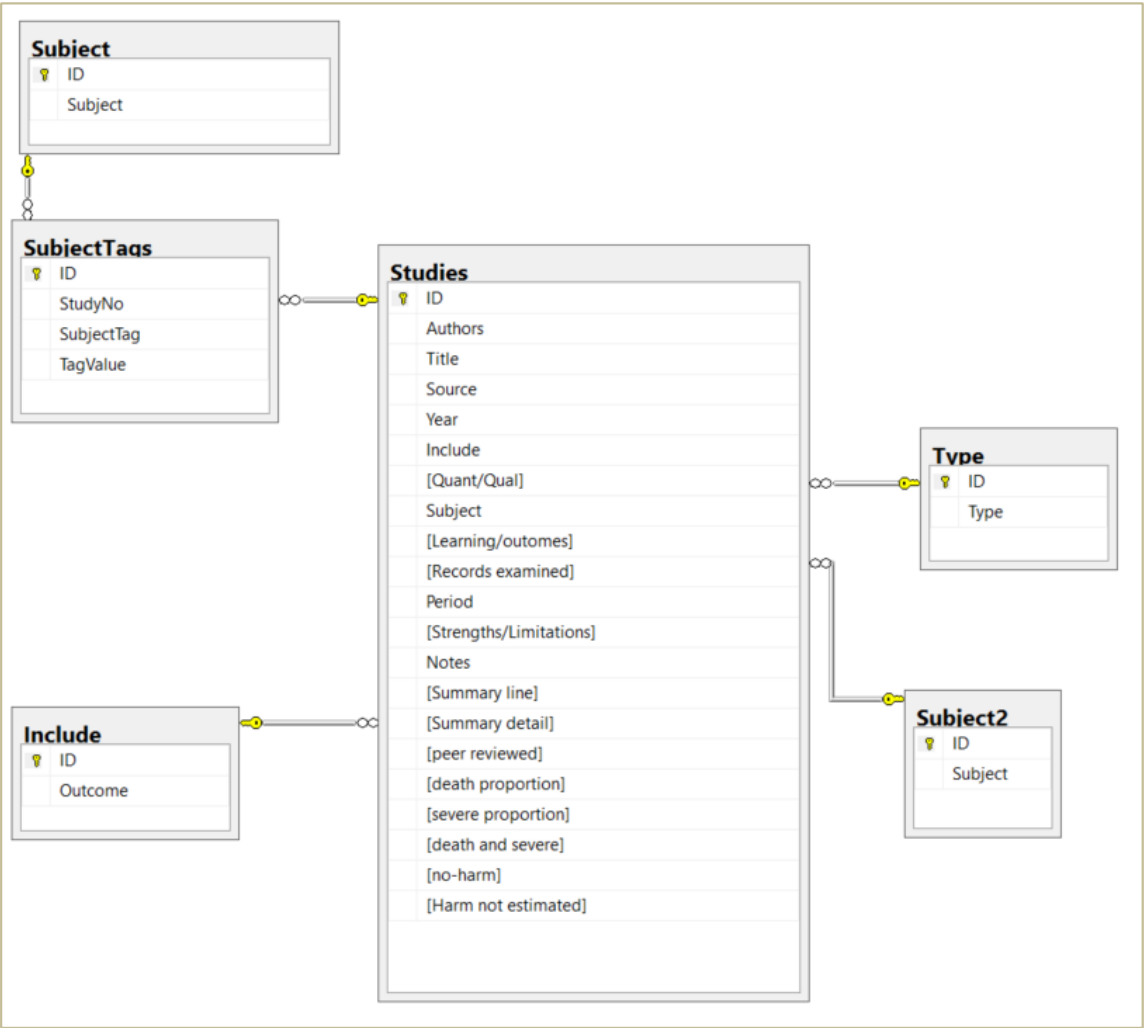
1. "NRLS" OR "reporting and learning system" OR "NPSA" OR "patient safety agency"

ProQuest Environmental Sciences and Pollution Management (Risk Abstracts)

The Risk abstracts database includes articles on risk management, but has been combined with the ProQuest Environmental Sciences and Pollution Management database. Searches were conducted using the 'Advanced search' option on the ProQuest search engine using the 'All- anywhere except full text' option. Final search conducted on 10/03/2016.

1. "NRLS" or "reporting and learning system" or (("NPSA" or "patient safety agency") and "report*")

Appendix B: Screen captures



B.1 Entity Diagram of system designed for literature review

Database structure was created in MS SQL Server 2014, with an entity diagram extracted using database diagramming tools. Database structure was designed for multiple tagging of articles with subjects and biases database diagramming tools with one-to-many relationships represented by key and infinity symbols.

dbo_Studies

Authors: Cousins DH, Gerrett D, Warner B

Outcome: **Include** (dropdown)

Year: 2012

Title: A review of medication incidents reported to the National Reporting and Learning System in England and Wales over 6 years (2005-2010)

Source: Br J Clin Pharmacol

Main subject: Medication (dropdown)

Learning/outcomes: Represented 9.68% of incidents, 75% from acute and only 8.5% in primary care. 0.9% resulted in death or severe harm. Largest process steps were Medication administration (50%) and prescribing (18%, with omitted/delayed (16%) and wrong dose (15%) representing largest categories. Strong increases in reporting each year demonstrated. Patient accidents 32.83 %, Treatment/procedure 9.59% access, admissions, transfer, discharge, missing patients 7.99. Identified drug classes.

Records examined: 526,186 of 5,437,99

Period: Jan 05 - Dec-10

Strengths/Limitations: Extracted from NRLS, using a medication specific-map (what is this?). Clinical outcome codes of severe or death potentially altered by author to be more accurate. Excluded if adverse drug reaction that was considered unavoidable. Small decrease in numbers observed on a second extract due to resolved misclassification/de-duplication

Subject tags:

Subject	Select
A&E	<input type="checkbox"/>
Abscondment	<input type="checkbox"/>
Access admission, transfer and discharge	<input type="checkbox"/>
Acupuncture	<input type="checkbox"/>
Administration (of drugs or treatment)	<input checked="" type="checkbox"/>
Administration (clerical)	<input type="checkbox"/>
Airway	<input type="checkbox"/>
Alternative Therapies	<input type="checkbox"/>
Ambulance	<input type="checkbox"/>
Anaesthesia	<input type="checkbox"/>
Anaphylaxis	<input type="checkbox"/>

Description:

Description	Select
Allocation of specialty	<input type="checkbox"/>
Alternative reporting route	<input checked="" type="checkbox"/>
Anonymisation	<input type="checkbox"/>
Ascertainment of reports	<input type="checkbox"/>
Classification of harm	<input checked="" type="checkbox"/>
Duplication	<input checked="" type="checkbox"/>
Lack of detail	<input checked="" type="checkbox"/>
Missing data	<input checked="" type="checkbox"/>
Poor/lack of search terms	<input type="checkbox"/>
Potential vs. actual harm	<input checked="" type="checkbox"/>
Re-classification by authors	<input checked="" type="checkbox"/>
Search terms not specified	<input type="checkbox"/>

Notes: Cross-validated some reports. Difference of percentage calculation of serious harm and death (0.9 in abstract, 0.15 in results). The recoding was due to some organisations reporting potential harm rather than actual as the outcome.

Record: 1 of 160

No Filter

Search

B.2 Data input and analysis form for NRLS literature review

Form designed in MS Access and connected to database described in B1, with tag windows for subjects and biases.

Cover Page
1| Total Incidents reports
2| Death & Severe Harm incident reports
3| National & Local Comparison
4| Export Table

NRLS Incident Reporting Ratios

Standardised Incident Reporting Ratios

General Summary

This module uses incident reporting data (NRLS) grouped and correlated with HES based on incident month, organisation and treatment specialty.

Standardised Incident Ratios (SIRR) compare reporting of incidents, adjusting for case-mix, compared to the England average, in both 'All Incident's and 'Severe Harm or Death' (DS) groupings.

Guidance Notes

Cover Page - Details data presented, exclusion and methodology.

1| Total Incident reports - This page contains filter panels, an overview of incident ratios as a funnel plot with 95% and 99% over-dispersed control limits. After making filter selections, please click on the 'Update funnel limits' button to recalculate the funnel plot limits (this is applied to other funnel plots in this module). Selections in Figure 1.1 or 1.2 will highlight organisations for display of cusum plots in Figure 1.3, or and in subsequent pages.

2| Death & Severe Harm incident reports - This page and equivalent funnel plot to page 1 for Death and Severe Harm (DS) incidents. Overdispersed funnel limits are the same as described for page 1, calculated with the same refresh button from page one. Markings on this page correspond with page 1, and changing this marking will alter marking on all other pages.

3| National & Local Comparisons - This page contains two scatter plots where organisation's SIRR's are plotted based on national average case-mix adjustment (marginal) and with an adjustment for local reporting behaviour (conditional). Figure 3.1 relates to All incident reports and Figure 3.2 relates to DS incidents. Markings on this page correspond with pages 1 and 2, and changing this marking will alter marking on all other pages.

4| Export Table - This page contains a single export table, linked to user's login details that allows them to view the record level model data, and aggregated predictors, per month, for your trust. Data can be exported for re-use by right-clicking on the table and selecting 'Export.'

Small number
Small numbers (1-5), and sensitive conditions have been censored in this module. Organisations accessing their own data may view small numbers. Censored cells are indicated in grey.

Analyst's Summary

Data Source



B.3 HED prototype module Cover page (1 of 3)

Cover Page	1 Total Incidents reports	2 Death & Severe Harm incident reports	3 National & Local Comparison	4 Export Table
<div> <div>Analyst's Summary</div> <div> <div>Data Source</div> <p>National Reporting and Learning System (NRLS) – via NHS Improvement.</p> <p>Hospital Episode Statistics (HES) © NHS Digital - 2018. Reused with permission of NHS Digital.</p> </div> <div> <div>Reporting Frequency</div> <p>Model is currently experimental and refreshed quarterly.</p> </div> <div> <div>Dataset</div> <p>Patient-level linkage is not feasible for NRLS data and HES data, as they are collected at different levels of focus, and contains no identifiers. E.g. an incident report may relate to a patient, staff member, equipment, near-miss, potential for an incident etc., whilst each HES episode described the period of inpatient care per patients, under a given consultant/team. NRLS and HES data have therefore been used for create aggregate datasets using monthly numbers of incident reports in harm categories at hospitals and comparing with monthly aggregated casemix data from HES inpatient, outpatient and A&E datasets.</p> </div> <div> <div>Bed-days</div> <p>Bed days</p> </div> <div> <div>Risk Modelling</div> <p>Models are based on extensions of Poisson regression. Data are overdispersed due to clustering at organisations, aggregation and predictors being comparatively poor. Random-intercept models are used to account for repeated measures/clustering at trusts, and non-linearity of predictors was trends represented using Generalized Additive Models with cubic regression splines. The effects of aggregation are countered by extending the Poisson models to use negative binomial distributions that give higher weight to smaller trusts when assessing the overdispersion (i.e. one incident is a greater proportion for a trust with 100 incidents than for a trust with 1000 incidents. Therefore 'noise' is proportionally higher in small trusts).</p> <ul style="list-style-type: none"> • Total number of bed-days • Proportion of bed day in age groups: <1, 1-17, 18-29, 30-49, 50-69,70-84, 85+ (with 30-49 dropped for multicollinearity) • Proportion of bed days with Charlson Comorbidity score: 0, 1-4, 5+ (with 1-4 dropped for multicollinearity) • Proportion of bed days in sex group: male, female ('other' dropped due to low numbers and 'male' fitted as binary, otherwise assumed to be female). • Proportion of bed days in admission method groups: Elective, Non-Elective, Maternity, Transfer (with Elective dropped for multicollinearity). • Proportion of bed-days admitted to a surgical specialty (Tretspet beginning with '1') • Proportion of bed-days on first day of admission • Interaction of emergency admission and admission day • Total number of outpatient attenders • Proportion of outpatient attenders in age groups: <1, 1-17, 18-29, 30-49, 50-69,70-84, 85+ (with 30-49 dropped for multicollinearity) • Total number of A&E attenders • Proportion of A&E attenders arriving by ambulance • Percentile ranges of waiting times for 25th, 50th, and 75th percentiles. • Seasonal fluctuation fitted as a cyclic cubic spline. </div> </div>				

B.3 HED prototype module Cover page (2 of 3)

Cover Page

1| Total Incidents reports

2| Death & Severe Harm incident reports

3| National & Local Comparison

4| Export Table

- Total number of bed-days
- Proportion of bed day in age groups: <1, 1-17, 18-29, 30-49, 50-69, 70-84, 85+ (with 30-49 dropped for multicollinearity)
- Proportion of bed days with Charlson Comorbidity score: 0, 1-4, 5+ (with 1-4 dropped for multicollinearity)
- Proportion of bed days in sex group: male, female ('other' dropped due to low numbers and 'male' fitted as binary, otherwise assumed to be female).
- Proportion of bed days in admission method groups: Elective, Non-Elective, Maternity, Transfer (with Elective dropped for multicollinearity).
- Proportion of bed-days admitted to a surgical specialty (Tretspef beginning with '1')
- Proportion of bed-days on first day of admission
- Interaction of emergency admission and admission day
- Total number of outpatient attenders
- Proportion of outpatient attenders in age groups: <1, 1-17, 18-29, 30-49, 50-69, 70-84, 85+ (with 30-49 dropped for multicollinearity)
- Total number of A&E attenders
- Proportion of A&E attenders arriving by ambulance
- Percentile ranges of waiting times for 25th, 50th, and 75th percentiles.
- Seasonal fluctuation fitted as a cyclic cubic spline.

The number of expected incident reports is predicted by the models and used to create an indirectly-standardised ratio.

Conditional v.s. marginal: Adjusted for organisation or England average

Random intercept models such as the ones used in this module may be used to predict risk in two ways:

- "Conditional" - meaning conditional on the random effects, i.e. with an adjustment for an organisations deviation away from the national average.
- "Marginal" - in this context refer to the average prediction, without the random effects, representing prediction for the English average hospital, given the case mix.

The comparison of these two measures allow users to assess if their reporting ratio is high or low due to their case-mix on it's own (marginal) or their organisational behaviour and case mix (conditional).

Model Fit & Control limits

Models display considerable over-dispersion. Control limits have been calculated to factor in for this using the additive random effects technique described in the Summary Hospital Mortality Index (SHMI specification) [1], developed by Spiegelhalter [2][3].

The OD banding is described as:

2 standard deviations (2SD) from the target, corresponding to a 95% control limit derived from a random effects model [2][3] applying a 10% trim from the top and bottom of all providers for over-dispersion.

[1] www.hscic.gov.uk/SHMI

[2] Spiegelhalter D J (2005) Funnel plots for comparing institutional performance. *Statistics in Medicine*, Apr 24(8): 1185-1202

[3] Spiegelhalter D J (2005) Handling over-dispersion of performance indicators. *Quality & Safety in Health Care*, Oct 14(5): 347-351

Online

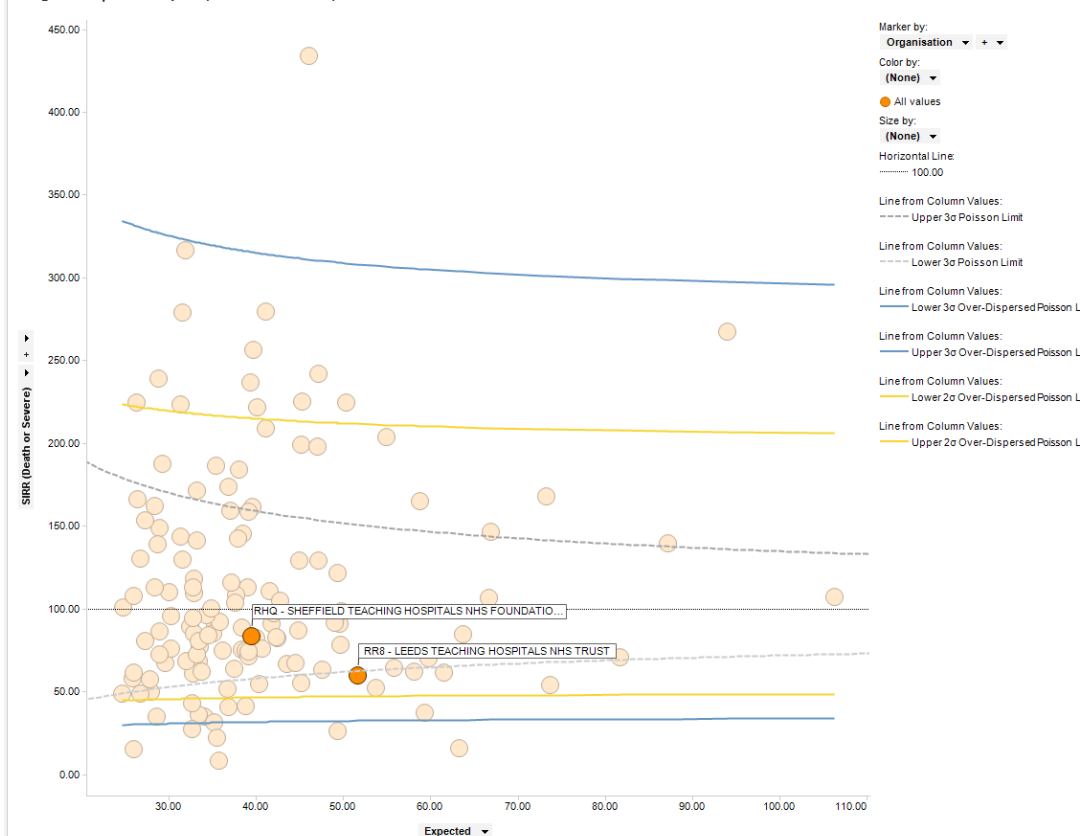
No active visualization

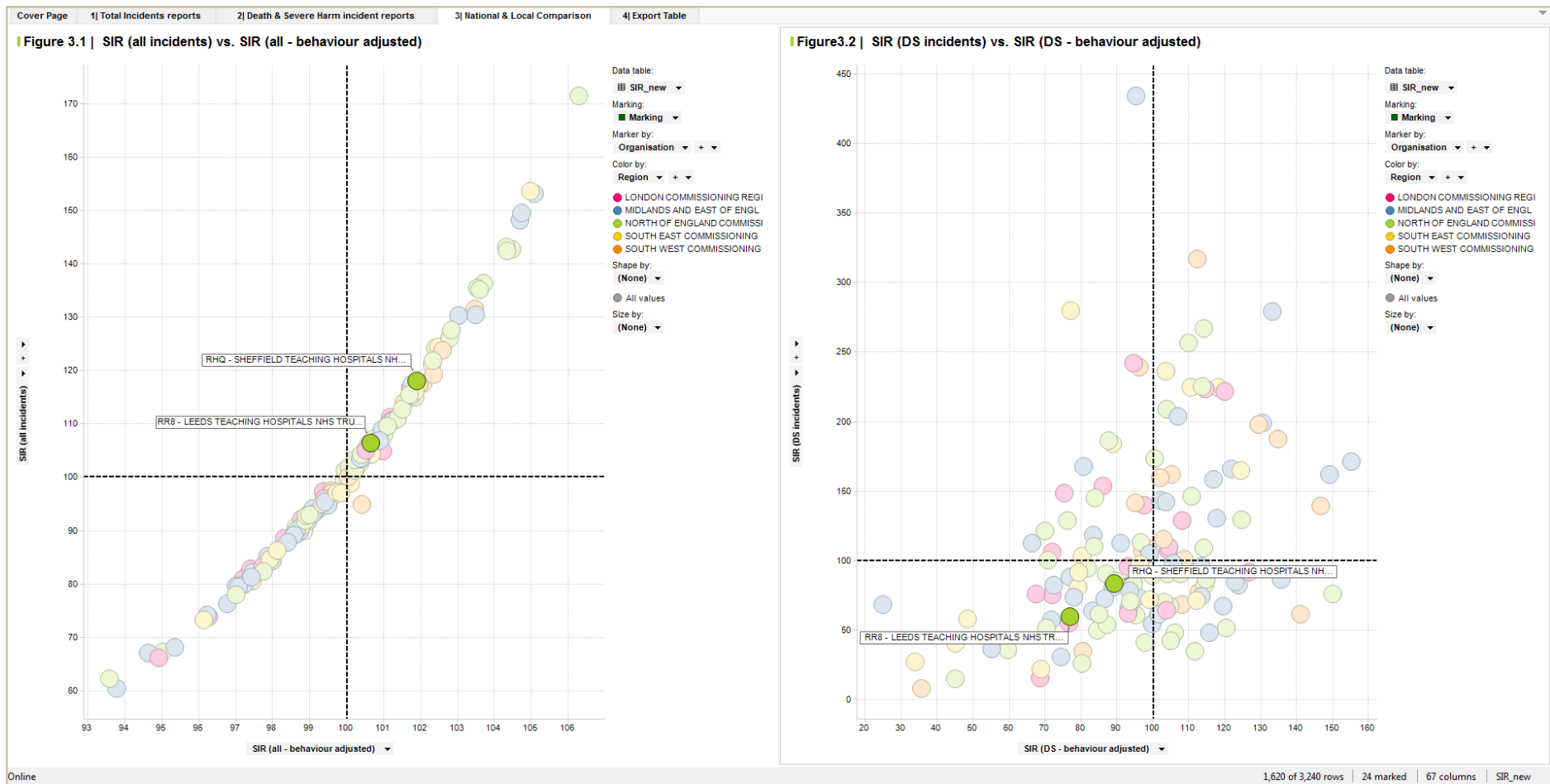
B.3 HED prototype module Cover page (3 of 3) screenshots

Figure 2.1| Standardised Incident Reporting Ratio (SIRR) Summary

(Column Names)					
Organisation	Bed-days	Observed	Expected	SIR	SIR (adjusted)
R1F - ISLE OF WIGHT NHS TRUST	131307	15	25.80	109.28	31.02
R1H - BART'S HEALTH NHS TRUST	920297	122	87.15	263.19	125.15
R1K - LONDON NORTH WEST UNIVERSITY...	544638	114	106.25	201.72	109.32
RA2 - ROYAL SURREY COUNTY HOSPITA...	234711	59	26.25	422.59	49.91
RA3 - WESTON AREA HEALTH NHS TRUST	117236	16	25.95	115.93	11.34
RA4 - YEovil DISTRICT HOSPITAL NHS F...	159039	25	24.75	189.92	22.99
RA7 - UNIVERSITY HOSPITALS BRISTOL ...	422704	64	39.52	304.45	60.84
RA9 - TORBAY AND SOUTH DEVON NHS F...	255692	23	33.41	129.43	21.29
RAE - BRADFORD TEACHING HOSPITALS ...	335058	12	34.09	66.18	10.74
RAJ - SOUTHEND UNIVERSITY HOSPITAL ...	287791	44	38.97	212.27	48.33
RAL - ROYAL FREE LONDON NHS FOUND...	672435	58	44.90	242.83	53.65
RAP - NORTH MIDDLESEX UNIVERSITY H...	255665	25	45.13	104.14	32.61
RAS - THE HILLINGDON HOSPITALS NHS ...	227779	33	29.97	207.02	31.57
RAX - KINGSTON HOSPITAL NHS FOUND...	231317	29	30.23	180.34	31.77
RBA - TAUNTON AND SOMERSET NHS FO...	277163	10	28.63	65.66	12.41
RBD - DORSET COUNTY HOSPITAL NHS F...	206476	28	25.97	202.69	28.81
RBK - WAL SALL HEALTHCARE NHS TRUST	259476	62	39.11	298.01	53.06
RBL - WIRRAL UNIVERSITY TEACHING H...	397023	38	41.77	171.05	43.67
RBN - ST HELENS AND KNOWSLEY HOSPI...	355289	45	49.49	170.95	43.27
RBT - MID CHESHIRE HOSPITALS NHS FO...	250328	36	32.88	205.86	31.52
RBZ - NORTHERN DEVON HEALTHCARE N...	163964	69	28.82	450.10	71.70
RC1 - BEDFORD HOSPITAL NHS TRUST	196225	32	28.34	212.27	48.18
RC9 - LUTON AND DUNSTABLE UNIVERSI...	335656	20	29.57	127.17	16.73
RCB - YORK TEACHING HOSPITAL NHS F...	516226	86	41.12	393.21	82.77
RCD - HARROGATE AND DISTRICT NHS F...	160963	12	24.64	91.55	11.30
RCF - AIREDALE NHS FOUNDATION TRUST	181580	14	27.94	94.21	16.57
RCX - THE QUEEN ELIZABETH HOSPITAL...	231979	35	26.77	245.75	29.70
RD1 - ROYAL UNITED HOSPITALS BATH ...	314453	55	29.30	352.87	40.83
RD3 - POOLE HOSPITAL NHS FOUNDATIO...	264028	26	33.56	145.66	23.06
RD8 - MILTON KEYNES UNIVERSITY HOSP...	226169	16	27.80	108.21	22.28
RDD - BASILDON AND THURROCK UNIVER...	320086	90	45.18	374.50	68.98
RDE - EAST SUFFOLK AND NORTH ESSEX ...	299585	39	32.91	222.80	46.79
RDU - FRIMLEY HEALTH NHS FOUNDATIO...	670309	54	63.65	159.49	57.25
RDZ - THE ROYAL BOURNEMOUTH AND C...	295377	41	37.70	204.46	41.59
RE9 - SOUTH TYNESIDE NHS FOUNDATIO...	148013	4	25.92	29.02	8.87
REF - ROYAL CORNWALL HOSPITALS NH...	365451	59	36.98	299.91	57.85
REM - AINTREE UNIVERSITY HOSPITAL N...	342926	20	32.76	114.79	21.01
RF4 - BARKING, HAVERING AND REDBRID...	493028	71	66.68	200.18	98.50
RFF - BARNSELY HOSPITAL NHS FOUNDA...	200322	29	32.38	168.40	29.03
RFR - THE ROTHERHAM NHS FOUNDATIO...	224326	41	31.59	244.01	32.89
RFS - CHESTERFIELD ROYAL HOSPITAL ...	253252	46	28.40	304.52	30.81
RGN - NORTH WEST ANGLIA NHS FOUND...	416458	112	54.89	383.63	104.67
RGP - JAMES PAGET UNIVERSITY HOSPI...	229684	13	26.69	91.56	11.24
RGQ - IPSWICH HOSPITAL NHS TRUST	302029	45	31.33	270.01	44.08
RGR - WEST SUFFOLK NHS FOUNDATION ...	209834	44	26.45	312.76	36.13

Figure 2.2| Funnel plot (Death or Severe)





B.3 HED prototype module Page 3 National and Local comparisons

Cover Page

1) Total Incidents reports

2) Death & Severe Harm incident reports

3) National & Local Comparison

4) Export Table

4.1|

Export Table

id	trusttype	Organisation	Region	finyr	quarter	month	IP Age <1	IP Age 1 - 17	IP Age 18 - 29	IP Age 30 - 49	IP Age 50 - 69	IP Age 70 - 84	IP Age 85+	age_null	IP Charlson Score 0	IP Charlson Score 1-4	IP Charlson Score 5+	IP male
----	-----------	--------------	--------	-------	---------	-------	-----------	---------------	----------------	----------------	----------------	----------------	------------	----------	---------------------	-----------------------	----------------------	---------

Online1,620 of 3,240 rows0 marked67 columnsSIR_new

Data have been censored to protect small numbers from HED data, and only column headers are available.

Appendix C: Supplementary tables

Organisation	Year																	
	2010/11		2011/12		2012/13		2013/14		2014/15		2015/16		2016/17		Incorrect		Total	
	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents
Clinical Commissioning Groups			0	6	0	6	0	26	0	13	0	30	0	18	0	8	0	107
06H - NHS CAMBRIDGESHIRE AND PETERBOROUGH CCG			0	6	0	6	0	24	0	6	0	13	0	14	0	8	0	77
09Y - NHS NORTH WEST SURREY CCG											0	2					0	2
99H - NHS SURREY DOWNS CCG							0	2	0	7	0	15	0	4			0	28
Independent Sector Healthcare Providers	0	175	44	5,280	65	8,027	137	12,537	141	14,512	134	15,496	66	10,192	0	2	587	66,221
NAX - EAST COAST COMMUNITY HEALTHCARE C.I.C									7	184	9	213	2	41			18	438
NDJ - FIRST COMMUNITY HEALTH AND CARE CIC							0	1	0	107	1	46	0	59			1	213
NHM - SUFFOLK COMMUNITY HEALTHCARE											1	451					1	451
NL3 - CARE PLUS GROUP	0	1	6	158	0	316	1	303	4	256	10	530	11	808	0	1	32	2,373
NL8 - LOCALA COMMUNITY PARTNERSHIPS CIC	0	5	1	730	1	841					0	826	0	499			2	2,901
NLL - PENINSULA COMMUNITY HEALTH C.I.C	0	1	2	7	6	78	82	3,209	72	3,405	39	3,362					201	10,062
NLT - NORTH SOMERSET COMMUNITY PARTNERSHIP COMMUNITY INTEREST COMPANY			7	127	3	314	4	398	2	244	1	259	2	387			19	1,729
NLW - BRISTOL COMMUNITY HEALTH			0	31	0	569	0	814	0	1,040	0	1,439	0	1,839			0	5,732
NLY - SEQOL	0	3	3	298	16	561	11	506	9	394	21	452	5	254			65	2,468
NQ7 - MEDWAY COMMUNITY HEALTHCARE					0	1	0	2									0	3
NQA - PROVIDE	0	34	1	1,025	0	1,322	1	1,414	2	1,652	1	1,197	2	457			7	7,101
NQE - NENE COMMISSIONING COMMUNITY INTEREST COMPANY							0	3									0	3
NQL - NAVIGO HEALTH AND SOCIAL CARE CIC			2	157	10	242	10	227	13	186	6	426	4	291			45	1,529
NQV - BROMLEY HEALTHCARE			15	691	15	963	13	1,287	8	1,413	10	1,530	1	1,018			62	6,902
NR3 - NOTTINGHAM CITYCARE PARTNERSHIP	0	127	7	1,625	8	2,084	4	2,946	17	3,436	6	2,391	6	1,614			48	14,223
NRS - LIVEWELL SOUTHWEST	0	4	0	431	6	736	11	1,427	7	2,195	29	2,374	33	2,925	0	1	86	10,093
NHS Care Trusts	16	840	26	1,219	1	121											43	2,180
TAL - TORBAY CARE TRUST	16	840	26	1,219	1	121											43	2,180
NHS Primary Care Trusts	1,653	131,976	643	37,310	208	14,623	2	81									2,506	183,990
SA3 - SOUTH GLOUCESTERSHIRE PCT	0	247			0	3											0	250
SA4 - HAVERING PCT	1	1,275	0	810	0	11											1	2,096
SA5 - KINGSTON PCT	10	106															10	106
SA7 - BROMLEY PCT	20	564	0	5	0	1											20	570
SA8 - GREENWICH TEACHING PCT	0	371	0	1	0	1											0	373
SA9 - BARNET PCT	7	269			0	2											7	271
SA7 - HILLINGDON PCT	8	1,068	1	19	0	14											9	1,101
SC1 - ENFIELD PCT	10	137	1	23													11	160
SC2 - BARKING AND DAGENHAM PCT	5	320	1	2	0	10											6	332
SC3 - CITY AND HACKNEY TEACHING PCT	1	20															1	20
SC4 - TOWER HAMLETS PCT	5	497	0	5	0	2											5	504
SC5 - NEWHAM PCT	4	182	0	2	0	2											4	186
SC9 - HARINGEY TEACHING PCT	2	127															2	127
SCN - HEREFORDSHIRE PCT	22	2,147			0	3											22	2,150
SCQ - MILTON KEYNES PCT	27	1,436	16	1,669	23	1,925	2	11									68	5,041
SD7 - NEWCASTLE PCT	1	2	0	10	1	9	0	1									2	22
SD8 - NORTH TYNESIDE PCT	5	651	0	2	0	1											5	654
SD9 - HARTLEPOOL PCT	0	40	3	62	0	11											3	113
SEF - NORTH LINCOLNSHIRE PCT	3	250	0	8	0	13	0	2									3	273
SEM - NOTTINGHAM CITY PCT	0	1,357	0	2	0	10											0	1,369
SET - BASSETLAW PCT	0	55	0	30	0	1											0	86
SF1 - PLYMOUTH TEACHING PCT	1	922	7	897	0	4											8	1,823
SF5 - SALFORD PCT	10	239	0	1	0	2											10	242

Organisation	Year																	
	2010/11		2011/12		2012/13		2013/14		2014/15		2015/16		2016/17		Incorrect		Total	
	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents
5F7 - STOCKPORT PCT	2	284	4	363	1	4											7	651
5FE - PORTSMOUTH CITY TEACHING PCT	1	5	0	1													1	6
5FL - BATH AND NORTH EAST SOMERSET PCT	4	356	4	252													8	608
5GC - LUTON PCT	8	254	1	13	0	3	0	2									9	272
5H1 - HAMMERSMITH AND FULHAM PCT	1	22	0	14	0	6											1	42
5H8 - ROTHERHAM PCT	2	193	0	3	0	20	0	1									2	217
5HG - ASHTON, LEIGH AND WIGAN PCT	1	636	0	47	0	39											1	722
5HP - BLACKPOOL PCT	9	581	9	333	0	1											18	915
5HQ - BOLTON PCT	3	327	0	117	0	2											3	446
5HX - EALING PCT	6	208	1	17	0	7											7	232
5HY - HOUNSLOW PCT	1	38	2	23	0	10											3	71
5J2 - WARRINGTON PCT	1	74	1	13	0	2											2	89
5J4 - KNOWSLEY PCT	2	53	0	1	0	7											2	61
5J5 - OLDHAM PCT	0	458	0	122	0	10											0	590
5J6 - CALDERDALE PCT	23	776	8	195	5	163											36	1,134
5J9 - DARLINGTON PCT	3	618	0	7	3	444	0	1									6	1,070
5JE - BARNSLEY PCT	14	1,189	0	145	0	24	0	1									14	1,359
5JX - BURY PCT	0	166	0	2													0	168
5K3 - SWINDON PCT	8	1,121	3	537	0	22	0	1									11	1,681
5K5 - BRENT TEACHING PCT	4	184	0	4	0	3											4	191
5K6 - HARROW PCT	4	135	0	3	0	1	0	1									4	140
5K7 - CAMDEN PCT	1	794															1	794
5K8 - ISLINGTON PCT	15	441	1	76	0	138											16	655
5K9 - CROYDON PCT	0	13	0	8	0	21	0	1									0	43
5KF - GATESHEAD PCT	2	398	1	184	1	97											4	679
5KG - SOUTH TYNESIDE PCT	1	437	2	156	1	79											4	672
5KL - SUNDERLAND TEACHING PCT	0	513	2	302	1	213	0	2									3	1,030
5KM - MIDDLESBROUGH PCT	0	29	0	54	0	45	0	6									0	134
5L1 - SOUTHAMPTON CITY PCT	51	2,034	0	91	0	11											51	2,136
5L3 - MEDWAY PCT	1	617	0	6	0	3	0	1									1	627
5LA - KENSINGTON AND CHELSEA PCT	7	854			0	1											7	855
5LC - WESTMINSTER PCT	0	3															0	3
5LD - LAMBETH PCT	5	284	2	161	0	25											7	470
5LE - SOUTHWARK PCT	1	60	1	40	0	3											2	103
5LF - LEWISHAM PCT	0	27	1	7	0	6											1	40
5LG - WANDSWORTH PCT	13	530															13	530
5LH - TAMESIDE AND GLOSSOP PCT	2	425	0	1	0	5											2	431
5LQ - BRIGHTON AND HOVE CITY PCT	4	10	0	2	1	10											5	22
5M1 - SOUTH BIRMINGHAM PCT	13	2,620	7	1,171	0	4	0	2									20	3,797
5M2 - SHROPSHIRE COUNTY PCT	4	1,288	2	345	0	3											6	1,636
5M3 - WALSALL TEACHING PCT	21	768	0	52	0	34											21	854
5M6 - RICHMOND AND TWICKENHAM PCT	6	388	0	1													6	389
5M7 - SUTTON AND MERTON PCT	0	117	0	3	0	22	0	1									0	143
5M8 - NORTH SOMERSET PCT	3	101	3	73	0	1											6	175
5MD - COVENTRY TEACHING PCT	3	583					0	1									3	584
5MK - TELFORD AND WREKIN PCT	3	44	0	26	0	1											3	71

Organisation	Year																	
	2010/11		2011/12		2012/13		2013/14		2014/15		2015/16		2016/17		Incorrect		Total	
	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents
5MV - WOLVERHAMPTON CITY PCT	45	1,389	9	231	0	6											54	1,626
5MX - HEART OF BIRMINGHAM TEACHING PCT	11	880	0	3													11	883
5N1 - LEEDS PCT	39	3,368	9	830	4	564											52	4,762
5N2 - KIRKLEES PCT	5	1,103	0	631	0	8											5	1,742
5N3 - WAKEFIELD DISTRICT PCT	5	2,009	2	111	2	70											9	2,190
5N4 - SHEFFIELD PCT	4	737	0	8	0	7											4	752
5N5 - DONCASTER PCT	38	3,029	0	3	0	2											38	3,034
5N6 - DERBYSHIRE COUNTY PCT	17	3,942	0	38	0	146	0	3									17	4,129
5N7 - DERBY CITY PCT	4	265	0	9	0	4											4	278
5N8 - NOTTINGHAMSHIRE COUNTY TEACHING PCT	10	1,264	0	77	0	66	0	1									10	1,408
5N9 - LINCOLNSHIRE TEACHING PCT	3	1,275	0	5	0	6											3	1,286
5NA - REDBRIDGE PCT	1	14	0	8	0	18											1	40
5NC - WALTHAM FOREST PCT	2	42			1	6											3	48
5ND - COUNTY DURHAM PCT	31	665	9	575	11	1,292											51	2,532
5NE - CUMBRIA TEACHING PCT	2	1,570	0	15	0	7											2	1,592
5NF - NORTH LANCASHIRE TEACHING PCT	6	575	7	709	0	9											13	1,293
5NG - CENTRAL LANCASHIRE PCT	7	1,304	0	127	0	5											7	1,436
5NH - EAST LANCASHIRE TEACHING PCT	25	960	2	50	0	8											27	1,018
5NJ - SEFTON PCT	9	479	0	16	0	5											9	500
5NK - WIRRAL PCT	4	1,081	0	5	0	15	0	1									4	1,102
5NL - LIVERPOOL PCT	0	415	1	145	0	3											1	563
5NM - HALTON AND ST HELENS PCT	0	395	0	4	0	12											0	411
5NN - WESTERN CHESHIRE PCT	15	912	13	241	0	24											28	1,177
5NP - CENTRAL AND EASTERN CHESHIRE PCT	10	1,484	3	209	0	1											13	1,694
5NQ - HEYWOOD, MIDDLETON AND ROCHDALE PCT	0	36	0	1													0	37
5NR - TRAFFORD PCT	0	107	0	1													0	108
5NT - MANCHESTER PCT	2	397	0	8	0	8											2	413
5NV - NORTH YORKSHIRE AND YORK PCT	16	3,557	3	1,629	3	142											22	5,328
5NW - EAST RIDING OF YORKSHIRE PCT	20	583	2	157	4	63	0	1									26	804
5NX - HULL TEACHING PCT	14	324	4	184	4	185	0	12									22	705
5NY - BRADFORD AND AIREDALE TEACHING PCT	6	1,417	1	41	0	15											7	1,473
5P1 - SOUTH EAST ESSEX PCT	22	331	4	69	0	2											26	402
5P2 - BEDFORDSHIRE PCT	2	1,407	12	608	2	9											16	2,024
5P5 - SURREY PCT	11	2,309	30	2,545	5	1,039	0	1									46	5,894
5P6 - WEST SUSSEX PCT	27	1,226	0	1	0	11	0	1									27	1,239
5P7 - EAST SUSSEX DOWNS AND WEALD PCT	5	807	0	1	3	6											8	814
5P8 - HASTINGS AND ROTHER PCT	1	529	0	15	0	3											1	547
5P9 - WEST KENT PCT	6	1,172	0	3	0	3											6	1,178
5PA - LEICESTERSHIRE COUNTY AND RUTLAND PCT	10	2,223	0	70	1	74	0	1									11	2,368
5PC - LEICESTER CITY PCT	12	572	0	34	0	55	0	1									12	662
5PD - NORTHAMPTONSHIRE TEACHING PCT	35	1,650	5	200	0	13	0	1									40	1,864
5PE - DUDLEY PCT	0	194	0	22	1	22											1	238
5PF - SANDWELL PCT	48	700	29	1,247	6	158											83	2,105
5PG - BIRMINGHAM EAST AND NORTH PCT	20	1,397	0	5													20	1,402
5PH - NORTH STAFFORDSHIRE PCT	3	543	12	590													15	1,133
5PJ - STOKE ON TRENT PCT	43	869	6	249	0	3											49	1,121

Organisation	Year																	
	2010/11		2011/12		2012/13		2013/14		2014/15		2015/16		2016/17		Incorrect		Total	
	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents
SPK - SOUTH STAFFORDSHIRE PCT	6	756	9	407	0	53	0	1									15	1,217
SPL - WORCESTERSHIRE PCT	15	1,729	0	372	0	5											15	2,106
SPM - WARWICKSHIRE PCT	22	1,022	20	720	0	5	0	2									42	1,749
SPN - PETERBOROUGH PCT	7	1,305	4	1,042	0	20	0	4									11	2,371
SPP - CAMBRIDGESHIRE PCT	19	1,446	0	59	0	34											19	1,539
SPQ - NORFOLK PCT	5	3,955	1	12	0	43	0	1									6	4,011
SPR - GREAT YARMOUTH AND WAVENEY PCT	11	698	24	408	0	6											35	1,112
SPT - SUFFOLK PCT	1	981	4	741	3	271	0	1									8	1,994
SPV - WEST ESSEX PCT	23	568	9	123	5	38											37	729
SPW - NORTH EAST ESSEX PCT	7	987	0	223	0	96											7	1,306
SPX - MID ESSEX PCT	4	764	0	7	0	13											4	784
SPY - SOUTH WEST ESSEX PCT	5	1,103	4	124	0	1											9	1,228
SQA - EASTERN AND COASTAL KENT PCT	15	1,939	2	184	0	102	0	6									17	2,231
SQC - HAMPSHIRE PCT	81	2,456	0	18	1	8	0	2									82	2,484
SQD - BUCKINGHAMSHIRE PCT	2	192	0	1	0	5											2	198
SQE - OXFORDSHIRE PCT	45	2,524	8	108	0	26											53	2,658
SQF - BERKSHIRE WEST PCT	5	1,453	0	73	0	58	0	2									5	1,586
SQG - BERKSHIRE EAST PCT	2	399	0	25	0	12	0	1									2	437
SQH - GLOUCESTERSHIRE PCT	44	3,074	68	2,588	59	3,156											171	8,818
SQJ - BRISTOL PCT	2	624	0	640	1	119											3	1,383
SQK - WILTSHIRE PCT	6	2,197	2	352	0	7											8	2,556
SQL - SOMERSET PCT	15	2,489	6	1,477	0	31											21	3,997
SQM - DORSET PCT	89	2,611	47	765	0	12											136	3,388
SQN - BOURNEMOUTH AND POOLE TEACHING PCT	0	832	0	102	0	12	0	1									0	947
SQP - CORNWALL AND ISLES OF SCILLY PCT	13	2,563	66	2,602	55	2,738											134	7,903
SQQ - DEVON PCT	113	2,954	11	389	0	121											124	3,464
SQR - REDCAR AND CLEVELAND PCT	7	515	0	22	0	4											7	541
SQT - ISLE OF WIGHT NHS PCT	95	2,636	109	3,446													204	6,082
SQV - HERTFORDSHIRE PCT	3	2,359	2	30	0	17	0	2									5	2,408
SQW - SOLIHULL PCT			0	1													0	1
NHS Trusts (acute trust)	6,762	821,192	7,111	933,947	6,869	1,030,798	6,427	1,129,209	6,512	1,257,849	6,508	1,305,897	5,248	1,169,877	10	742	45,447	7,649,511
R1F - ISLE OF WIGHT NHS TRUST	0	1	0	12	94	3,043	102	2,620	95	3,795	41	3,801	13	3,710	0	16	345	16,998
R1H - BARTS HEALTH NHS TRUST	3	19	8	117	113	14,190	55	20,219	74	22,907	123	23,797	112	23,822	0	15	488	105,086
R1K - LONDON NORTH WEST UNIVERSITY HEALTHCARE NHS TRUST	1	3	1	2	2	9	4	446	14	942	113	11,492	98	12,905	0	17	233	25,816
RA2 - ROYAL SURREY COUNTY HOSPITAL NHS FOUNDATION TRUST	5	4,013	14	4,915	7	4,155	39	4,415	52	5,464	44	5,445	51	4,675	1	4	213	33,086
RA3 - WESTON AREA HEALTH NHS TRUST	50	1,752	17	2,019	5	2,064	12	2,583	9	3,306	5	3,708	13	3,157			111	18,589
RA4 - YEOVIL DISTRICT HOSPITAL NHS FOUNDATION TRUST	16	2,186	14	2,773	2	2,667	6	3,508	14	3,508	20	3,719	25	3,730			97	22,091
RA7 - UNIVERSITY HOSPITALS BRISTOL NHS FOUNDATION TRUST	142	8,354	100	9,073	77	11,236	41	12,154	51	13,079	55	13,594	57	13,492	0	5	523	80,987
RA9 - TORBAY AND SOUTH DEVON NHS FOUNDATION TRUST	58	4,001	48	4,238	27	4,352	13	4,516	9	5,260	18	5,684	19	5,106			192	33,157
RAE - BRADFORD TEACHING HOSPITALS NHS FOUNDATION TRUST	56	5,886	66	6,833	49	7,215	14	7,727	20	9,061	8	9,663	12	7,601	0	19	225	54,005
RAJ - SOUTHEND UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	25	4,175	22	5,365	37	5,964	30	6,862	60	7,508	52	8,691	43	7,907	0	10	269	46,482
RAL - ROYAL FREE LONDON NHS FOUNDATION TRUST	74	4,857	65	4,781	53	4,742	33	4,869	66	9,433	50	11,081	53	11,161	0	1	394	50,925
RAP - NORTH MIDDLESEX UNIVERSITY HOSPITAL NHS TRUST	48	3,025	50	3,281	27	2,752	12	4,204	22	7,055	39	6,754	18	6,296	0	8	216	33,375
RAS - THE HILLINGDON HOSPITALS NHS FOUNDATION TRUST	19	3,700	42	4,250	47	4,686	54	5,310	45	5,633	30	5,645	28	5,028			265	34,252
RAX - KINGSTON HOSPITAL NHS FOUNDATION TRUST	38	2,902	14	1,757	26	3,081	43	4,739	25	4,700	32	4,957	12	4,292			190	26,428
RBA - TAUNTON AND SOMERSET NHS FOUNDATION TRUST	47	3,504	29	3,995	16	5,352	10	6,120	23	6,463	13	6,123	10	5,901			148	37,458

Organisation	Year																	
	2010/11		2011/12		2012/13		2013/14		2014/15		2015/16		2016/17		Incorrect		Total	
	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents
RBD - DORSET COUNTY HOSPITAL NHS FOUNDATION TRUST	32	2,777	30	3,001	25	3,078	1	3,276	38	4,116	26	4,542	17	2,793	0	1	169	23,584
RBK - WALSALL HEALTHCARE NHS TRUST	11	3,457	115	4,677	133	9,835	52	9,661	29	10,438	47	11,486	56	9,257			443	58,811
RBL - WIRRAL UNIVERSITY TEACHING HOSPITAL NHS FOUNDATION TRUST	13	8,875	25	9,849	15	12,317	20	10,174	32	9,731	44	9,415	36	10,082			185	70,443
RBN - ST HELENS AND KNOWSLEY HOSPITAL SERVICES NHS TRUST	2	4,220	11	4,879	11	7,870	4	6,008	46	8,158	38	9,067	42	8,411	0	3	154	48,616
RBT - MID CHESHIRE HOSPITALS NHS FOUNDATION TRUST	6	6,171	8	6,393	11	6,546	7	6,746	13	6,156	26	6,541	33	5,877			104	44,430
RBZ - NORTHERN DEVON HEALTHCARE NHS TRUST	55	4,527	138	7,480	112	9,239	82	8,230	89	7,831	89	7,539	68	5,370	0	6	633	50,222
RC1 - BEDFORD HOSPITAL NHS TRUST	12	1,759	8	1,850	27	2,755	22	3,223	41	4,743	64	5,131	30	4,992			204	24,453
RC3 - EALING HOSPITAL NHS TRUST	16	2,411	22	2,649	26	2,473	44	3,176	36	3,981	6	181					150	14,871
RC9 - LUTON AND DUNSTABLE UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	22	5,113	26	7,977	24	8,782	24	7,527	43	6,841	14	6,917	10	3,346	0	1	163	46,504
RCB - YORK TEACHING HOSPITAL NHS FOUNDATION TRUST	48	9,895	42	9,265	67	9,573	88	11,089	71	10,680	93	12,189	73	8,656	0	2	482	71,349
RCC - SCARBOROUGH AND NORTH EAST YORKSHIRE HEALTH CARE NHS TRUST	21	2,630	17	2,639	6	503											44	5,772
RCD - HARROGATE AND DISTRICT NHS FOUNDATION TRUST	2	2,164	2	2,729	7	3,348	8	2,514	5	3,230	9	3,997	10	4,120	0	2	43	22,104
RCF - AIREDALE NHS FOUNDATION TRUST	19	3,486	7	3,857	7	4,563	9	4,914	12	4,726	18	5,044	12	4,165			84	30,755
RCX - THE QUEEN ELIZABETH HOSPITAL, KING'S LYNN, NHS FOUNDATION TRUST	13	4,739	13	5,460	20	5,485	17	7,160	29	7,154	27	6,307	34	4,730			153	41,035
RD1 - ROYAL UNITED HOSPITALS BATH NHS FOUNDATION TRUST	21	2,630	40	2,669	27	3,049	65	4,888	21	7,735	25	6,850	42	5,881			241	33,702
RD3 - POOLE HOSPITAL NHS FOUNDATION TRUST	17	6,902	22	7,895	18	8,020	19	7,434	4	8,645	21	8,676	26	7,118	0	4	127	54,694
RD7 - HEATHERWOOD AND WEXHAM PARK HOSPITALS NHS FOUNDATION TRUST	58	5,743	65	5,182	78	6,320	133	6,000	43	3,859	0	14					377	27,118
RD8 - MILTON KEYNES UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	37	3,398	22	3,185	22	3,096	25	3,307	33	4,276	24	4,629	6	4,575	0	1	169	26,467
RDD - BASILDON AND THURROCK UNIVERSIT HOSPITALS NHS FOUNDATION TRUST	46	4,324	81	3,488	67	1,952	140	7,993	30	11,097	49	10,479	85	9,606			498	48,939
RDE - EAST SUFFOLK AND NORTH ESSEX NHS FOUNDATION TRUST	29	3,966	16	4,303	29	5,387	44	7,459	26	7,127	53	7,605	36	7,051			233	42,898
RDU - FRIMLEY HEALTH NHS FOUNDATION TRUST	29	4,398	19	5,601	17	5,171	25	5,019	54	8,398	56	12,424	36	11,818	0	3	236	52,832
RDZ - THE ROYAL BOURNEMOUTH AND CHRISTCHURCH HOSPITALS NHS FOUNDATION TRUST	19	4,227	29	5,192	30	6,134	42	7,163	26	7,198	42	8,038	32	7,501			220	45,453
RE9 - SOUTH TYNESIDE NHS FOUNDATION TRUST	6	2,818	31	3,257	19	4,455	11	4,122	20	4,545	12	3,776	4	2,811			103	25,784
REF - ROYAL CORNWALL HOSPITALS NHS TRUST	87	6,279	46	6,878	57	7,731	60	8,982	56	9,032	51	9,151	39	8,263	0	6	396	56,322
REM - AINTREE UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	14	4,824	9	6,065	12	5,915	17	5,759	14	5,485	20	5,943	18	5,788	0	2	104	39,781
RF4 - BARKING, HAVERING AND REDBRIDGE UNIVERSITY HOSPITALS NHS TRUST	13	7,556	21	7,983	27	7,427	30	8,099	84	8,347	117	9,523	58	8,274			350	57,209
RFF - BARNSELY HOSPITAL NHS FOUNDATION TRUST	36	2,980	42	3,401	44	3,262	38	6,281	25	7,173	30	7,033	25	6,343			240	36,473
RFR - THE ROTHERHAM NHS FOUNDATION TRUST	5	4,729	15	5,406	8	6,186	30	6,116	47	6,680	24	6,427	32	6,047	0	10	161	41,601
RFS - CHESTERFIELD ROYAL HOSPITAL NHS FOUNDATION TRUST	26	4,809	29	5,023	27	5,503	28	5,843	17	6,067	15	5,846	42	5,288	1	5	185	38,384
RFW - WEST MIDDLESEX UNIVERSITY HOSPITAL NHS TRUST	15	2,175	17	2,339	24	2,371	32	3,115	36	4,345	39	3,317			0	1	163	17,663
RGC - WHIPPS CROSS UNIVERSITY HOSPITAL NHS TRUST	61	2,851	69	4,791													130	7,642
RGN - NORTH WEST ANGLIA NHS FOUNDATION TRUST	27	5,983	11	7,226	20	8,068	22	7,958	54	7,881	47	8,077	50	7,392			231	52,585
RGP - JAMES PAGET UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	25	3,097	21	3,840	32	4,971	14	5,991	18	5,058	8	4,337	12	3,989			130	31,283
RGQ - IPSWICH HOSPITAL NHS TRUST	12	6,593	18	6,777	25	8,057	23	5,364	42	5,539	39	6,344	35	6,613	0	1	194	45,288
RGR - WEST SUFFOLK NHS FOUNDATION TRUST	52	2,690	34	2,866	36	3,504	42	4,136	25	4,089	27	4,289	40	3,629			256	25,203
RGT - CAMBRIDGE UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	26	10,734	21	12,533	40	12,090	42	12,067	33	12,104	18	11,428	51	8,977			231	79,933
RH8 - ROYAL DEVON AND EXETER NHS FOUNDATION TRUST	18	6,205	28	6,452	19	8,609	16	10,029	9	13,059	11	11,553	1	9,606	0	8	102	65,521
RHM - UNIVERSITY HOSPITAL SOUTHAMPTON NHS FOUNDATION TRUST	85	6,268	91	8,929	82	9,265	108	12,792	155	14,298	178	14,931	100	14,735			799	81,218
RHQ - SHEFFIELD TEACHING HOSPITALS NHS FOUNDATION TRUST	56	10,453	51	10,833	58	11,136	75	11,720	62	17,746	41	20,529	25	17,677	0	20	368	100,114
RHU - PORTSMOUTH HOSPITALS NHS TRUST	70	8,636	115	7,687	114	7,976	112	7,800	125	8,959	75	8,991	97	12,195	0	5	708	62,249
RHW - ROYAL BERKSHIRE NHS FOUNDATION TRUST	44	5,351	38	4,940	19	5,529	17	4,939	18	8,477	26	10,253	35	9,059	0	7	197	48,555
RJ1 - GUY'S AND ST THOMAS' NHS FOUNDATION TRUST	71	8,736	35	9,126	35	10,284	14	12,203	48	14,614	54	16,894	78	15,421	0	17	335	87,295
RJ2 - LEWISHAM AND GREENWICH NHS TRUST	38	2,873	40	4,045	27	3,603	26	7,323	69	11,302	17	13,187	6	10,518			223	52,851
RJ6 - CROYDON HEALTH SERVICES NHS TRUST	46	3,282	40	5,189	163	5,176	89	4,519	57	4,404	44	5,042	66	6,263	0	1	505	33,876
RJ7 - ST GEORGE'S UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	70	6,874	24	9,707	16	9,505	14	9,867	33	10,324	51	10,744	26	7,185	0	3	234	64,209
RJC - SOUTH WARWICKSHIRE NHS FOUNDATION TRUST	82	2,670	53	3,776	45	4,510	94	5,557	194	5,285	153	6,810	20	6,737	0	1	641	35,346

Organisation	Year																	
	2010/11 Death or Severe	Total Incidents	2011/12 Death or Severe	Total Incidents	2012/13 Death or Severe	Total Incidents	2013/14 Death or Severe	Total Incidents	2014/15 Death or Severe	Total Incidents	2015/16 Death or Severe	Total Incidents	2016/17 Death or Severe	Total Incidents	Incorrect Death or Severe	Total Incidents	Total Death or Severe	Total Incidents
RJD - MID STAFFORDSHIRE NHS FOUNDATION TRUST	131	3,612	168	4,772	113	5,082	21	4,919	12	2,714							445	21,099
RJE - UNIVERSITY HOSPITALS OF NORTH MIDLANDS NHS TRUST	74	6,482	63	6,322	71	7,964	22	8,226	48	12,141	18	13,671	28	11,382			324	66,188
RJF - BURTON HOSPITALS NHS FOUNDATION TRUST	28	4,006	67	5,421	93	5,883	40	4,969	39	4,692	15	5,504	27	3,408			309	33,883
RJL - NORTHERN LINCOLNSHIRE AND GOOLE NHS FOUNDATION TRUST	10	7,450	18	8,330	14	9,437	17	9,703	26	10,703	16	10,617	26	11,076	0	6	127	67,322
RJN - EAST CHESHIRE NHS TRUST	3	2,585	18	4,478	23	4,515	14	5,491	8	6,687	8	7,774	23	6,417	0	8	97	37,955
RJR - COUNTESS OF CHESTER HOSPITAL NHS FOUNDATION TRUST	7	7,211	17	7,068	22	4,529	18	7,158	12	7,679	29	9,734	24	8,248			129	51,627
RJZ - KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	89	7,830	31	6,918	41	8,117	124	14,480	147	19,130	151	21,494	94	11,189	0	26	677	89,184
RKS - SHERWOOD FOREST HOSPITALS NHS FOUNDATION TRUST	7	4,263	7	5,724	9	6,302	16	6,441	16	7,120	18	7,275	6	5,430	0	8	79	42,563
RK9 - UNIVERSITY HOSPITALS PLYMOUTH NHS TRUST	44	5,346	56	7,405	93	9,755	92	10,403	81	12,614	54	13,023	41	11,153	0	12	461	69,711
RKB - UNIVERSITY HOSPITALS COVENTRY AND WARWICKSHIRE NHS TRUST	28	8,889	23	10,671	30	10,206	46	10,751	34	11,696	61	12,117	34	11,593			256	75,923
RKE - WHITTINGTON HEALTH NHS TRUST	70	2,625	110	3,579	75	2,853	73	3,817	78	3,902	72	4,130	31	3,244			509	24,150
RL4 - THE ROYAL WOLVERHAMPTON NHS TRUST	23	6,379	45	8,805	28	8,616	10	9,024	34	9,913	32	9,966	24	7,310			196	60,013
RLN - CITY HOSPITALS SUNDERLAND NHS FOUNDATION TRUST	124	6,291	62	5,658	94	8,672	20	10,389	16	12,799	40	14,135	30	12,274			386	70,218
RLQ - WYE VALLEY NHS TRUST	14	2,595	19	3,425	24	3,567	10	3,575	9	5,958	11	6,452	21	6,171	0	2	108	31,745
RLT - GEORGE ELIOT HOSPITAL NHS TRUST	126	2,068	82	1,971	20	2,182	69	3,678	5	4,013	26	4,456	21	4,203	0	1	349	22,572
RM1 - NORFOLK AND NORWICH UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	73	8,700	263	11,096	128	12,700	11	13,345	16	15,152	32	15,487	20	10,798	0	6	543	87,284
RM2 - UNIVERSITY HOSPITAL OF SOUTH MANCHESTER NHS FOUNDATION TRUST	18	4,332	34	5,446	52	7,442	34	7,724	31	9,881	33	11,027	29	12,293	0	2	231	58,147
RM3 - SALFORD ROYAL NHS FOUNDATION TRUST	33	7,264	40	7,278	28	7,470	37	7,616	27	9,760	24	9,074	21	7,434			210	55,896
RM4 - TRAFFORD HEALTHCARE NHS TRUST	0	911	2	1,624													2	2,535
RM6 - BOLTON NHS FOUNDATION TRUST	17	4,141	23	3,103	19	4,870	36	5,623	40	7,585	25	8,990	19	8,672			179	42,984
RMP - TAMESIDE AND GLOSSOP INTEGRATED CARE NHS FOUNDATION TRUST	7	3,588	8	3,454	17	4,135	8	6,094	10	5,798	14	7,791	15	7,067			79	37,927
RN3 - GREAT WESTERN HOSPITALS NHS FOUNDATION TRUST	27	4,710	44	6,516	42	6,908	44	6,870	43	6,663	35	6,165	36	6,161	0	60	271	44,053
RN5 - HAMPSHIRE HOSPITALS NHS FOUNDATION TRUST	49	4,231	66	5,507	103	6,753	56	6,362	69	6,788	57	8,457	34	7,786	0	4	434	45,888
RN7 - DARTFORD AND GRAVESHAM NHS TRUST	2	3,935	10	3,886	23	4,175	16	4,598	12	6,269	12	7,333	8	5,952	0	2	83	36,150
RNA - THE DUDLEY GROUP NHS FOUNDATION TRUST	73	9,330	99	10,732	48	9,128	19	10,489	32	9,574	48	7,380	30	7,346	0	5	349	63,984
RNH - NEWHAM UNIVERSITY HOSPITAL NHS TRUST	11	3,898	10	3,788													21	7,686
RNI - BARTS AND THE LONDON NHS TRUST	82	6,578	52	8,082	0	3											134	14,663
RNL - NORTH CUMBRIA UNIVERSITY HOSPITALS NHS TRUST	11	2,548	18	3,339	50	5,092	59	5,736	79	6,171	64	8,076	56	7,295	0	2	337	38,259
RNQ - KETTERING GENERAL HOSPITAL NHS FOUNDATION TRUST	15	2,910	22	2,935	37	4,397	98	5,661	67	6,562	44	6,175	86	5,755	0	3	369	34,398
RNS - NORTHAMPTON GENERAL HOSPITAL NHS TRUST	23	4,599	37	5,629	47	6,764	44	7,653	31	6,522	28	7,272	13	4,749	0	6	223	43,194
RNZ - SALISBURY NHS FOUNDATION TRUST	41	4,826	36	5,140	34	5,566	22	6,220	17	6,348	13	6,029	33	6,254			196	40,383
RP5 - DONCASTER AND BASSETLAW TEACHING HOSPITALS NHS FOUNDATION TRUST	70	2,005	122	4,208	170	5,186	186	6,696	157	10,332	104	11,259	57	8,840	0	1	866	48,527
RPA - MEDWAY NHS FOUNDATION TRUST	35	5,082	36	3,714	39	4,385	67	5,868	69	6,650	76	5,451	53	7,161	0	17	375	38,328
RQ6 - ROYAL LIVERPOOL AND BROADGREEN UNIVERSITY HOSPITALS NHS TRUST	33	4,806	5	4,565	7	4,261	28	5,053	47	8,163	35	10,657	9	11,241	0	2	164	48,748
RQ8 - MID ESSEX HOSPITAL SERVICES NHS TRUST	9	3,277	40	4,614	40	6,395	31	5,838	26	5,655	22	5,066	15	5,977	0	13	183	36,835
RQM - CHELSEA AND WESTMINSTER HOSPITAL NHS FOUNDATION TRUST	14	4,962	1	4,976	4	5,154	2	5,892	13	6,472	12	5,812	29	6,925	0	2	75	40,195
RQQ - HINCHINGBROOKE HEALTH CARE NHS TRUST	10	3,241	28	3,030	13	2,534	36	3,320	35	4,314	26	3,519	9	3,023			157	22,981
RQW - THE PRINCESS ALEXANDRA HOSPITAL NHS TRUST	48	3,927	86	4,834	23	4,657	39	5,521	21	6,409	20	7,391	24	7,017	0	3	261	39,759
RQX - HOMERTON UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	38	3,418	45	4,118	40	4,714	38	6,377	23	5,891	24	5,846	14	4,368			222	34,732
RR1 - HEART OF ENGLAND NHS FOUNDATION TRUST	186	12,952	173	12,790	238	15,026	161	15,427	176	14,689	177	15,428	98	14,806	0	1	1,209	101,119
RR7 - GATESHEAD HEALTH NHS FOUNDATION TRUST	37	4,907	30	4,924	28	4,765	40	4,509	35	5,062	40	5,408	34	5,061	0	1	244	34,637
RR8 - LEEDS TEACHING HOSPITALS NHS TRUST	49	15,659	74	17,395	34	19,882	62	19,603	54	20,469	48	22,386	23	19,012	0	4	344	134,410
RRF - WRIGHTINGTON, WIGAN AND LEIGH NHS FOUNDATION TRUST	51	3,450	28	2,539	22	2,381	33	2,850	33	5,906	82	7,444	58	7,057	1	11	308	31,638
RRK - UNIVERSITY HOSPITALS BIRMINGHAM NHS FOUNDATION TRUST	49	8,940	113	8,591	88	9,620	24	10,152	24	16,750	20	21,365	21	21,293	0	20	339	96,731
RRV - UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	54	5,604	55	6,218	38	7,319	44	8,630	35	9,765	55	9,837	20	7,869	0	8	301	55,250
RTD - THE NEWCASTLE UPON TYNE HOSPITALS NHS FOUNDATION TRUST	61	7,971	70	8,767	55	10,054	40	11,966	72	14,900	81	15,861	37	11,002			416	80,521

Organisation	Year																	
	2010/11 Death or Severe	Total Incidents	2011/12 Death or Severe	Total Incidents	2012/13 Death or Severe	Total Incidents	2013/14 Death or Severe	Total Incidents	2014/15 Death or Severe	Total Incidents	2015/16 Death or Severe	Total Incidents	2016/17 Death or Severe	Total Incidents	Incorrect Death or Severe	Total Incidents	Total Death or Severe	Total Incidents
RTE - GLOUCESTERSHIRE HOSPITALS NHS FOUNDATION TRUST	34	10,619	77	10,309	165	9,533	125	8,423	67	9,932	39	11,360	77	12,784	0	1	584	72,961
RTF - NORTHUMBRIA HEALTHCARE NHS FOUNDATION TRUST	88	8,368	79	9,182	98	10,076	54	11,044	19	11,421	38	12,102	42	10,161	0	24	418	72,378
RTG - UNIVERSITY HOSPITALS OF DERBY AND BURTON NHS FOUNDATION TRUST	16	10,946	18	13,406	12	12,367	30	11,120	33	12,596	26	11,238	30	10,574	0	11	165	82,258
RTH - OXFORD UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	177	10,251	168	11,610	48	8,657	50	16,706	47	17,887	50	18,001	11	15,761	1	8	552	98,881
RTK - ASHFORD AND ST PETER'S HOSPITALS NHS FOUNDATION TRUST	30	4,441	11	5,486	20	4,491	31	4,473	26	5,438	22	6,266	19	5,568			159	36,163
RTP - SURREY AND SUSSEX HEALTHCARE NHS TRUST	41	3,923	45	3,782	49	4,028	42	4,722	41	5,931	39	6,389	27	6,488	0	2	284	35,265
RTR - SOUTH TEES HOSPITALS NHS FOUNDATION TRUST	78	9,356	52	11,133	66	12,541	41	12,302	38	12,690	45	10,851	7	6,783			327	75,656
RTX - UNIVERSITY HOSPITALS OF MORECAMBE BAY NHS FOUNDATION TRUST	22	6,671	43	7,656	65	11,715	48	9,009	37	9,599	32	8,605	26	7,983	1	2	274	61,240
RV8 - NORTH WEST LONDON HOSPITALS NHS TRUST	59	4,688	65	4,636	51	5,109	37	6,761	49	8,033	10	846					271	30,073
RVJ - NORTH BRISTOL NHS TRUST	48	5,932	41	7,220	61	9,286	78	9,905	87	10,374	79	10,034	86	10,394			480	63,145
RVL - BARNET AND CHASE FARM HOSPITALS NHS TRUST	37	6,836	51	7,024	42	6,167	49	5,820	18	2,209							197	28,056
RVR - EPSOM AND ST HELIER UNIVERSITY HOSPITALS NHS TRUST	52	4,529	49	4,374	41	4,117	81	5,165	59	7,797	61	8,444	29	8,239	0	3	372	42,668
RVV - EAST KENT HOSPITALS UNIVERSITY NHS FOUNDATION TRUST	63	4,090	128	6,546	106	8,605	41	11,217	28	12,253	55	13,096	94	13,386	0	3	515	69,196
RVW - NORTH TEES AND HARTLEPOOL NHS FOUNDATION TRUST	36	3,881	53	4,639	47	5,250	30	5,961	20	6,292	11	5,980	12	5,518			209	37,521
RVY - SOUTHPORT AND ORMSKIRK HOSPITAL NHS TRUST	2	3,356	7	3,725	19	4,040	20	3,884	23	4,547	33	5,501	16	4,998	0	4	120	30,055
RW3 - CENTRAL MANCHESTER UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	135	8,449	25	13,696	63	21,848	47	22,846	56	24,944	76	23,782	95	22,541	1	6	498	138,112
RW6 - PENNINE ACUTE HOSPITALS NHS TRUST	349	10,792	192	11,311	26	10,611	34	12,007	73	14,028	187	13,403	216	13,802	1	1	1,078	85,955
RWA - HULL AND EAST YORKSHIRE HOSPITALS NHS TRUST	71	11,879	38	13,740	20	11,353	51	11,541	60	11,008	58	12,134	39	10,937	0	16	337	82,608
RWD - UNITED LINCOLNSHIRE HOSPITALS NHS TRUST	87	10,307	108	10,319	122	10,878	152	10,635	170	9,178	207	9,347	179	10,109	0	11	1,025	70,784
RWE - UNIVERSITY HOSPITALS OF LEICESTER NHS TRUST	173	15,815	168	17,558	159	22,310	119	23,810	70	24,531	38	22,257	33	19,961	0	6	760	146,248
RWF - MAIDSTONE AND TUNBRIDGE WELLS NHS TRUST	58	4,595	38	4,668	60	5,461	65	5,567	60	5,883	75	6,624	80	6,243	0	5	436	39,046
RWG - WEST HERTFORDSHIRE HOSPITALS NHS TRUST	29	4,888	17	4,323	41	6,000	41	5,771	35	6,546	46	8,656	38	7,286	0	1	247	43,471
RWH - EAST AND NORTH HERTFORDSHIRE NHS TRUST	31	8,816	42	9,585	31	9,132	48	5,863	34	5,242	49	7,012	49	5,968	0	1	284	51,619
RWJ - STOCKPORT NHS FOUNDATION TRUST	122	6,845	72	7,279	81	8,603	145	9,210	189	10,986	91	10,752	74	6,591			774	60,266
RWP - WORCESTERSHIRE ACUTE HOSPITALS NHS TRUST	49	8,155	41	9,353	34	11,296	47	10,333	62	10,160	43	10,754	39	8,349	0	8	315	68,408
RWQ - WARRINGTON AND HALTON HOSPITALS NHS FOUNDATION TRUST	17	5,362	22	6,327	9	6,899	14	7,443	9	6,952	61	7,080	26	6,137	0	4	158	46,204
RWY - CALDERDALE AND HUDDERSFIELD NHS FOUNDATION TRUST	90	7,669	135	7,597	174	7,357	151	7,352	142	8,903	112	10,363	26	7,466	1	13	831	56,720
RX1 - NOTTINGHAM UNIVERSITY HOSPITALS NHS TRUST	49	12,745	46	16,145	104	18,549	76	22,873	50	23,047	47	20,651	33	17,093	0	90	405	131,193
RXC - EAST SUSSEX HEALTHCARE NHS TRUST	51	5,639	73	7,813	82	8,565	36	8,791	21	7,452	59	9,353	40	13,725	0	16	362	61,354
RXF - MID YORKSHIRE HOSPITALS NHS TRUST	25	3,865	34	8,746	75	8,662	39	10,188	73	13,359	63	15,057	53	15,143			362	75,020
RXH - BRIGHTON AND SUSSEX UNIVERSITY HOSPITALS NHS TRUST	20	6,731	22	6,747	12	6,872	15	8,129	21	8,631	19	8,980	22	10,432	0	27	131	56,549
RXK - SANDWELL AND WEST BIRMINGHAM HOSPITALS NHS TRUST	81	5,812	107	7,861	38	11,261	46	13,943	43	14,312	58	11,985	16	10,037			389	75,211
RXL - BLACKPOOL TEACHING HOSPITALS NHS FOUNDATION TRUST	24	6,114	12	6,145	18	9,743	39	10,873	30	11,742	17	14,122	7	14,741			147	73,480
RXN - LANCASHIRE TEACHING HOSPITALS NHS FOUNDATION TRUST	37	5,146	30	6,690	48	9,988	26	11,044	56	12,889	83	12,747	99	11,415	0	5	379	69,924
RXP - COUNTY DURHAM AND DARLINGTON NHS FOUNDATION TRUST	17	6,876	21	7,410	22	8,543	28	9,381	20	10,417	55	11,538	37	9,558			200	63,723
RXQ - BUCKINGHAMSHIRE HEALTHCARE NHS TRUST	129	6,394	115	7,241	63	7,760	56	8,233	63	8,615	45	8,432	8	6,553	0	3	479	53,231
RXR - EAST LANCASHIRE HOSPITALS NHS TRUST	5	8,077	3	10,122	3	11,634	53	15,227	66	15,850	48	13,440	31	9,789	0	17	209	84,156
RXW - SHREWSBURY AND Telford HOSPITAL NHS TRUST	40	7,406	76	7,724	61	8,098	34	7,590	53	7,651	40	7,651	24	5,459	1	3	329	51,582
RYJ - IMPERIAL COLLEGE HEALTHCARE NHS TRUST	42	9,039	37	11,121	31	11,587	44	12,571	40	14,861	37	14,998	31	15,350	0	9	262	89,536
RYQ - SOUTH LONDON HEALTHCARE NHS TRUST	121	8,406	130	9,605	64	7,904	45	4,064									360	29,979
RYR - WESTERN SUSSEX HOSPITALS NHS FOUNDATION TRUST	28	7,493	17	8,008	11	8,349	19	8,860	24	8,355	41	8,082	38	8,301	1	1	179	57,449
NHS Trusts (ambulance)	111	4,927	56	5,445	116	6,209	206	7,652	331	11,280	306	15,857	228	10,018	0	30	1,354	61,418
RRU - LONDON AMBULANCE SERVICE NHS TRUST	2	393	7	731	14	1,454	4	441	0	731	9	1,689	64	492	0	4	100	5,935
RT4 - WELSH AMBULANCE SERVICES NHS TRUST	1	618	0	544	1	796	10	956	10	806	12	1,072	4	497			38	5,289
RX6 - NORTH EAST AMBULANCE SERVICE NHS FOUNDATION TRUST	1	48	7	546	21	615	54	1,184	70	1,256	53	2,098	45	1,179	0	26	251	6,952
RX7 - NORTH WEST AMBULANCE SERVICE NHS TRUST	21	869	2	454	9	173	8	715	15	1,771	11	967	11	1,211			77	6,160

Organisation	2010/11		2011/12		2012/13		2013/14		Year 2014/15		2015/16		2016/17		Incorrect		Total	
	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents
RX8 - YORKSHIRE AMBULANCE SERVICE NHS TRUST	10	580	6	586	18	367	23	184	60	1,801	66	1,659	59	1,548			242	6,725
RX9 - EAST MIDLANDS AMBULANCE SERVICE NHS TRUST	31	415	2	199	1	171	3	490	8	567	38	718	13	870			96	3,430
RYA - WEST MIDLANDS AMBULANCE SERVICE NHS FOUNDATION TRUST	5	451	2	540	16	624	10	585	12	588	7	752	6	1,128			58	4,668
RYC - EAST OF ENGLAND AMBULANCE SERVICE NHS TRUST	8	339	2	540	9	224	7	326	5	580	2	1,787	1	1,292			34	5,088
RYD - SOUTH EAST COAST AMBULANCE SERVICE NHS FOUNDATION TRUST	7	227	6	215	11	498	19	628	34	577	34	528	8	167			119	2,840
RYE - SOUTH CENTRAL AMBULANCE SERVICE NHS FOUNDATION TRUST	9	148	15	218	7	249	46	674	94	1,069	31	849	9	139			211	3,346
RYF - SOUTH WESTERN AMBULANCE SERVICE NHS FOUNDATION TRUST	16	839	7	872	9	1,038	22	1,469	23	1,534	43	3,738	8	1,495			128	10,985
NHS Trusts (community)	36	4,605	422	44,833	549	59,179	655	68,307	736	72,712	653	68,235	457	58,639	0	24	3,508	376,534
AXG - WILTSHIRE HEALTH & CARE											0	8	14	677			14	685
R1A - WORCESTERSHIRE HEALTH AND CARE NHS TRUST	0	6	22	1,482	34	1,325	158	4,155	88	5,985	129	5,775	86	4,878			517	23,606
R1D - SHROPSHIRE COMMUNITY HEALTH NHS TRUST			19	1,298	62	2,145	42	1,829	36	1,727	25	1,810	22	1,202			206	10,011
R1E - STAFFORDSHIRE AND STOKE ON TRENT PARTNERSHIP NHS TRUST	0	6	16	1,385	38	3,133	20	3,349	25	6,139	58	10,544	42	9,567			199	34,123
R1G - TORBAY AND SOUTHERN DEVON HEALTH AND CARE NHS TRUST	0	3	2	100	10	2,183	10	3,291	1	2,919	2	998					25	9,494
R1J - GLOUCESTERSHIRE CARE SERVICES NHS TRUST					4	69	25	2,760	12	1,846	11	2,486	6	1,750			58	8,911
RDR - SUSSEX COMMUNITY NHS FOUNDATION TRUST	6	706	23	2,830	11	3,791	20	4,684	23	4,265	31	2,937	3	2,405	0	2	117	21,620
RY1 - LIVERPOOL COMMUNITY HEALTH NHS TRUST			1	622	2	321	35	1,479	131	3,968	69	4,069	43	2,909	0	2	281	13,370
RY2 - BRIDGEWATER COMMUNITY HEALTHCARE NHS FOUNDATION TRUST	0	337	5	1,192	7	1,627	8	1,091	14	1,291	4	1,250	11	1,090			49	7,878
RY3 - NORFOLK COMMUNITY HEALTH AND CARE NHS TRUST	1	26	94	4,745	77	6,953	73	6,818	49	6,359	57	5,585	56	4,825	0	2	407	35,313
RY4 - HERTFORDSHIRE COMMUNITY NHS TRUST	0	386	30	4,338	24	4,502	31	4,543	9	4,586	10	4,658	7	4,486	0	2	111	27,501
RY5 - LINCOLNSHIRE COMMUNITY HEALTH SERVICES NHS TRUST			1	1,916	0	1,776	0	2,101	2	2,614	1	3,320	0	2,792			4	14,519
RY6 - LEEDS COMMUNITY HEALTHCARE NHS TRUST	5	348	32	2,383	17	2,336	28	3,193	31	3,670	60	4,108	14	3,555	0	3	187	19,596
RY7 - WIRRAL COMMUNITY NHS FOUNDATION TRUST	0	39	5	1,088	13	1,631	9	2,307	17	2,110	27	2,725	26	2,327	0	2	97	12,229
RY8 - DERBYSHIRE COMMUNITY HEALTH SERVICES NHS FOUNDATION TRUST	0	3	26	5,274	11	6,778	14	8,185	4	6,471	14	6,860	4	6,412			73	39,983
RY9 - HOUNSLOW AND RICHMOND COMMUNITY HEALTHCARE NHS TRUST			0	415	16	565	18	1,529	10	1,782	31	2,063	16	1,871	0	1	91	8,226
RYV - CAMBRIDGESHIRE COMMUNITY SERVICES NHS TRUST	15	802	80	4,475	97	5,732	38	3,317	17	3,426	1	1,082	3	1,083	0	7	251	19,924
RYW - BIRMINGHAM COMMUNITY HEALTHCARE NHS FOUNDATION TRUST	0	44	18	2,995	42	4,372	23	4,757	43	4,677	13	3,733	8	3,076	0	2	147	23,656
RYX - CENTRAL LONDON COMMUNITY HEALTHCARE NHS TRUST	7	906	35	3,321	74	3,944	95	4,799	206	3,854	101	2,401	90	2,464	0	1	608	21,690
RYY - KENT COMMUNITY HEALTH NHS FOUNDATION TRUST	2	993	13	4,974	10	5,996	8	4,120	18	5,023	9	1,823	6	1,270			66	24,199
NHS Trusts (mental health/learning disability)	1,634	177,917	2,419	222,030	3,023	232,273	3,096	257,220	3,114	276,586	3,696	295,149	3,737	285,572	6	62	20,725	1,746,809
R1C - SOLENT NHS TRUST	4	194	6	2,459	12	2,516	25	2,945	39	3,277	69	2,143	172	2,776			327	16,310
RAT - NORTH EAST LONDON NHS FOUNDATION TRUST	27	1,343	76	3,828	62	4,906	36	5,389	21	5,740	43	7,232	47	7,529	0	1	312	35,968
RDY - DORSET HEALTHCARE UNIVERSITY NHS FOUNDATION TRUST	16	3,239	122	5,578	55	5,483	94	5,781	60	5,652	53	5,809	34	5,734	1	1	435	37,277
RGD - LEEDS AND YORK PARTNERSHIP NHS FOUNDATION TRUST	1	5,149	14	5,242	61	6,308	25	6,186	32	6,286	29	5,760	24	4,640	0	1	186	39,572
RHS - SOMERSET PARTNERSHIP NHS FOUNDATION TRUST	26	1,021	18	2,144	31	2,992	34	2,654	26	3,157	69	3,024	72	2,140			276	17,132
RHA - NOTTINGHAMSHIRE HEALTHCARE NHS FOUNDATION TRUST	58	5,426	72	7,698	41	8,939	39	9,360	67	10,385	57	10,998	47	11,601			381	64,407
RHX - OXFORDSHIRE LEARNING DISABILITY NHS TRUST	0	1,960	1	2,477	1	1,066											2	5,503
RJB - CORNWALL PARTNERSHIP NHS FOUNDATION TRUST	20	966	15	1,146	24	1,179	34	1,668	37	2,342	67	2,862	69	5,669			266	15,832
RJX - CALDERSTONES PARTNERSHIP NHS FOUNDATION TRUST	0	1,773	5	1,701	3	1,560	7	1,501	2	1,782	2	2,140	1	453			20	10,910
RKL - WEST LONDON NHS TRUST	55	2,002	18	2,035	31	2,034	24	2,021	17	2,078	39	3,476	26	3,165			210	16,811
RLY - NORTH STAFFORDSHIRE COMBINED HEALTHCARE NHS TRUST	16	888	39	1,509	71	2,537	60	2,191	40	2,506	44	2,605	45	2,485			315	14,721
RMV - NORFOLK AND SUFFOLK NHS FOUNDATION TRUST	4	3,805	10	4,034	42	6,031	49	7,553	47	8,618	64	9,275	47	7,801			263	47,117
RNK - TAVISTOCK AND PORTMAN NHS FOUNDATION TRUST	0	61	0	69	1	27	0	41	0	12	2	37	4	97			7	344
RNN - CUMBRIA PARTNERSHIP NHS FOUNDATION TRUST	9	1,587	52	4,790	98	5,963	69	4,833	145	4,257	74	3,713	46	2,469			493	27,612
RNU - OXFORD HEALTH NHS FOUNDATION TRUST	58	2,447	150	5,375	109	6,807	85	7,833	62	7,635	61	7,713	134	6,071			659	43,881
RP1 - NORTHAMPTONSHIRE HEALTHCARE NHS FOUNDATION TRUST	27	2,468	19	2,629	54	2,795	43	3,497	38	3,748	14	4,004	34	4,186	0	3	229	23,350
RP7 - LINCOLNSHIRE PARTNERSHIP NHS FOUNDATION TRUST	20	1,576	41	1,507	47	1,633	55	1,701	34	2,174	47	2,365	11	1,655	0	1	255	12,612

Organisation	Year																	
	2010/11		2011/12		2012/13		2013/14		2014/15		2015/16		2016/17		Incorrect		Total	
	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents
RPG - OXLEAS NHS FOUNDATION TRUST	23	2,203	25	2,939	50	4,859	54	7,656	64	8,032	84	7,665	75	7,823			375	41,177
ROY - SOUTH WEST LONDON AND ST GEORGE'S MENTAL HEALTH NHS TRUST	9	1,446	14	1,306	50	2,518	29	3,542	28	3,891	64	4,614	100	4,282	0	9	294	21,608
RRD - NORTH ESSEX PARTNERSHIP UNIVERSITY NHS FOUNDATION TRUST	24	2,780	7	2,582	8	2,444	2	2,436	10	2,424	4	1,912	15	1,226	0	1	70	15,805
RRE - MIDLANDS PARTNERSHIP NHS FOUNDATION TRUST	50	2,962	154	2,642	87	2,484	73	3,506	60	3,114	64	2,755	213	4,204			701	21,667
RRP - BARNET, ENFIELD AND HARINGEY MENTAL HEALTH NHS TRUST	14	2,349	19	2,654	24	2,389	41	3,501	76	2,698	41	2,859	89	4,970			304	21,420
RT1 - CAMBRIDGESHIRE AND PETERBOROUGH NHS FOUNDATION TRUST	47	4,111	77	4,551	63	4,598	54	5,226	38	6,357	41	6,980	24	5,934			344	37,757
RT2 - PENNINE CARE NHS FOUNDATION TRUST	0	3,281	14	4,247	29	4,469	43	4,094	54	4,572	207	7,076	136	7,270			483	35,009
RT5 - LEICESTERSHIRE PARTNERSHIP NHS TRUST	12	2,893	11	5,849	19	6,117	18	7,517	46	8,509	48	9,518	35	9,839			189	50,242
RTQ - ZGETHER NHS FOUNDATION TRUST	2	1,349	6	2,532	28	3,227	26	3,242	22	2,550	10	2,641	57	4,347			151	19,888
RTV - NORTH WEST BOROUGHs HEALTHCARE NHS FOUNDATION TRUST	2	2,381	8	2,568	17	3,537	47	5,712	32	5,354	56	5,166	58	5,348			220	30,066
RV3 - CENTRAL AND NORTH WEST LONDON NHS FOUNDATION TRUST	19	3,999	40	5,122	29	5,750	44	7,684	48	8,963	21	7,754	40	9,002	0	4	241	48,278
RV5 - SOUTH LONDON AND MAUDSLEY NHS FOUNDATION TRUST	114	7,509	61	8,053	92	6,996	75	6,596	55	6,273	70	6,895	39	3,786	0	2	506	46,110
RV9 - HUMBER TEACHING NHS FOUNDATION TRUST	2	1,306	10	2,430	17	2,805	18	2,795	10	3,603	52	4,191	82	4,334			191	21,464
RVN - AVON AND WILTSHIRE MENTAL HEALTH PARTNERSHIP NHS TRUST	108	5,304	87	5,498	93	5,858	59	6,968	60	7,433	70	7,947	51	8,086			528	47,094
RW1 - SOUTHERN HEALTH NHS FOUNDATION TRUST	34	5,556	87	6,567	100	6,543	73	8,523	219	11,266	242	12,513	120	10,628	3		878	61,600
RW4 - MERSEY CARE NHS FOUNDATION TRUST	5	1,763	4	4,002	15	4,644	44	5,112	60	4,907	88	5,775	85	5,389	0	13	301	31,605
RW5 - LANCASHIRE CARE NHS FOUNDATION TRUST	22	8,407	38	10,682	63	9,094	66	10,362	48	10,333	49	10,525	152	9,291	2	6	440	68,700
RWK - EAST LONDON NHS FOUNDATION TRUST	78	1,915	50	2,378	11	2,661	145	4,424	66	2,325	194	6,855	59	4,866	0	1	603	25,425
RWN - SOUTH ESSEX PARTNERSHIP UNIVERSITY NHS FOUNDATION TRUST	15	3,403	117	5,438	125	7,546	75	7,468	90	9,409	39	7,619	32	6,089			493	46,972
RWR - HERTFORDSHIRE PARTNERSHIP UNIVERSITY NHS FOUNDATION TRUST	2	2,957	8	3,119	17	2,811	25	4,450	27	3,997	33	3,871	44	3,986			156	25,191
RWW - DEVON PARTNERSHIP NHS TRUST	24	1,619	83	1,619	78	2,026	83	1,614	46	1,936	30	2,382	81	3,017			425	14,213
RWX - BERKSHIRE HEALTHCARE NHS FOUNDATION TRUST	13	1,251	28	3,882	40	3,706	33	3,730	47	3,592	67	3,662	37	2,891			265	22,714
RX2 - SUSSEX PARTNERSHIP NHS FOUNDATION TRUST	8	3,431	17	3,237	44	2,978	104	2,856	80	3,253	124	3,809	123	4,005			500	23,569
RX3 - TEES, ESK AND WEAR VALLEYS NHS FOUNDATION TRUST	4	6,999	29	6,075	82	5,988	61	6,482	66	7,568	171	7,618	190	10,937			603	51,667
RX4 - NORTHUMBERLAND, TYNE AND WEAR NHS FOUNDATION TRUST	142	11,551	165	12,558	277	13,764	179	12,560	183	11,093	138	10,038	185	12,047	0	2	1,269	83,613
RXA - CHESHIRE AND WIRRAL PARTNERSHIP NHS FOUNDATION TRUST	29	1,924	16	2,311	58	3,774	100	3,197	122	2,104	178	6,138	223	4,834			726	24,282
RXE - ROTHERHAM DONCASTER AND SOUTH HUMBER NHS FOUNDATION TRUST	55	2,879	75	6,443	64	5,389	82	5,182	59	5,724	89	4,494	75	3,920			499	34,031
RXG - SOUTH WEST YORKSHIRE PARTNERSHIP NHS FOUNDATION TRUST	52	3,765	74	4,459	76	4,051	79	4,605	66	4,717	58	6,098	55	6,144	0	1	460	33,840
RXM - DERBYSHIRE HEALTHCARE NHS FOUNDATION TRUST	9	2,458	26	2,619	37	2,728	77	2,972	79	2,715	119	2,879	97	2,638			444	19,009
RXT - BIRMINGHAM AND SOLIHULL MENTAL HEALTH NHS FOUNDATION TRUST	55	8,921	71	8,905	122	6,745	95	9,030	82	10,782	70	10,200	57	8,234	0	1	552	62,818
RXV - GREATER MANCHESTER MENTAL HEALTH NHS FOUNDATION TRUST	20	4,708	15	4,288	49	3,831	35	3,682	37	4,913	40	6,530	36	8,524	0	5	232	36,481
RXX - SURREY AND BORDERS PARTNERSHIP NHS FOUNDATION TRUST	66	1,665	84	1,411	66	1,701	84	2,300	82	2,215	102	2,622	47	2,135	0	1	531	14,050
RXY - KENT AND MEDWAY NHS AND SOCIAL CARE PARTNERSHIP TRUST	21	4,669	13	3,923	71	3,855	53	3,828	82	3,427	110	3,302	67	3,500	0	1	417	26,505
RYG - COVENTRY AND WARWICKSHIRE PARTNERSHIP NHS TRUST	38	3,775	74	4,533	104	5,338	132	6,408	155	10,122	56	9,411	30	7,470			589	47,057
RYK - DUDLEY AND WALSALL MENTAL HEALTH PARTNERSHIP NHS TRUST	16	935	19	997	36	1,257	35	1,527	28	1,554	16	1,722	14	2,125			164	10,117
TAD - BRADFORD DISTRICT CARE NHS FOUNDATION TRUST	14	4,738	17	5,233	31	2,771	59	2,923	31	3,857	26	4,234	35	4,588			213	28,344
TAE - MANCHESTER MENTAL HEALTH AND SOCIAL CARE TRUST	52	1,861	19	1,535	11	860	24	1,892	26	2,669	20	2,492	8	2,430	0	1	160	13,740
TAF - CAMDEN AND ISLINGTON NHS FOUNDATION TRUST	43	1,455	27	1,127	41	1,285	43	1,409	24	2,268	21	2,840	9	2,251	0	2	208	12,637
TAH - SHEFFIELD HEALTH & SOCIAL CARE NHS FOUNDATION TRUST	28	3,404	42	3,633	44	3,408	36	3,164	25	4,585	27	4,116	30	3,337			232	25,647
TAJ - BLACK COUNTRY PARTNERSHIP NHS FOUNDATION TRUST	22	2,130	30	3,862	12	2,692	12	1,891	14	1,813	23	2,275	19	1,374	0	1	132	16,038
NHS Trusts (other)	43	7,369	32	8,513	9	5,507	6	2,029	2	5	0	7	0	10	3	3	95	23,443
RYH - NHS DIRECT NHS TRUST	43	6,944	30	8,201	9	5,445	2	2,023									84	22,613
RYT - PUBLIC HEALTH WALES NHS TRUST	0	425	2	312	0	62	4	6	2	5	0	7	0	10	3	3	11	830
NHS Trusts (specialist)	189	26,223	150	30,916	148	32,659	201	37,572	180	43,964	98	46,941	77	44,938	1	5	1,044	263,218
RAN - ROYAL NATIONAL ORTHOPAEDIC HOSPITAL NHS TRUST	12	492	12	734	11	806	29	923	37	873	4	663	3	705	0	1	108	5,197
RBB - ROYAL NATIONAL HOSPITAL FOR RHEUMATIC DISEASES NHS FOUNDATION TRUST	1	134	0	194	2	273	0	210	0	109							3	920

Organisation	Year																	
	2010/11		2011/12		2012/13		2013/14		2014/15		2015/16		2016/17		Incorrect		Total	
	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents	Death or Severe	Total Incidents
RBQ - LIVERPOOL HEART AND CHEST HOSPITAL NHS FOUNDATION TRUST	3	1,338	7	1,097	2	1,339	2	1,352	1	1,099	3	1,538	0	1,244			18	9,007
RBS - ALDER HEY CHILDREN'S NHS FOUNDATION TRUST	0	853	0	1,170	1	1,698	2	2,096	10	3,720	4	4,698	2	4,131			19	18,366
RBV - THE CHRISTIE NHS FOUNDATION TRUST	2	2,103	1	3,155	2	2,320	6	1,793	6	1,537	3	1,716	2	1,611			22	14,235
RCU - SHEFFIELD CHILDREN'S NHS FOUNDATION TRUST	0	2,020	10	2,207	0	2,468	13	2,715	1	3,151	1	2,749	0	3,324			25	18,634
REN - THE CLATTERBRIDGE CANCER CENTRE NHS FOUNDATION TRUST	0	1,087	0	1,017	0	1,338	0	1,531	0	1,861	0	2,467	0	1,905	0	1	0	11,207
REP - LIVERPOOL WOMEN'S NHS FOUNDATION TRUST	86	2,743	35	2,740	53	3,087	56	2,117	35	2,717	14	3,372	7	4,040	0	1	286	20,817
RET - THE WALTON CENTRE NHS FOUNDATION TRUST	7	818	2	564	0	533	1	627	2	865	2	1,011	6	999			20	5,417
RGM - ROYAL PAPWORTH HOSPITAL NHS FOUNDATION TRUST	4	1,120	4	1,508	8	1,969	8	1,834	1	2,359	6	2,192	5	1,886			36	12,868
RL1 - THE ROBERT JONES AND AGNES HUNT ORTHOPAEDIC HOSPITAL NHS FOUNDATION TRUST	5	968	14	1,022	12	1,272	6	1,415	3	1,445	3	1,615	3	1,373			46	9,110
RLU - BIRMINGHAM WOMEN'S NHS FOUNDATION TRUST	15	1,045	15	1,470	4	1,468	4	1,884	10	1,385	6	1,847	5	1,804			59	10,903
RP4 - GREAT ORMOND STREET HOSPITAL FOR CHILDREN NHS FOUNDATION TRUST	18	2,813	5	2,787	16	3,434	17	4,627	15	4,778	10	3,698	4	1,761	0	1	85	23,899
RP6 - MOORFIELDS EYE HOSPITAL NHS FOUNDATION TRUST	5	581	7	833	9	1,206	12	3,446	9	6,201	14	5,716	13	7,464	1	1	70	25,448
RPC - QUEEN VICTORIA HOSPITAL NHS FOUNDATION TRUST	3	467	0	619	0	638	2	888	3	794	1	874	3	645			12	4,925
RPY - THE ROYAL MARSDEN NHS FOUNDATION TRUST	7	1,598	7	3,535	3	2,854	3	3,383	2	3,379	4	4,004	2	3,969			28	22,722
RQ3 - BIRMINGHAM WOMEN'S AND CHILDREN'S NHS FOUNDATION TRUST	7	2,514	9	2,797	12	2,342	18	2,706	31	3,201	10	3,226	14	3,293			101	20,079
RQF - VELINDRE NHS TRUST	0	691	0	646	1	619	1	750	2	699	0	789	1	721			5	4,915
RRJ - THE ROYAL ORTHOPAEDIC HOSPITAL NHS FOUNDATION TRUST	6	695	21	596	7	845	17	889	8	929	10	948	6	714			75	5,616
RT3 - ROYAL BROMPTON & HAREFIELD NHS FOUNDATION TRUST	8	2,143	1	2,225	5	2,150	4	2,386	4	2,862	3	3,818	1	3,349			26	18,933
NULL	525	72,273	486	74,575	419	80,152	546	88,018	611	89,232	723	96,668	536	85,290	5	82	3,851	586,290
NULL	525	72,273	486	74,575	419	80,152	546	88,017	611	89,232	723	96,667	536	85,290	5	82	3,851	586,288
to be updated							0	1			0	1					0	2
Other	78	2,752	61	1,968	0	4	0	57	0	10	0	2					139	4,793
O2Y - NHS EAST RIDING OF YORKSHIRE CCG					0	4	0	57	0	10	0	2					0	73
RN1 - To be updated	78	2,752	61	1,968													139	4,720
Grand Total	11,047	1,250,249	11,450	1,366,042	11,407	1,469,558	11,276	1,602,708	11,627	1,766,163	12,118	1,844,282	10,349	1,664,554	25	958	79,299	10,964,514

C.1 NRLS reports submitting by organisation, grouped by fiscal year for incidents reported to NRLS 2010/11 - 2016/17

Location Levels (1/2/3)	Total number incidents	Incidents as a percentage of total
Ambulance (including call / control centre)	27,706	0.25%
Call / control centre	8,550	0.08%
NULL	8,550	0.08%
In vehicle / in transit	13,061	0.12%
NULL	13,061	0.12%
NHS Direct	13	0.00%
NULL	13	0.00%
NULL	15	0.00%
NULL	15	0.00%
Other	6,067	0.06%
NULL	6,067	0.06%
Community hospital	572,122	5.22%
Day care services	9,497	0.09%
NULL	9,495	0.09%
Other	2	0.00%
General areas	72,975	0.67%
Hospital buildings (inside)	46,045	0.42%
Hospital grounds (outside)	5,538	0.05%
NULL	18,305	0.17%
Other	3,087	0.03%
Inpatient areas	391,753	3.57%
NULL	47,364	0.43%
Other	6,901	0.06%
Ward	337,488	3.08%
NULL	507	0.00%
NULL	507	0.00%
Other	43,416	0.40%
NULL	43,416	0.40%
Outpatient department	26,054	0.24%
NULL	26,053	0.24%
Other	1	0.00%
Support Services	27,920	0.25%
Hospital transport (car)	253	0.00%
Laboratory	2,977	0.03%
NULL	9,541	0.09%
Other	4,297	0.04%
Pharmacy	3,036	0.03%
Therapy department	7,816	0.07%
General / acute hospital	7,671,154	69.96%
Accident (A) / minor injury unit / medical assessment unit	610,631	5.57%
NULL	610,586	5.57%
Other	45	0.00%
Ambulatory care treatment centre	14,548	0.13%
NULL	14,548	0.13%

Location Levels (1/2/3)	Total number incidents reported	Incidents as a percentage of total
Day care pre-assessment clinic	145	0.00%
NULL	145	0.00%
Day care services	122,102	1.11%
NULL	122,089	1.11%
Other	13	0.00%
General areas	286,145	2.61%
Hospital buildings (inside)	239,762	2.19%
Hospital grounds (outside)	15,371	0.14%
Mortuary	3,405	0.03%
NULL	4,932	0.04%
Other	22,675	0.21%
Inpatient areas	5,650,166	51.53%
Anaesthetic room	16,667	0.15%
Intensive care unit / high dependency unit	338,800	3.09%
NULL	55,757	0.51%
Operating theatre	356,103	3.25%
Other	53,449	0.49%
Recovery room	31,374	0.29%
Ward	4,798,016	43.76%
NULL	496	0.00%
NULL	496	0.00%
Other	83,928	0.77%
NULL	83,923	0.77%
Other	5	0.00%
Outpatient department	446,565	4.07%
NULL	446,409	4.07%
Other	143	0.00%
Therapy department	13	0.00%
Outpatient pre-assessment clinic	121	0.00%
NULL	121	0.00%
Support Services	456,307	4.16%
Hospital transport (car)	2,495	0.02%
Laboratory	153,141	1.40%
NULL	18,915	0.17%
Other	25,141	0.23%
Pharmacy	68,731	0.63%
Radiology	159,348	1.45%
Therapy department	28,536	0.26%
Mental health unit / facility	1,389,480	12.67%
Community mental health facility	136,662	1.25%
NULL	136,662	1.25%
Day care services	17,386	0.16%
NULL	17,386	0.16%

Location Levels (1/2/3)	Total number incidents reported	Incidents as a percentage of total
General areas	105,854	0.97%
Hospital buildings (inside)	74,711	0.68%
Hospital grounds (outside)	14,627	0.13%
NULL	12,904	0.12%
Other	3,612	0.03%
Inpatient areas	1,082,612	9.87%
ECT Suite	965	0.01%
Intensive care unit / high dependency unit	28,105	0.26%
NULL	38,043	0.35%
Other	12,190	0.11%
Secure unit	114,076	1.04%
Ward	889,233	8.11%
NULL	1,269	0.01%
NULL	1,269	0.01%
Other	22,871	0.21%
NULL	22,871	0.21%
Outpatient department	10,263	0.09%
NULL	10,263	0.09%
Support Services	12,563	0.11%
Hospital transport	988	0.01%
NULL	2,306	0.02%
Other	2,917	0.03%
Pharmacy	6,352	0.06%
Not applicable	18,330	0.17%
NULL	18,330	0.17%
NULL	18,330	0.17%
NULL	4	0.00%
NULL	4	0.00%
NULL	4	0.00%
Other	151,474	1.38%
NULL	151,472	1.38%
NULL	151,472	1.38%
Other	2	0.00%
NULL	2	0.00%
Primary care setting	339,863	3.10%
Ambulatory care treatment centre	2,164	0.02%
NULL	2,164	0.02%
Community pharmacy	84,891	0.77%
NULL	84,891	0.77%
Dental surgery	20,951	0.19%
NULL	3,219	0.03%
Other	12,779	0.12%
Treatment / consulting room	4,382	0.04%
Waiting room / reception	571	0.01%

Location Levels (1/2/3)	Total number incidents reported	Incidents as a percentage of total
GP Surgery	37,136	0.34%
Dispensary	1,070	0.01%
NULL	23,126	0.21%
Other	4,142	0.04%
Treatment / consulting room	5,604	0.05%
Waiting room / reception	3,194	0.03%
Health centre / out-of-hours centre	125,370	1.14%
NULL	125,365	1.14%
Other	5	0.00%
NHS Direct	22,736	0.21%
NULL	22,736	0.21%
NULL	29	0.00%
NULL	29	0.00%
Optician / optometrist	3,650	0.03%
Dispensing area	2	0.00%
NULL	165	0.00%
Other	622	0.01%
Treatment / consulting room	2,527	0.02%
Waiting room / reception	334	0.00%
Other	34,232	0.31%
NULL	34,232	0.31%
Rehabilitation centre	8,704	0.08%
NULL	8,704	0.08%
Public place (specify)	37,035	0.34%
NULL	37,035	0.34%
NULL	37,035	0.34%
Residence / home	639,884	5.84%
Hospice	7,962	0.07%
NULL	7,962	0.07%
Intermediate care setting	27,570	0.25%
NULL	27,570	0.25%
NULL	290	0.00%
NULL	290	0.00%
Nursing home	62,813	0.57%
NULL	62,810	0.57%
Other	3	0.00%
Other	15,481	0.14%
NULL	15,481	0.14%
Prison / remand centre	35,914	0.33%
NULL	35,914	0.33%
Private house / flat etc.	489,854	4.47%
NULL	489,818	4.47%
Other	36	0.00%

Location Levels (1/2/3)	Total number incidents reported	Incidents as a percentage of total
Social care facility	96,141	0.88%
Day care services	1,907	0.02%
NULL	1,907	0.02%
Local Authority (non-residential)	558	0.01%
NULL	558	0.01%
NULL	126	0.00%
NULL	126	0.00%
Other	7,152	0.07%
NULL	7,152	0.07%
Residential care home	86,398	0.79%
Unknown	21,321	0.19%
Grand Total	10,964,514	100.00%

C.2 NRLS levels 1,2 and 3 incident locations for incidents reported to NRLS 2015/16-2016/17

Specialty Levels 1/2)	Incidents	% of total
Accident and Emergency (A)	672,352	6.13%
NULL	672,352	6.13%
Anaesthesia Pain Management and Critical Care	149,591	1.36%
Anaesthesia	1,645	0.02%
Critical Care	8,974	0.08%
NULL	204	0.00%
Other	132,025	1.20%
Pain service	6,743	0.06%
Children's Specialties	12,026	0.11%
Critical Care	1,351	0.01%
Medical specialties	1,161	0.01%
Neonatology	3,557	0.03%
Other	1,059	0.01%
Paediatrics (not specified)	3,184	0.03%
Surgical specialties	1,714	0.02%
Dentistry - General and Community	12,695	0.12%
Endodontics	966	0.01%
NULL	10	0.00%
Oral surgery	1,257	0.01%
Orthodontics	1,478	0.01%
Other	7,850	0.07%
Paedodontics	606	0.01%
Periodontics	11	0.00%
Restorative dentistry	517	0.00%
Diagnostic services	400,177	3.65%
Blood transfusion	16,912	0.15%
Chemical pathology	37,451	0.34%
Haematology	51,515	0.47%
Histopathology	24,810	0.23%
Immunopathology	3,215	0.03%
Microbiology	20,249	0.18%
Neuropathology	2,067	0.02%
NULL	149	0.00%
Other	69,637	0.64%
Radiology	172,153	1.57%
Virology	2,019	0.02%
Learning disabilities	226,858	2.07%
Community teams	33,217	0.30%
Day care	6,537	0.06%
Forensic	28,693	0.26%
Inpatient assessment and treatment	81,970	0.75%
NULL	27	0.00%
Other	24,684	0.23%
Residential care	24,596	0.22%
Respite care	8,150	0.07%
Supported living	17,964	0.16%
(blank)	1,020	0.01%

Specialty Levels 1/2)	Incidents	% of total
Medical specialties	3,369,003	30.73%
Audiological medicine	6,944	0.06%
Cardiology	228,760	2.09%
Care of older people	681,071	6.21%
Clinical cytogenetics and molecular genetics	1,258	0.01%
Clinical haematology	47,157	0.43%
Clinical immunology and allergy	1,826	0.02%
Clinical oncology (previously radiotherapy)	81,144	0.74%
Dental medicine	2,875	0.03%
Dermatology	18,779	0.17%
Endocrinology	84,165	0.77%
Gastroenterology	192,124	1.75%
General medicine	896,479	8.18%
Genetics	3,309	0.03%
Genito-urinary medicine	17,027	0.16%
Infectious diseases	18,106	0.17%
Medical oncology	122,940	1.12%
Medical ophthalmology	18,439	0.17%
Neonatology	106,613	0.97%
Nephrology / renal	111,715	1.02%
Neurology	82,626	0.75%
Nuclear medicine	4,608	0.04%
NULL	451	0.00%
Other	195,600	1.78%
Palliative medicine	24,309	0.22%
Rehabilitation	211,296	1.93%
Rheumatology	20,696	0.19%
Thoracic / respiratory medicine	188,686	1.72%
Mental health	1,380,566	12.59%
Adult mental health	666,439	6.08%
Child and adolescent mental health	107,362	0.98%
Drug and alcohol service	19,540	0.18%
Forensic mental health	149,470	1.36%
Mental health rehabilitation	37,699	0.34%
NULL	197	0.00%
Older adult mental health	348,265	3.18%
Other	51,594	0.47%
Not applicable	106,360	0.97%
NULL	106,360	0.97%
NULL	79,621	0.73%
NULL	79,621	0.73%
Obstetrics and gynaecology	998,749	9.11%
Community midwifery	29,348	0.27%
Fertility treatment	3,245	0.03%
Gynaecology	131,188	1.20%
NULL	60	0.00%
Obstetrics	817,684	7.46%
Other	17,224	0.16%

Specialty Levels 1/2)	Incidents	% of total
Other	647,546	5.91%
General medicine	1	0.00%
Neonatology	2	0.00%
NULL	647,542	5.91%
Other	1	0.00%
Other specialties	296,734	2.71%
NULL	1	0.00%
Nutrition and dietetics	8,931	0.08%
Occupational therapy	16,086	0.15%
Other	113,363	1.03%
Pharmacy (inpatient)	108,692	0.99%
Physiotherapy	40,646	0.37%
Speech and language therapy	9,015	0.08%
Primary care / Community	891,989	8.14%
Chiropody / podiatry	11,847	0.11%
Community medicine	36,821	0.34%
Community midwifery	8,553	0.08%
Community nursing	575,911	5.25%
Community paediatrics	14,507	0.13%
General practice - no specialism	32,211	0.29%
General practice - with specialism relevant to this pat	1,185	0.01%
Health visiting / school nursing	38,039	0.35%
Intermediate care	77,953	0.71%
NULL	102	0.00%
Other	82,062	0.75%
Sexual health / family planning	12,798	0.12%
PTS (Patient Transport Service)	38,018	0.35%
NULL	38,018	0.35%
Surgical specialties	1,599,356	14.59%
Breast surgery	11,152	0.10%
Burns surgery	6,272	0.06%
Cardiac surgery	45,346	0.41%
Colorectal surgery	33,092	0.30%
Dental surgery	3,350	0.03%
ENT	48,164	0.44%
General surgery	432,459	3.94%
Maxillofacial / oral surgery	16,563	0.15%
Neurosurgery	56,566	0.52%
NULL	69	0.00%
Ophthalmology	75,127	0.69%
Orthodontics	2,049	0.02%
Other	233,635	2.13%
Paedodontics	1,218	0.01%
Plastic surgery	23,369	0.21%
Renal surgery	8,265	0.08%
Thoracic surgery	18,998	0.17%
Trauma and orthopaedics	451,390	4.12%
Urology	77,791	0.71%
Vascular surgery	54,481	0.50%
Unknown	82,873	0.76%
NULL	82,873	0.76%
Grand Total	10,964,514	100.00%

C.3 NRLS levels 1 and 2 incident specialties for incidents reported to NRLS 2015/16-2016/17

Incident type (Level 1/2)	Incidents	% of total
Access, admission, transfer, discharge (including missing patient)	990,024	9.029%
Absconder / missing patient	166,681	1.520%
Access / admission - delay / failure in access to hospital / care	87,589	0.799%
Access / admission - unexpected readmission / reattendance	57,291	0.523%
Access / admission - unplanned admission / transfer to specialist care unit	92,601	0.845%
Delay / difficulty in obtaining clinical assistance	2	0.000%
Discharge - delay / failure	42,586	0.388%
Discharge - inappropriate	32,634	0.298%
Discharge - planning failure	51,942	0.474%
Discharge - self or against medical advice	20,868	0.190%
Documentation / missing / inadequate / wrong / illegible healthcare record / card	3	0.000%
Failure in referral process	49,024	0.447%
Failure to return from authorised leave	29,556	0.270%
NULL	23	0.000%
Other	155,145	1.415%
Transfer / delay / failure / inappropriate	112,641	1.027%
Transport - delay / failure	46,915	0.428%
Unsafe / inappropriate clinical environment (including clinical waste)	44,523	0.406%
Clinical assessment (including diagnosis, scans, tests, assessments)	543,290	4.955%
Assessment - lack of clinical or risk assessment	59,624	0.544%
Cross-matching error	4,670	0.043%
Diagnosis - delay / failure to	72,171	0.658%
Diagnosis - wrong	11,568	0.106%
Documentation / missing / inadequate / wrong / illegible healthcare record / card	3	0.000%
NULL	44	0.000%
Other	66,693	0.608%
Patient incorrectly identified	2	0.000%
Scans / X-rays / specimens - inadequate / incomplete	39,160	0.357%
Scans / X-rays / specimens - mislabelled / unlabelled	81,177	0.740%
Scans / X-rays / specimens - missing	25,662	0.234%
Scans / X-rays / specimens - wrong	10,269	0.094%
Test results / reports - failure / delay to interpret or act on	24,379	0.222%
Test results / reports - failure / delay to receive	57,629	0.526%
Test results / reports - incorrect	33,698	0.307%
Test results / reports - missing	8,923	0.081%
Tests - failure / delay to undertake	47,618	0.434%
Consent, communication, confidentiality	400,101	3.649%
Breach of patient confidentiality	67,364	0.614%
Communication failure - outside of immediate team	101,347	0.924%
Communication failure - with patient / parent / carer	54,853	0.500%
Communication failure - within team	76,955	0.702%
Delay / difficulty in obtaining clinical assistance	2	0.000%
Documentation / missing / inadequate / wrong / illegible healthcare record / card	2	0.000%
Failure to receive informed consent (includes doctrine of necessity)	16,767	0.153%
NULL	94	0.001%
Other	82,717	0.754%
Disruptive, aggressive behaviour (includes patient-to-patient)	348,538	3.179%
NULL	41	0.000%
Other	96,608	0.881%
Physical	186,816	1.704%
Racial	2,805	0.026%
Sexual	11,385	0.104%
Verbal	50,883	0.464%

Incident type (Level 1/2)	Incidents	% of total
Documentation (including electronic & paper records, identification and drug charts)	690,025	6.293%
Appointment recording error	39,600	0.361%
Documentation - delay in obtaining healthcare record / card	41,572	0.379%
Documentation - healthcare record / card - mislabelled	13,097	0.119%
Documentation - misfiled	65,829	0.600%
Documentation - no access to	38,727	0.353%
Documentation / missing / inadequate / wrong / illegible healthcare record / card	218,722	1.995%
Documentation / missing / inadequate / wrong / illegible referral letter	11,687	0.107%
NULL	26	0.000%
Other	131,724	1.201%
Patient incorrectly identified	96,591	0.881%
Test request form - none / incomplete	20,825	0.190%
Test results / reports - mislabelled	11,625	0.106%
Implementation of care and ongoing monitoring / review	1,257,205	11.466%
Delay / difficulty in obtaining clinical assistance	48,537	0.443%
Delay / failure in recognising complication of treatment	28,893	0.264%
Delay or failure to monitor	345,512	3.151%
Failure to discontinue treatment	1	0.000%
Failure to follow up missed appointment	22,769	0.208%
NULL	70	0.001%
Other	811,422	7.400%
Patient incorrectly identified	1	0.000%
Infection Control Incident	194,714	1.776%
Diagnosis - delay / failure to	2,675	0.024%
Diagnosis - wrong	132	0.001%
Failure of sterilisation or contamination of equipment	16,634	0.152%
Infection - cross / healthcare associated	56,128	0.512%
Infection - wound	22,297	0.203%
NULL	78	0.001%
Other	59,969	0.547%
Test results / reports - failure / delay to interpret or act on	1,068	0.010%
Test results / reports - failure / delay to receive	286	0.003%
Test results / reports - incorrect	52	0.000%
Test results / reports - missing	138	0.001%
Tests - failure / delay to undertake	781	0.007%
Treatment / procedure - delay / failure	4,170	0.038%
Treatment / procedure - inappropriate	4,678	0.043%
Unsafe / inappropriate clinical environment (including clinical waste)	25,628	0.234%
Infrastructure (including staffing, facilities, environment)	621,828	5.671%
Exposure to cold / heat (includes fire)	18	0.000%
Failure / delay in collection / delivery systems	23,638	0.216%
Inadequate check on equipment / supplies	10,420	0.095%
IT / telecommunications failure / overload	16,077	0.147%
Lack of / delayed availability of beds (general)	91,943	0.839%
Lack of / delayed availability of beds (high dependency / intensive care)	13,803	0.126%
Lack of / delayed availability of operating theatre	4,522	0.041%
Lack of suitably trained / skilled staff	336,188	3.066%
NULL	59	0.001%
Other	58,803	0.536%
Unsafe / inappropriate clinical environment (including clinical waste)	36,669	0.334%
Unsafe environment (light, temperature, noise, air quality) - personal safety	29,688	0.271%
Medical device / equipment	307,573	2.805%
Failure of device / equipment	119,343	1.088%
Lack / unavailability of device / equipment	103,889	0.948%
NULL	150	0.001%
Other	44,475	0.406%
User error	32,340	0.295%
Wrong device / equipment used	7,376	0.067%

Incident type (Level 1/2)	Incidents	% of total
Medication	1,230,715	11.225%
Delay or failure to monitor	1	0.000%
NULL	1,230,712	11.225%
Other	2	0.000%
NULL	1	0.000%
NULL	1	0.000%
Other	426,007	3.885%
Other	426,007	3.885%
Patient abuse (by staff / third party)	57,916	0.528%
NULL	273	0.002%
Other	34,359	0.313%
Physical	15,089	0.138%
Racial	149	0.001%
Sexual	2,117	0.019%
Verbal	5,929	0.054%
Patient accident	2,357,183	21.498%
Ambulance / patient in road traffic accident	1,390	0.013%
Collision / contact with an object	81,258	0.741%
Contact with sharps (includes needle stick)	21,718	0.198%
Exposure to cold / heat (includes fire)	22,857	0.208%
Exposure to hazardous substance	16,738	0.153%
Inappropriate patient handling / positioning	9,250	0.084%
NULL	97	0.001%
Other	193,294	1.763%
Slips, trips, falls	2,010,581	18.337%
Pressure Ulcer	24	0.000%
NULL	24	0.000%
Self-harming behaviour	406,280	3.705%
Other	39,129	0.357%
Self-harm	337,036	3.074%
Suspected suicide (actual)	6,939	0.063%
Suspected suicide (attempted)	23,176	0.211%
Treatment, procedure	1,133,090	10.334%
Delay / difficulty in obtaining clinical assistance	19,017	0.173%
Delay / failure in recognising complication of treatment	1	0.000%
Extended stay / episode of care	51,761	0.472%
Failure to discontinue treatment	2,287	0.021%
Inappropriate patient handling / positioning	14,957	0.136%
Infusion injury (extravasation)	21,905	0.200%
Missing needle / swab / instrument	8,413	0.077%
NULL	84	0.001%
Other	602,827	5.498%
Patient incorrectly identified	1	0.000%
Restraint	4,809	0.044%
Retained needle / swab / instrument	4,598	0.042%
Theatre list details incorrect	13,020	0.119%
Transfer / delay / failure / inappropriate	1	0.000%
Treatment / procedure - delay / failure	243,641	2.222%
Treatment / procedure - inappropriate / wrong	126,137	1.150%
Treatment not clinically indicated	7,121	0.065%
Unplanned return to theatre	12,510	0.114%
Grand Total	10,964,514	100.000%

C.4 NRLS levels 1 and 2 incident type for incidents reported to NRLS 2015/16-2016/17

Parametrization	Type of scale predictor used	Model Class	Family / Distribution	zero-inflation	Parameter Selection	Training MAE	Testing MAE
Full	IN	GAM	NB2	0	0	1.908	2.055
Full	NM	GAM	NB2	0	1	1.915	2.061
Full	NM	GAM	TW	0	1	1.784	2.074
Full	NM	GAM	Pois	1	1	1.792	2.075
Full	NM	GAM	Pois	0	1	1.783	2.075
Full	IN	GAM	Pois	0	0	1.772	2.078
Full	BD	GAM	Pois	0	1	1.788	2.085
Param 1	IN	GLMM	NB1	0	0	1.742	2.092
Param 2	IN	GLMM	NB1	0	0	1.736	2.098
Param 1	IN	GLMM	NB2	0	0	1.742	2.099
Param 1	IN	GLMM	NB1	1	0	1.744	2.099
Param 2	IN	GLMM	NB1	1	0	1.737	2.105
Param 2	IN	GLMM	NB2	0	0	1.737	2.105
Param 1	IN	GLMM	NB2	1	0	1.744	2.106
Param 2	IN	GAM	NB2	0	0	1.735	2.106
Param 2	IN	GLMM	NB2	1	0	1.739	2.111
Param 1	IN	GAM	NB2	0	0	1.740	2.111
Param 1	IN	GLMM	Pois	0	0	1.733	2.121
Param 1	IN	GLMM	Pois	1	0	1.737	2.123
Param 3	IN	GLMM	NB2	0	0	1.736	2.124
Param 2	IN	GLMM	Pois	0	0	1.733	2.124
Param 2	IN	GLMM	Pois	1	0	1.737	2.126
Param 3	IN	GAM	NB2	0	0	1.734	2.126
Param 2	IN	GAM	TW	0	0	1.732	2.126
Param 1	IN	GAM	TW	0	0	1.737	2.128
Param 4	IN	GLMM	Pois	0	0	1.715	2.129
Param 4	IN	GLMM	Pois	1	0	1.719	2.129
Param 3	IN	GLMM	NB2	1	0	1.738	2.130
Param 2	IN	GAM	Pois	1	0	1.734	2.131
Param 1	IN	GAM	Pois	1	0	1.739	2.133
Param 1	IN	GAM	Pois	0	0	1.736	2.137
Param 2	IN	GAM	Pois	0	0	1.729	2.139
Param 3	IN	GLMM	Pois	0	0	1.732	2.143
Param 3	IN	GLMM	Pois	1	0	1.737	2.144
Param 3	IN	GAM	TW	0	0	1.731	2.147
Param 3	IN	GAM	Pois	1	0	1.732	2.149
Full	BD	RF	-	0	1	0.806	2.155
Full	IN	RF	-	0	1	0.806	2.156
Full	NM	RF	-	0	1	0.813	2.158
Param 3	IN	GAM	Pois	0	0	1.729	2.165
Full	IN	GLMM	NB1	0	0	1.721	2.292
Param 4	IN	GAM	NB2	0	0	2.273	2.340

Parametrization	Type of scale predictor used	Model Class	Family / Distribution	zero-inflation	Parameter Selection	Training MAE	Testing MAE
Full	NM	GLMM	Pois	total bed-days, op & ae attendaers	0	1.715	2.342
Param 4	IN	GAM	TW	0	0	2.258	2.363
Full	BD	GLMM	Pois	0	0	1.718	2.367
Full	NM	GLMM	NB2	0	0	1.726	2.368
Full	NM	GLMM	Pois	0	0	1.714	2.370
Full	NM	GLMM	Pois	total bed-days	0	1.715	2.370
Full	NM	GLMM	Pois	1	0	1.715	2.379
Param 4	IN	GAM	Pois	1	0	2.225	2.398
Param 4	IN	GLMM	NB1	1	0	2.471	2.415
Param 4	IN	GLMM	NB1	0	0	2.473	2.426
Param 4	IN	GLMM	NB2	1	0	2.490	2.426
Param 4	IN	GAM	Pois	0	0	2.201	2.433
Param 4	IN	GLMM	NB2	0	0	2.491	2.437
Param 3	IN	GLMM	NB1	1	0	2.532	2.446
Param 3	IN	GLMM	NB1	0	0	2.534	2.463
Full	NM	GAM	TW	0	0	1.634	2.566
Full	NM	GAM	Pois	0	0	1.632	2.571
Full	BD	GAM	Pois	0	0	1.635	2.576

C.5 Model summary for death or severe harm incident reports

NRLS data for 2015/16 (training) and 2016/17 (testing) data sets. Parametrizations are detailed in Chapter 7. MAE = Mean Absolute Error. Type of scale predictors: NB=Non-mandatory incidents, BD=total bed-days and IN=total incidents. Zero-inflation indicates presences or absence, or specific formulae for zero inflation used.

Trust Name	All												DS Incidents											
	Poisson GLMM		NB1 GLMM		NB2 GLMM		Poisson GAM		NB2 GAM		Random Forest		Poisson GAM (Full)		NB2 GAM (Full)		Poisson GLMM (Param 1)		NB1 GLMM (Param 1)		NB1 GLMM (Param 2)			
	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI		
RIF - ISLE OF WIGHT NHS TRUST	-4.602	-3.385	-2.310	-1.530	-3.934	-2.842	-3.465	-2.545	-1.941	-1.174	-3.454	-2.253	3.427	1.416	5.080	1.878	2.573	1.175	2.581	1.179	3.631	1.480		
R1H - BARTS HEALTH NHS TRUST	-1.963	-1.266	-2.824	-1.922	-3.353	-2.354	-0.582	-0.361	-1.010	-0.590	0.417	0.263	1.238	0.622	1.172	0.592	3.052	1.377	2.715	1.259	-0.240	-0.143		
R1K - LONDON NORTH WEST UNIVERSITY HEALTHCARE NHS TRUST	-2.274	-1.487	-3.933	-2.849	-2.652	-1.799	-4.702	-3.788	-3.705	-2.423	-0.545	-0.346	-0.184	-0.106	0.051	0.028	5.785	2.158	5.492	2.088	0.625	0.339		
RA2 - ROYAL SURREY COUNTY HOSPITAL NHS FOUNDATION TRUST	-2.515	-1.663	-1.797	-1.160	-2.122	-1.404	-2.608	-1.811	-1.569	-0.935	-2.651	-1.719	1.138	0.567	1.014	0.510	1.493	0.754	1.468	0.743	2.388	1.075		
RA3 - WESTON AREA HEALTH NHS TRUST	-2.637	-1.752	-4.227	-3.115	-1.862	-1.217	-3.841	-2.897	-2.339	-1.439	-2.596	-1.669	-3.153	-3.057	-3.023	-2.783	-3.271	-4.095	-3.273	-4.096	-3.254	-3.687		
RA4 - YEOVIL DISTRICT HOSPITAL NHS FOUNDATION TRUST	-3.857	-2.724	-3.635	-2.587	-3.350	-2.349	-2.745	-1.923	-3.147	-2.005	-3.173	-2.060	0.431	0.225	0.403	0.211	-1.093	-0.758	-1.092	-0.757	-0.520	-0.316		
RA7 - UNIVERSITY HOSPITALS BRISTOL NHS FOUNDATION TRUST	2.688	1.450	2.712	1.453	3.003	1.634	-0.610	-0.378	1.163	0.627	1.156	0.719	1.532	0.741	1.760	0.834	1.731	0.860	1.667	0.834	1.092	0.560		
RA9 - TORBAY AND SOUTH DEVON NHS FOUNDATION TRUST	-3.428	-2.368	-2.986	-2.050	-3.367	-2.364	-0.924	-0.582	-2.665	-1.663	-3.501	-2.299	-2.769	-2.315	-2.836	-2.360	-1.770	-1.390	-1.786	-1.406	-1.824	-1.371		
RAE - BRADFORD TEACHING HOSPITALS NHS FOUNDATION TRUST	1.092	0.623	-0.396	-0.240	1.774	1.006	-2.066	-1.389	-0.948	-0.552	0.028	0.018	-3.538	-3.678	-3.479	-3.465	-2.947	-3.183	-2.946	-3.176	-3.375	-3.880		
RAJ - SOUTHEND UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	2.465	1.339	2.800	1.495	2.732	1.499	2.093	1.145	2.935	1.495	1.918	1.176	1.278	0.631	1.377	0.673	1.423	0.728	1.330	0.687	2.335	1.063		
RAL - ROYAL FREE LONDON NHS FOUNDATION TRUST	-2.443	-1.610	-4.210	-3.101	-2.477	-1.666	-3.512	-2.588	-4.153	-2.775	-2.073	-1.344	0.510	0.271	0.698	0.363	2.011	0.970	2.002	0.967	-0.854	-0.550		
RAP - NORTH MIDDLESEX UNIVERSITY HOSPITAL NHS TRUST	0.503	0.293	-1.394	-0.883	0.503	0.299	-2.369	-1.621	-1.654	-0.990	-2.194	-1.417	0.470	0.249	0.583	0.305	0.537	0.300	0.495	0.278	-0.043	-0.025		
RAS - THE HILLINGDON HOSPITALS NHS FOUNDATION TRUST	-2.147	-1.396	-1.116	-0.698	-2.127	-1.407	-0.213	-0.130	-1.209	-0.711	-1.098	-0.697	-0.565	-0.335	-0.590	-0.348	-0.298	-0.184	-0.320	-0.198	0.109	0.061		
RAX - KINGSTON HOSPITAL NHS FOUNDATION TRUST	-0.573	-0.348	-1.906	-1.237	0.110	0.066	-2.302	-1.569	-1.681	-1.006	-2.130	-1.368	0.923	0.465	0.549	0.287	0.136	0.079	0.124	0.072	0.115	0.065		
RBA - TAUNTON AND SOMERSET NHS FOUNDATION TRUST	-1.774	-1.133	-1.575	-1.006	-1.347	-0.861	-1.831	-1.214	-1.943	-1.176	-1.834	-1.177	-2.605	-2.124	-2.777	-2.303	-2.432	-2.228	-2.448	-2.250	-2.176	-1.761		
RBD - DORSET COUNTY HOSPITAL NHS FOUNDATION TRUST	-3.085	-2.095	-2.248	-1.485	-2.399	-1.607	-0.610	-0.378	-2.641	-1.646	-1.175	-0.743	0.035	0.020	0.135	0.073	-0.285	-0.176	-0.291	-0.179	0.561	0.299		
RBK - WALSALL HEALTHCARE NHS TRUST	6.360	3.067	7.795	3.561	6.618	3.231	4.682	2.320	8.890	3.855	8.503	4.835	1.187	0.589	1.460	0.706	0.077	0.046	-0.037	-0.022	1.463	0.719		
RBL - WIRRAL UNIVERSITY TEACHING HOSPITAL NHS FOUNDATION TRUST	-0.248	-0.149	1.176	0.667	-0.296	-0.181	1.583	0.885	0.412	0.228	-0.367	-0.232	-0.050	-0.028	-0.739	-0.446	0.297	0.171	0.207	0.120	1.242	0.623		
RBN - ST HELENS AND KNOWSLEY HOSPITAL SERVICES NHS TRUST	0.473	0.276	1.422	0.799	0.477	0.283	1.459	0.821	0.356	0.198	-1.011	-0.645	-0.717	-0.434	-0.737	-0.444	-0.111	-0.067	-0.187	-0.114	-0.156	-0.091		
RBT - MID CHESHIRE HOSPITALS NHS FOUNDATION TRUST	-0.287	-0.172	0.642	0.372	-0.373	-0.229	0.833	0.482	-0.272	-0.154	-0.775	-0.490	-0.353	-0.203	-0.304	-0.173	-0.949	-0.644	-0.978	-0.667	-0.481	-0.291		
RBZ - NORTHERN DEVON HEALTHCARE NHS TRUST	1.706	0.951	3.311	1.736	1.679	0.955	3.010	1.587	6.313	2.919	1.568	0.965	5.164	1.951	4.092	1.658	5.808	2.142	5.692	2.116	7.719	2.497		
RC1 - BEDFORD HOSPITAL NHS TRUST	-2.285	-1.494	-1.030	-0.641	-1.929	-1.265	-1.078	-0.685	-1.077	-0.630	-3.297	-2.155	5.653	2.043	6.108	2.158	4.188	1.700	4.163	1.694	5.582	2.012		
RC9 - LUTON AND DUNSTABLE UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	-0.919	-0.566	-0.484	-0.294	-0.633	-0.393	-0.044	-0.027	-1.610	-0.961	-1.359	-0.868	-2.387	-1.861	-2.764	-2.281	-2.409	-2.191	-2.431	-2.222	-2.326	-1.944		
RCB - YORK TEACHING HOSPITAL NHS FOUNDATION TRUST	0.859	0.494	1.409	0.792	0.548	0.325	2.252	1.225	1.417	0.758	1.831	1.128	2.656	1.188	2.360	1.080	3.802	1.611	3.576	1.542	4.409	1.743		
RCD - HARROGATE AND DISTRICT NHS FOUNDATION TRUST	-2.401	-1.578	-2.259	-1.492	-2.153	-1.425	-1.648	-1.081	-2.309	-1.419	-4.680	-3.111	-2.784	-2.386	-2.859	-2.446	-2.746	-2.770	-2.750	-2.773	-2.591	-2.331		
RCF - AIREDALE NHS FOUNDATION TRUST	-1.643	-1.044	-0.466	-0.283	-1.395	-0.894	1.497	0.840	-0.830	-0.481	-1.182	-0.749	-0.711	-0.426	-0.650	-0.384	-1.690	-1.306	-1.702	-1.317	-1.247	-0.846		
RCX - THE QUEEN ELIZABETH HOSPITAL, KING'S LYNN, NHS FOUNDATION TRUST	-0.951	-0.587	0.222	0.131	-0.795	-0.497	0.352	0.208	-0.224	-0.127	0.693	0.430	-0.035	-0.020	0.147	0.080	-0.783	-0.518	-0.811	-0.539	-0.069	-0.039		
RD1 - ROYAL UNITED HOSPITALS BATH NHS FOUNDATION TRUST	-1.183	-0.737	-0.105	-0.063	-1.505	-0.969	-1.432	-0.928	-0.827	-0.479	-0.521	-0.329	-1.109	-0.704	-1.315	-0.851	-0.851	-0.570	-0.890	-0.599	-0.237	-0.139		
RD3 - POOLE HOSPITAL NHS FOUNDATION TRUST	2.884	1.545	2.776	1.484	2.813	1.539	2.938	1.554	3.432	1.722	3.844	2.300	-1.183	-0.757	-1.303	-0.841	-1.847	-1.474	-1.888	-1.518	-1.266	-0.862		
RD8 - MILTON KEYNES UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	-3.342	-2.299	-3.607	-2.563	-2.827	-1.932	-2.746	-1.923	-3.647	-2.378	-4.561	-3.033	-0.324	-0.186	-0.347	-0.199	-0.304	-0.188	-0.311	-0.192	0.024	0.013		
RDD - BASILDON AND THURROCK UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	1.552	0.870	3.405	1.781	1.995	1.123	2.163	1.181	3.893	1.927	2.786	1.694	0.143	0.079	0.324	0.175	0.597	0.333	0.485	0.275	1.324	0.661		
RDE - EAST SUFFOLK AND NORTH ESSEX NHS FOUNDATION TRUST	-0.159	-0.095	-0.385	-0.233	0.112	0.067	-1.695	-1.115	-0.343	-0.195	-1.486	-0.952	1.761	0.834	1.941	0.906	2.014	0.973	1.937	0.943	3.163	1.345		
RDU - FRIMLEY HEALTH NHS FOUNDATION TRUST	-0.524	-0.318	-1.380	-0.874	-1.080	-0.683	-0.624	-0.388	-1.339	-0.791	-0.550	-0.350	-0.764	-0.467	-1.048	-0.657	0.483	0.274	0.336	0.194	-0.414	-0.252		
RDZ - THE ROYAL BOURNEMOUTH AND CHRISTCHURCH HOSPITALS NHS FOUNDATION TRUST	3.063	1.630	3.147	1.660	2.940	1.602	3.894	1.985	2.709	1.389	0.554	0.345	1.841	0.859	1.504	0.723	0.510	0.286	0.445	0.252	1.362	0.674		

Trust Name	All												DS Incidents											
	Poisson GLMM		NB1 GLMM		NB2 GLMM		Poisson GAM		NB2 GAM		Random Forest		Poisson GAM (Full)		NB2 GAM (Full)		Poisson GLMM (Param 1)		NB1 GLMM (Param 1)		NB1 GLMM (Param 2)			
	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI		
RE9 - SOUTH TYNESIDE NHS FOUNDATION TRUST	-4.951	-3.716	-4.310	-3.193	-4.175	-3.056	-4.711	-3.798	-3.337	-2.145	-4.640	-3.079	-2.313	-1.785	-2.100	-1.539	-2.315	-2.059	-2.315	-2.058	-2.195	-1.786		
REF - ROYAL CORNWALL HOSPITALS NHS TRUST	0.094	0.056	0.670	0.388	0.266	0.159	0.096	0.057	-0.041	-0.023	-0.620	-0.394	1.259	0.624	1.163	0.581	1.664	0.833	1.557	0.788	3.020	1.301		
REM - AINTREE UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	-3.368	-2.320	-4.907	-3.777	-2.312	-1.543	-0.858	-0.539	-4.674	-3.204	-5.081	-3.418	0.418	0.219	0.216	0.115	-0.707	-0.460	-0.647	-0.417	-1.928	-1.478		
RF4 - BARKING, HAVERING AND REDBRIDGE UNIVERSITY HOSPITALS NHS TRU	-0.848	-0.521	-1.292	-0.815	-1.164	-0.739	-3.586	-2.657	-1.931	-1.169	-2.377	-1.546	3.309	1.419	3.099	1.351	7.493	2.534	7.274	2.492	4.767	1.852		
RFF - BARNSELEY HOSPITAL NHS FOUNDATION TRUST	-2.387	-1.569	-2.390	-1.590	-2.237	-1.488	-0.126	-0.076	-0.465	-0.266	-0.026	-0.016	0.260	0.140	0.386	0.205	-0.589	-0.379	-0.627	-0.406	-0.026	-0.015		
RFR - THE ROTHERHAM NHS FOUNDATION TRUST	-1.526	-0.965	-0.100	-0.060	-1.064	-0.672	-0.362	-0.221	-0.490	-0.280	-1.261	-0.803	-1.359	-0.892	-1.295	-0.835	-1.158	-0.813	-1.184	-0.835	-0.740	-0.465		
RFS - CHESTERFIELD ROYAL HOSPITAL NHS FOUNDATION TRUST	-2.625	-1.745	-1.646	-1.055	-2.430	-1.630	-0.781	-0.489	-1.865	-1.125	-1.281	-0.815	-1.842	-1.303	-1.893	-1.335	-1.934	-1.571	-1.952	-1.591	-1.649	-1.199		
RGN - NORTH WEST ANGLIA NHS FOUNDATION TRUST	8.325	3.811	7.943	3.615	8.087	3.798	9.876	4.160	9.195	3.961	5.043	2.995	1.714	0.825	2.257	1.041	2.305	1.102	2.023	0.992	3.709	1.541		
RGP - JAMES PAGET UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	-3.900	-2.761	-3.019	-2.075	-3.589	-2.548	-1.444	-0.936	-3.559	-2.311	-2.595	-1.675	-3.231	-3.094	-3.336	-3.203	-2.912	-3.108	-2.917	-3.115	-2.885	-2.825		
RGQ - IPSWICH HOSPITAL NHS TRUST	-1.994	-1.287	-1.707	-1.098	-1.752	-1.140	-2.540	-1.757	-1.993	-1.209	-3.364	-2.208	-0.086	-0.048	-0.070	-0.039	0.763	0.415	0.725	0.397	1.342	0.664		
RGR - WEST SUFFOLK NHS FOUNDATION TRUST	-4.187	-3.010	-3.883	-2.804	-3.799	-2.726	-3.465	-2.546	-3.814	-2.507	-4.971	-3.322	-0.031	-0.017	0.025	0.014	-0.354	-0.220	-0.356	-0.221	0.146	0.082		
RGT - CAMBRIDGE UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	-0.167	-0.100	-0.972	-0.604	0.078	0.047	-1.991	-1.332	-0.869	-0.505	-0.890	-0.568	-2.536	-2.025	-2.657	-2.135	-1.973	-1.616	-1.985	-1.630	-2.525	-2.202		
RHB - ROYAL DEVON AND EXETER NHS FOUNDATION TRUST	4.656	2.359	3.317	1.740	4.367	2.276	1.124	0.642	3.908	1.933	1.971	1.212	-3.195	-2.983	-3.251	-3.007	-2.737	-2.739	-2.746	-2.751	-3.027	-3.054		
RHM - UNIVERSITY HOSPITAL SOUTHAMPTON NHS FOUNDATION TRUST	-0.948	-0.585	1.249	0.707	-1.399	-0.897	1.331	0.753	0.600	0.330	0.979	0.611	11.485	3.287	10.893	3.214	12.924	3.473	12.677	3.443	12.382	3.343		
RHQ - SHEFFIELD TEACHING HOSPITALS NHS FOUNDATION TRUST	0.765	0.441	2.358	1.280	0.031	0.019	5.895	2.802	3.555	1.778	2.064	1.274	-0.623	-0.373	-1.032	-0.645	-1.254	-0.900	-1.365	-0.998	-1.449	-1.020		
RHU - PORTSMOUTH HOSPITALS NHS TRUST	-2.733	-1.826	-1.672	-1.074	-3.219	-2.244	-1.454	-0.944	-2.482	-1.537	-4.151	-2.767	1.326	0.657	0.568	0.301	3.614	1.542	3.471	1.499	3.703	1.529		
RHW - ROYAL BERKSHIRE NHS FOUNDATION TRUST	3.717	1.939	4.100	2.094	4.270	2.232	4.668	2.314	2.746	1.407	0.248	0.156	-2.517	-1.992	-2.619	-2.078	-1.599	-1.216	-1.659	-1.276	-1.367	-0.947		
RJ1 - GUY'S AND ST THOMAS' NHS FOUNDATION TRUST	1.275	0.722	3.576	1.860	1.237	0.715	9.285	3.976	1.983	1.041	0.660	0.414	0.543	0.288	0.663	0.346	0.806	0.440	0.682	0.378	-0.085	-0.050		
RJ2 - LEWISHAM AND GREENWICH NHS TRUST	3.241	1.717	1.993	1.096	3.204	1.732	0.972	0.559	1.807	0.954	1.484	0.919	-3.697	-3.881	-3.837	-4.058	-2.736	-2.714	-2.790	-2.810	-3.034	-3.031		
RJ6 - CROYDON HEALTH SERVICES NHS TRUST	-2.493	-1.646	-2.295	-1.519	-1.796	-1.171	-4.193	-3.246	-3.255	-2.085	-5.049	-3.387	2.103	0.961	2.186	0.992	1.667	0.827	1.652	0.821	2.506	1.116		
RJ7 - ST GEORGE'S UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	-1.559	-0.987	-0.499	-0.303	-2.156	-1.429	-2.076	-1.396	-1.980	-1.201	-0.989	-0.632	0.689	0.359	0.190	0.104	2.207	1.045	2.212	1.047	0.623	0.334		
RJC - SOUTH WARWICKSHIRE NHS FOUNDATION TRUST	1.349	0.762	1.746	0.969	1.618	0.922	1.437	0.809	3.326	1.673	2.801	1.692	15.306	3.786	15.929	3.892	13.648	3.540	13.494	3.524	17.967	3.958		
RJE - UNIVERSITY HOSPITALS OF NORTH MIDLANDS NHS TRUST	-2.683	-1.788	-2.620	-1.764	-3.045	-2.105	-3.489	-2.567	-2.944	-1.860	-1.176	-0.754	-2.662	-2.179	-2.829	-2.351	-2.011	-1.661	-2.035	-1.690	-2.683	-2.432		
RJF - BURTON HOSPITALS NHS FOUNDATION TRUST	-0.858	-0.527	-1.043	-0.650	-0.936	-0.588	-0.313	-0.191	-0.182	-0.103	-1.968	-1.264	-2.058	-1.508	-2.042	-1.474	-1.767	-1.387	-1.783	-1.403	-1.942	-1.492		
RJL - NORTHERN LINCOLNSHIRE AND GOOLE NHS FOUNDATION TRUST	1.568	0.879	2.761	1.477	1.606	0.916	3.134	1.645	3.002	1.526	1.779	1.096	-3.258	-3.048	-3.365	-3.151	-2.605	-2.492	-2.649	-2.564	-2.676	-2.426		
RJN - EAST CHESHIRE NHS TRUST	6.812	3.242	7.181	3.335	6.822	3.310	4.974	2.438	9.936	4.201	5.485	3.213	-3.107	-2.885	-2.976	-2.626	-3.154	-3.671	-3.173	-3.715	-2.964	-2.971		
RJR - COUNTNESS OF CHESTER HOSPITAL NHS FOUNDATION TRUST	4.499	2.290	5.593	2.724	4.731	2.438	4.266	2.145	6.010	2.802	6.804	3.936	-1.513	-1.016	-1.436	-0.945	-1.211	-0.860	-1.273	-0.914	-0.510	-0.311		
RJZ - KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	2.564	1.389	1.828	1.012	2.141	1.199	1.125	0.642	2.842	1.453	1.450	0.902	5.557	2.085	5.906	2.183	5.755	2.169	5.380	2.077	3.901	1.615		
RK5 - SHERWOOD FOREST HOSPITALS NHS FOUNDATION TRUST	-1.411	-0.888	-1.682	-1.080	-0.917	-0.576	-1.006	-0.637	-1.027	-0.599	-1.346	-0.860	-2.512	-1.997	-2.616	-2.086	-1.995	-1.642	-2.023	-1.675	-2.031	-1.590		
RK9 - UNIVERSITY HOSPITALS PLYMOUTH NHS TRUST	2.505	1.359	2.730	1.462	3.031	1.648	-0.519	-0.320	3.300	1.663	1.996	1.229	2.527	1.124	2.687	1.181	1.626	0.815	1.576	0.794	1.557	0.761		
RKB - UNIVERSITY HOSPITALS COVENTRY AND WARWICKSHIRE NHS TRUST	-0.610	-0.371	-0.738	-0.454	-0.874	-0.548	-2.409	-1.653	-1.088	-0.637	-0.077	-0.049	1.804	0.855	1.926	0.903	2.946	1.317	2.916	1.307	0.914	0.477		
RKE - WHITTINGTON HEALTH NHS TRUST	-2.680	-1.785	-1.568	-1.001	-2.313	-1.543	-3.356	-2.447	-2.609	-1.624	-4.052	-2.669	6.869	2.330	6.568	2.276	5.759	2.109	5.767	2.113	6.316	2.188		
RL4 - THE ROYAL WOLVERHAMPTON NHS TRUST	-1.409	-0.887	-0.969	-0.602	-1.372	-0.879	-0.895	-0.563	-0.744	-0.430	-1.654	-1.066	-2.225	-1.669	-2.107	-1.530	-0.792	-0.527	-0.873	-0.588	-1.151	-0.772		
RLN - CITY HOSPITALS SUNDERLAND NHS FOUNDATION TRUST	6.765	3.227	8.232	3.717	6.067	3.009	9.795	4.134	8.027	3.556	4.654	2.775	-1.742	-1.210	-1.355	-0.883	-0.974	-0.668	-1.097	-0.767	-0.581	-0.359		
RLQ - WYE VALLEY NHS TRUST	-0.212	-0.127	0.301	0.177	0.122	0.073	0.981	0.564	1.249	0.672	2.134	1.299	-2.750	-2.322	-2.778	-2.318	-2.704	-2.678	-2.720	-2.705	-2.471	-2.146		

Trust Name	All												DS Incidents									
	Poisson GLMM		NB1 GLMM		NB2 GLMM		Poisson GAM		NB2 GAM		Random Forest		Poisson GAM (Full)		NB2 GAM (Full)		Poisson GLMM (Param 1)		NB1 GLMM (Param 1)		NB1 GLMM (Param 2)	
	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI	CQC	SHMI
RLT - GEORGE ELIOT HOSPITAL NHS TRUST	-2.608	-1.731	-2.271	-1.501	-2.346	-1.567	-2.852	-2.012	-0.795	-0.460	-2.407	-1.549	0.295	0.158	0.430	0.226	-0.388	-0.242	-0.392	-0.245	0.069	0.039
RM1 - NORFOLK AND NORWICH UNIVERSITY HOSPITALS NHS FOUNDATION TR	3.462	1.821	2.368	1.285	3.441	1.847	0.379	0.224	3.378	1.698	2.306	1.416	-0.986	-0.616	-1.087	-0.683	-0.995	-0.683	-1.051	-0.728	-1.231	-0.836
RM2 - UNIVERSITY HOSPITAL OF SOUTH MANCHESTER NHS FOUNDATION TRU	1.465	0.824	2.071	1.136	1.797	1.019	-1.057	-0.671	2.207	1.150	1.724	1.063	-0.065	-0.036	0.102	0.056	0.153	0.090	0.149	0.087	-0.235	-0.139
RM3 - SALFORD ROYAL NHS FOUNDATION TRUST	0.483	0.282	1.664	0.926	0.784	0.461	3.556	1.836	-0.603	-0.347	-0.542	-0.343	-0.802	-0.488	-1.439	-0.947	-1.623	-1.239	-1.672	-1.289	-1.292	-0.884
RMC - BOLTON NHS FOUNDATION TRUST	1.545	0.866	3.230	1.699	1.585	0.905	3.545	1.831	3.095	1.569	3.509	2.110	-1.504	-1.008	-1.608	-1.084	-1.485	-1.105	-1.535	-1.153	-1.260	-0.858
RMP - TAMESIDE AND GLOSSOP INTEGRATED CARE NHS FOUNDATION TRUST	1.592	0.891	1.409	0.792	2.597	1.431	1.465	0.824	3.343	1.681	2.831	1.713	-2.574	-2.082	-2.430	-1.883	-2.501	-2.330	-2.528	-2.370	-2.283	-1.890
RN3 - GREAT WESTERN HOSPITALS NHS FOUNDATION TRUST	-1.701	-1.084	-2.384	-1.586	-1.634	-1.058	-1.492	-0.970	-2.041	-1.240	-2.433	-1.575	-1.002	-0.628	-0.934	-0.576	0.323	0.185	0.291	0.167	0.228	0.127
RN5 - HAMPSHIRE HOSPITALS NHS FOUNDATION TRUST	0.992	0.568	0.040	0.024	0.821	0.482	0.083	0.050	0.757	0.414	0.162	0.101	0.845	0.436	0.477	0.254	2.324	1.095	2.222	1.057	2.789	1.227
RN7 - DARTFORD AND GRAVESHAM NHS TRUST	1.154	0.656	1.278	0.722	1.022	0.595	0.397	0.234	1.619	0.860	0.912	0.565	-3.209	-2.987	-3.285	-3.041	-2.566	-2.435	-2.589	-2.472	-2.512	-2.195
RNA - THE DUDLEY GROUP NHS FOUNDATION TRUST	-2.146	-1.395	-0.320	-0.193	-1.859	-1.216	-1.444	-0.936	-2.313	-1.422	-2.917	-1.906	0.778	0.402	0.857	0.439	1.440	0.733	1.376	0.706	2.124	0.982
RNL - NORTH CUMBRIA UNIVERSITY HOSPITALS NHS TRUST	0.677	0.392	1.647	0.918	0.969	0.565	0.406	0.239	2.014	1.056	2.035	1.245	4.097	1.635	4.151	1.655	2.849	1.285	2.753	1.253	4.237	1.672
RNQ - KETTERING GENERAL HOSPITAL NHS FOUNDATION TRUST	-1.407	-0.885	-1.709	-1.099	-1.327	-0.848	-1.332	-0.859	-1.711	-1.026	-4.232	-2.812	2.079	0.952	1.952	0.904	1.278	0.660	1.240	0.643	2.065	0.957
RNS - NORTHAMPTON GENERAL HOSPITAL NHS TRUST	-1.853	-1.189	-0.894	-0.554	-1.436	-0.922	-0.680	-0.423	-2.065	-1.256	-2.732	-1.780	-1.311	-0.856	-1.545	-1.032	-0.742	-0.489	-0.782	-0.518	-0.614	-0.380
RNZ - SALISBURY NHS FOUNDATION TRUST	-1.236	-0.772	0.949	0.543	-0.719	-0.448	2.521	1.355	0.605	0.332	1.473	0.904	-2.638	-2.165	-2.692	-2.196	-2.432	-2.227	-2.448	-2.250	-2.056	-1.622
RP5 - DONCASTER AND BASSETT LAW TEACHING HOSPITALS NHS FOUNDATIO	1.702	0.949	1.442	0.810	1.579	0.902	2.907	1.539	1.859	0.980	0.003	0.002	4.180	1.688	4.118	1.674	5.395	2.057	5.155	1.997	5.768	2.099
RPA - MEDWAY NHS FOUNDATION TRUST	-4.413	-3.212	-3.902	-2.821	-4.471	-3.327	-2.877	-2.032	-4.371	-2.951	-5.795	-3.937	5.490	2.022	5.243	1.967	5.617	2.083	5.580	2.076	6.193	2.169
RQ6 - ROYAL LIVERPOOL AND BROADGREEN UNIVERSITY HOSPITALS NHS TR	-0.498	-0.302	-2.029	-1.326	0.301	0.180	-0.728	-0.454	-1.037	-0.606	-1.430	-0.919	1.160	0.573	1.141	0.563	0.292	0.167	0.294	0.169	-1.476	-1.043
RQ8 - MID ESSEX HOSPITAL SERVICES NHS TRUST	-3.638	-2.540	-3.230	-2.247	-3.520	-2.490	-4.072	-3.124	-4.196	-2.809	-5.186	-3.488	-2.033	-1.479	-1.935	-1.370	-0.813	-0.540	-0.826	-0.550	-0.696	-0.435
RQ9 - CHELSEA AND WESTMINSTER HOSPITAL NHS FOUNDATION TRUST	-1.903	-1.223	-3.967	-2.879	-2.148	-1.423	-5.140	-4.296	-5.583	-4.016	-4.433	-2.952	-2.507	-2.019	-2.495	-1.975	-2.193	-1.897	-2.162	-1.854	-3.113	-3.238
RQW - THE PRINCESS ALEXANDRA HOSPITAL NHS TRUST	0.808	0.465	2.339	1.270	0.944	0.551	0.950	0.547	2.335	1.211	1.224	0.755	-1.604	-1.092	-1.743	-1.199	-1.778	-1.398	-1.809	-1.431	-1.607	-1.160
RQX - HOMERTON UNIVERSITY HOSPITAL NHS FOUNDATION TRUST	1.810	1.005	2.577	1.386	1.459	0.836	1.225	0.696	0.862	0.469	0.188	0.117	-1.291	-0.839	-1.720	-1.178	-1.069	-0.739	-1.090	-0.757	-1.125	-0.749
RR1 - HEART OF ENGLAND NHS FOUNDATION TRUST	-1.794	-1.149	-1.288	-0.812	-2.466	-1.658	1.576	0.882	-0.920	-0.535	0.222	0.140	5.369	2.043	4.163	1.706	8.261	2.725	7.760	2.628	6.188	2.229
RR7 - GATESHEAD HEALTH NHS FOUNDATION TRUST	-2.803	-1.878	-2.342	-1.554	-2.443	-1.640	-1.966	-1.313	-1.774	-1.066	-1.709	-1.093	0.862	0.439	0.876	0.445	1.015	0.537	0.993	0.527	1.074	0.546
RR8 - LEEDS TEACHING HOSPITALS NHS TRUST	-1.098	-0.682	0.050	0.030	-1.856	-1.214	1.682	0.937	0.954	0.519	2.268	1.396	-1.513	-1.017	-1.622	-1.096	-1.236	-0.886	-1.373	-1.006	-1.819	-1.365
RRF - WRIGHTINGTON, WIGAN AND LEIGH NHS FOUNDATION TRUST	-1.804	-1.154	1.347	0.759	-1.746	-1.136	1.361	0.769	0.937	0.509	0.002	0.001	4.654	1.809	4.872	1.874	5.104	1.963	4.996	1.937	6.604	2.264
RRK - UNIVERSITY HOSPITALS BIRMINGHAM NHS FOUNDATION TRUST	5.201	2.595	7.926	3.610	6.382	3.139	20.888	6.847	8.619	3.765	2.964	1.809	-2.296	-1.751	-1.937	-1.374	-2.818	-2.857	-2.878	-2.970	-2.755	-2.539
RRV - UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	-2.223	-1.450	-0.529	-0.322	-1.653	-1.071	2.578	1.383	-2.188	-1.338	-1.004	-0.641	2.378	1.072	2.366	1.068	3.043	1.344	3.063	1.351	1.450	0.717
RTD - THE NEWCASTLE UPON TYNE HOSPITALS NHS FOUNDATION TRUST	-3.210	-2.195	-1.453	-0.923	-3.907	-2.821	-1.269	-0.816	-1.957	-1.185	-1.491	-0.962	3.010	1.307	2.572	1.155	3.338	1.459	3.203	1.416	1.512	0.749
RTE - GLOUCESTERSHIRE HOSPITALS NHS FOUNDATION TRUST	0.387	0.226	0.970	0.555	0.536	0.318	1.901	1.050	0.835	0.456	1.122	0.697	-1.202	-0.774	-1.494	-0.992	-0.293	-0.182	-0.398	-0.251	0.098	0.056
RTF - NORTHUMBRIA HEALTHCARE NHS FOUNDATION TRUST	4.350	2.225	2.594	1.396	4.609	2.385	3.109	1.634	4.957	2.380	2.122	1.303	-1.281	-0.833	-1.130	-0.715	-0.695	-0.457	-0.800	-0.534	-0.892	-0.577
RTG - UNIVERSITY HOSPITALS OF DERBY AND BURTON NHS FOUNDATION TR	-0.435	-0.263	-1.336	-0.844	-0.239	-0.146	-2.377	-1.628	-0.670	-0.386	-0.517	-0.328	-1.482	-0.991	-1.602	-1.079	-0.948	-0.644	-0.957	-0.651	-1.686	-1.235
RTH - OXFORD UNIVERSITY HOSPITALS NHS FOUNDATION TRUST	1.373	0.775	3.238	1.704	1.013	0.590	1.927	1.063	3.034	1.541	0.980	0.612	-0.032	-0.018	-0.181	-0.102	0.012	0.007	-0.097	-0.059	-0.121	-0.071
RTK - ASHFORD AND ST PETER'S HOSPITALS NHS FOUNDATION TRUST	0.316	0.185	-1.115	-0.697	0.467	0.278	-1.227	-0.786	-0.046	-0.026	-1.348	-0.859	-1.823	-1.284	-1.976	-1.409	-1.374	-1.001	-1.397	-1.022	-1.203	-0.811
RTP - SURREY AND SUSSEX HEALTHCARE NHS TRUST	-1.982	-1.279	-1.514	-0.965	-1.896	-1.242	-1.636	-1.072	-2.341	-1.441	-3.092	-2.022	0.949	0.480	0.678	0.351	0.738	0.403	0.699	0.383	1.319	0.655

C.6 Adjusted z-score output for organisations from NRLS-HES models for 2015/16.

Row are organisations, and columns represent models, grouped by total or DS incidents. Purple indicates z-score below -3 and yellow indicates a z-score above 3.

Trust	Total Incidents										DS Incidents													
	Poisson GLMM		NB1 GLMM		NB2 GLMM		Poisson GAM		NB2 GAM		Random Forest	Poisson GAM (full)		NB2 GAM (full)		Poisson GLMM (parameterisation 1)		NB1 GLMM (parameterisation 1)		NB1 GLMM (parameterisation 2)				
	+	-	+	-	+	-	+	-	+	-		+	-	+	-	+	-	+	-	+	-			
R1F	0	11	0	5	0	8	0	12	0	3	0	3	1	0	1	0	0	0	0	0	1	0		
R1K	0	6	0	10	0	6	0	12	0	7	0	7	0	1	0	1	3	0	3	0	0	0		
RA2	0	5	0	5	0	4	0	10	0	2	0	2	0	0	0	0	0	0	0	0	0			
RA3	0	4	0	11	0	3	0	11	0	4	0	4	0	6	0	5	0	8	0	8	0	8		
RA4	0	7	0	9	0	5	0	8	0	5	0	5	0	2	0	2	0	2	0	2	0	2		
RA7	4	0	5	0	4	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0			
RA9	0	8	0	8	0	8	0	3	0	5	0	5	0	5	0	5	0	4	0	4	0	4		
RAE	0	0	0	1	1	0	0	10	0	2	0	2	0	8	0	8	0	8	0	8	0	8		
RAJ	1	0	3	0	2	0	2	0	3	0	3	0	1	0	1	0	0	0	0	0	1	0		
RAL	0	7	0	12	0	6	0	12	0	12	0	12	0	2	0	2	0	1	0	1	0	2		
RAP	0	0	0	3	0	0	0	8	0	3	0	3	0	0	0	0	0	0	0	0	0	0		
RAS	0	4	0	3	0	4	0	0	0	2	0	2	0	2	0	2	0	1	0	1	0	1		
RAX	0	0	0	5	0	0	0	7	0	3	0	3	0	2	0	2	0	2	0	2	0	2		
RBA	0	3	0	4	0	2	0	7	0	4	0	4	0	4	0	4	0	4	0	4	0	4		
RBD	0	6	0	6	0	4	0	3	0	6	0	6	0	0	0	0	0	0	0	0	0	0		
RBK	11	0	12	0	11	0	11	0	12	0	12	0	0	0	0	0	0	0	0	0	0	0		
RBL	1	5	3	2	2	5	3	0	3	3	3	3	0	2	0	2	0	1	0	1	0	1		
RBN	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	1	0	1	0	1	0	1		
RBT	0	0	0	0	0	0	0	0	0	0	0	0	1	3	1	3	1	3	1	3	1	3		
RBZ	0	0	4	0	0	0	5	0	7	0	7	0	2	0	1	0	2	0	2	0	3	0		
RC1	0	4	0	2	0	4	0	4	0	1	0	1	2	1	2	1	2	1	2	1	2	1		
RC9	0	1	0	1	0	1	0	1	0	2	0	2	0	6	0	6	0	6	0	6	0	6		
RCB	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	2	0	2	0	2	0		
RCD	0	4	0	5	0	3	0	5	0	4	0	4	0	5	0	5	0	5	0	5	0	5		
RCF	0	3	0	1	0	2	1	0	0	1	0	1	0	4	0	4	0	4	0	4	0	4		
RCX	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	2	0	2	0	1		
RD1	0	2	0	0	0	2	0	6	0	1	0	1	0	1	0	1	0	0	0	0	0	0		
RD3	2	0	3	0	3	0	4	0	3	0	3	0	0	5	0	5	0	5	0	5	0	5		
RD8	0	8	0	10	0	5	0	10	0	8	0	8	0	0	0	0	0	0	0	0	0	0		
RDD	0	0	5	0	2	0	3	0	5	0	5	0	0	0	0	0	0	0	0	0	0	0		
RDE	0	0	0	0	0	0	0	7	0	0	0	0	0	1	0	1	0	1	0	1	0	1		
RDU	0	1	0	5	0	2	0	3	0	4	0	4	0	1	0	2	0	0	0	0	0	1		
RDZ	3	0	4	0	3	0	7	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0		
RE9	0	11	0	12	0	8	0	12	0	6	0	6	0	3	0	3	0	3	0	3	0	3		
REF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
REM	0	8	0	12	0	5	0	4	0	11	0	11	0	1	0	1	0	1	0	1	0	3		
RF4	0	2	0	4	0	2	0	12	0	5	0	5	1	0	1	0	4	0	4	0	3	0		
RFF	0	6	0	7	0	5	0	1	0	1	0	1	0	2	0	2	0	2	0	2	0	2		
RFR	0	2	0	0	0	1	0	1	0	0	0	0	0	2	0	2	0	2	0	2	0	1		
RFS	0	6	0	4	0	5	0	3	0	4	0	4	0	4	0	4	0	4	0	4	0	3		
RGN	12	0	12	0	12	0	12	0	12	0	12	0	0	0	0	0	0	0	0	0	1	0		
RGP	0	9	0	7	0	7	0	4	0	7	0	7	0	7	0	7	0	6	0	6	0	6		
RGQ	0	4	0	4	0	3	0	11	0	4	0	4	0	0	0	0	0	0	0	0	0	0		
RGR	0	9	0	10	0	8	0	12	0	7	0	7	0	3	0	3	0	2	0	2	0	2		
RGT	1	3	0	5	1	2	0	9	0	5	0	5	0	5	0	4	0	3	0	3	0	5		
RH8	7	0	6	0	7	0	0	0	6	0	6	0	0	7	0	7	0	6	0	6	0	7		
RHM	0	4	3	1	0	5	3	0	2	2	2	2	7	0	7	0	7	0	7	0	7	0		
RHQ	1	0	5	0	1	1	12	0	7	0	7	0	0	2	0	3	0	3	0	3	0	3		
RHU	0	6	0	4	0	7	0	6	0	5	0	5	0	0	0	0	1	0	0	0	1	0		
RHW	5	0	6	0	6	0	9	0	3	0	3	0	0	4	0	4	0	3	0	3	0	3		
RJ2	4	0	3	0	4	0	0	0	2	0	2	0	0	8	0	8	0	5	0	5	0	5		
RJ6	0	5	0	5	0	3	0	12	0	6	0	6	0	0	0	0	0	0	0	0	0	0		
RJ7	0	4	0	1	0	6	0	11	0	5	0	5	0	0	0	0	0	0	0	0	0	0		
RJC	0	0	1	0	1	0	0	0	3	0	3	0	5	1	5	1	5	1	5	1	5	1		
RJF	0	1	0	2	0	1	0	1	0	0	0	0	0	4	0	4	0	4	0	4	0	4		
RJL	1	0	4	0	1	0	6	0	4	0	4	0	0	6	0	5	0	5	0	5	0	4		
RJN	9	0	9	0	9	0	9	0	12	0	12	0	0	6	0	6	0	7	0	7	0	6		
RJR	6	0	9	0	6	0	9	0	9	0	9	0	0	2	0	2	0	2	0	2	0	0		
RJZ	5	0	4	0	4	0	1	0	7	0	7	0	3	0	4	0	4	0	4	0	3	0		
RK5	0	3	0	5	0	1	0	4	0	2	0	2	0	5	0	5	0	4	0	4	0	3		
RK9	4	0	5	0	5	0	0	3	6	0	6	0	0	0	0	0	0	0	0	0	0	0		
RKB	0	1	0	2	0	2	0	12	0	2	0	2	0	0	0	0	0	0	0	0	0	0		
RKE	0	5	0	4	0	4	0	10	0	4	0	4	2	0	2	0	2	0	2	0	2	0		
RL4	0	3	0	3	0	3	0	4	0	2	0	2	0	4	0	3	0	0	0	0	0	2		
RLN	11	0	12	0	10	0	12	0	12	0	12	0	0	5	0	3	0	1	0	1	0	1		

Trust	Total Incidents												DS Incidents											
	Poisson GLMM		NB1 GLMM		NB2 GLMM		Poisson GAM		NB2 GAM		Random Forest		Poisson GAM (full)		NB2 GAM (full)		Poisson GLMM (parameterisation 1)		NB1 GLMM (parameterisation 1)		NB1 GLMM (parameterisation 2)			
	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-		
RLQ	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5	0	5	0	5	0	5		
RLT	0	5	0	6	0	4	0	10	0	1	0	1	0	0	0	0	1	0	1	0	0	0		
RM1	6	0	5	0	6	0	0	0	6	0	6	0	0	1	0	2	0	2	0	2	0	2		
RM2	0	0	2	0	1	0	0	4	2	0	2	0	0	1	0	1	0	1	0	1	0	1		
RM3	0	1	2	0	0	1	7	0	0	3	0	3	0	3	0	3	0	3	0	3	0	3		
RMC	0	0	4	0	0	0	7	0	4	0	4	0	0	3	0	2	0	3	0	3	0	2		
RMP	0	0	0	0	2	0	0	0	4	0	4	0	0	4	0	4	0	5	0	5	0	4		
RN3	0	3	0	6	0	3	0	5	0	4	0	4	0	1	0	1	0	0	0	0	0	0		
RN5	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	1	0	0		
RN7	0	0	1	0	0	0	0	0	1	0	1	0	0	6	0	6	0	6	0	6	0	6		
RNA	0	6	0	3	0	4	0	6	0	6	0	6	0	0	0	0	0	0	0	0	0	0		
RNL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
RNQ	0	3	0	4	0	2	0	5	0	4	0	4	0	1	0	1	0	1	0	1	0	1		
RNS	0	4	0	3	0	3	0	3	0	5	0	5	0	3	0	3	0	2	0	2	0	2		
RNZ	0	2	0	0	0	0	2	0	0	0	0	0	0	3	0	3	0	3	0	3	0	3		
RP5	2	0	2	0	2	0	6	0	2	0	2	0	3	1	3	1	3	0	3	0	3	0		
RPA	0	9	0	9	0	9	0	9	0	9	0	9	0	2	0	2	0	2	0	2	0	2		
RQ6	1	3	0	6	1	2	0	4	0	4	0	4	0	0	0	0	0	0	0	0	0	3		
RQ8	0	9	0	9	0	9	0	12	0	10	0	10	0	3	0	3	0	0	0	0	0	0		
RQW	1	0	1	0	1	0	0	0	2	0	2	0	0	2	0	2	0	2	0	3	0	2		
RQX	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	3	0	1	0	1	0	1		
RR1	0	5	0	4	0	9	3	0	0	2	0	2	4	0	3	0	5	0	5	0	4	0		
RR7	0	6	0	6	0	5	0	7	0	3	0	3	0	0	0	0	0	0	0	0	0	0		
RR8	0	5	0	0	0	7	5	0	1	0	1	0	0	2	0	3	0	2	0	2	0	4		
RRF	0	4	1	0	0	4	1	0	0	0	0	0	3	0	3	0	3	0	3	0	3	0		
RRK	10	0	12	0	11	0	12	0	12	0	12	0	0	5	0	5	0	6	0	7	0	5		
RRV	0	7	0	2	0	4	4	0	0	6	0	6	0	0	0	0	0	0	0	0	0	0		
RTD	0	11	0	5	0	12	0	8	0	8	0	8	1	0	1	0	1	0	1	0	0	0		
RTE	0	0	0	0	0	0	3	0	0	0	0	0	0	3	0	4	0	1	0	1	0	1		
RTF	6	0	5	0	6	0	7	0	8	0	8	0	0	2	0	2	0	2	0	2	0	2		
RTG	0	1	0	4	0	1	0	10	0	2	0	2	0	2	0	2	0	2	0	2	0	2		
RTH	1	0	8	0	0	0	5	0	6	0	6	0	0	1	0	0	0	0	1	0	0	0		
RTK	0	0	0	2	0	0	0	5	0	0	0	0	0	2	0	2	0	2	0	2	0	1		
RTP	0	4	0	4	0	4	0	6	0	5	0	5	0	1	0	1	0	1	0	1	0	1		
RTR	0	4	0	2	0	3	0	8	0	1	0	1	0	0	0	0	0	0	0	0	0	0		
RTX	0	0	1	0	0	0	0	0	1	0	1	0	0	3	0	3	0	2	0	3	0	2		
RVJ	0	1	0	0	0	2	12	0	0	0	0	0	3	0	1	0	2	0	2	0	1	0		
RVR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
RVV	0	0	0	2	0	1	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	1		
RVW	0	3	0	3	0	2	0	4	0	3	0	3	0	5	0	5	0	5	0	5	0	5		
RVY	0	3	0	0	0	3	0	0	0	0	0	0	0	2	0	2	0	2	0	2	0	2		
RW6	0	2	0	4	0	5	1	0	0	4	0	4	4	0	4	0	8	0	8	0	4	0		
RWA	0	4	0	1	0	5	0	4	0	2	0	2	1	0	1	0	1	0	1	0	0	0		
RWD	0	5	0	5	0	5	0	7	0	5	0	5	10	0	8	0	11	0	11	0	11	0		
RWE	1	0	1	0	0	0	1	0	7	0	7	0	0	4	0	4	0	4	0	4	0	6		
RWF	0	6	0	11	0	6	0	12	0	10	0	10	0	0	0	0	2	0	2	0	2	0		
RWG	4	0	4	0	3	0	2	0	0	0	0	0	0	2	0	2	0	1	0	1	0	1		
RWH	0	4	0	5	0	4	0	10	0	5	0	5	0	1	0	1	0	1	0	1	0	1		
RWJ	6	0	11	0	7	0	9	0	7	0	7	0	3	1	4	1	1	1	1	3	1	1		
RWP	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	2	0	1	0	1	0	1		
RWW	0	0	0	0	0	0	0	5	0	0	0	0	1	0	1	0	1	0	1	0	1	0		
RWY	4	0	5	1	3	1	9	0	2	1	2	1	4	0	4	0	4	0	4	0	4	0		
RX1	0	3	1	0	0	6	2	0	2	0	2	0	0	3	0	3	0	1	0	1	0	3		
RXC	0	0	0	0	0	0	3	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0		
RXF	6	0	4	0	6	0	10	0	6	0	6	0	0	2	0	2	0	0	0	0	0	0		
RXH	0	1	0	7	0	2	0	12	0	5	0	5	0	4	0	4	0	2	0	2	0	4		
RXK	6	0	3	0	8	0	3	0	3	0	3	0	0	2	0	1	0	1	0	1	0	1		
RXL	11	0	12	0	11	0	12	0	12	0	12	0	0	6	0	6	0	7	0	7	0	5		
RXN	0	1	7	0	0	0	12	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0		
RXP	1	0	0	3	1	0	0	0	0	1	0	1	0	4	0	4	1	0	1	0	0	2		
RXQ	2	1	4	0	2	1	3	1	2	1	2	1	0	1	0	1	0	0	0	0	0	0		
RXR	0	0	5	0	0	0	0	0	2	0	2	0	0	2	0	3	0	2	0	2	0	2		
RXW	0	4	0	2	0	4	0	1	0	5	0	5	0	1	0	1	0	1	0	1	0	1		
RYJ	0	0	0	0	0	0	0	6	0	0	0	0	0	3	0	4	0	2	0	2	0	5		
RYR	0	3	0	4	0	3	0	3	0	4	0	4	0	2	0	2	0	1	0	1	0	1		
Total	156	315	221	324	166	290	255	444	214	280	214	280	64	255	60	253	79	208	77	212	72	219		

C.7 Marginal CUSUM alerts for organisations from NRLS-HES models for 2015/16.

Row are organisations, and columns represent number of alerts for increasing (=) and decreasing (-) weights, grouped by total or DS incidents.

Appendix D: Interactive module development process flow chart

