The Secondary Use of Longitudinal Critical Care Data

Dr Edward S. Palmer

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Division of Medicine

University College London

June 23, 2021

I, Dr Edward S. Palmer, 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 work.

A thesis is a collaborative endeavor. The role of collaborators, along with details of my personal involvement are detailed below.

Chapters 3 and 4

The review of both the CC-HIC data model (chapter 3) and the quality of data it contains (chapter 4) are all my own work, with supervision from Roma Klapaukh, Steve Harris and Mervyn Singer. The inspectEHR and wranglEHR software packages were written in their entirety by myself, with supervision from Roma Klapaukh.

This research was conducted using National Institute for Health Research (NIHR) Health Informatics Collaborative (HIC) data resources [1], representing a collaboration with the Biomedical Research Centres based at: University College Hospitals National Health Service (NHS) Foundation Trust and University College London (UCL), Imperial College Healthcare NHS Trust and Imperial College London, Cambridge University Hospitals NHS Foundation Trust and the University of Cambridge, Oxford University Hospitals NHS Foundation Trust and the University of Oxford, and Guy's and St. Thomas' NHS Foundation Trust and King's College London. The CC-HIC data model and the governance arrangements for data transfer were developed prior to my involvement and included support from Jonathon Cooper, Nicola Cooper, Charles Crichton, Jim Davies, Ben Glampson, James Hetherington, Finola Higgins, Shaun Hyett, Leonardo de Jongh, Hasmita Maisuria, Abdul Mulla, Theresa Noble, Dimitri Papadimitriou, David Perez-Suarez, Angela Poland, Luis Romao, Sinan Shi, Andrew Tsui, Kinga Varnai, James Welch, Ray Wells and Kerrie Woods.

Chapter 5: Cumulative Exposure to Excess Oxygen

I designed the study outline, imported data from the CC-HIC data model into the CC-HIC research database, evaluated its quality, extracted data suitable for analysis and performed the statistical analysis. Research outputs were co-authored with Ben Post, Roma Klapaukh, Giampiero Marra, Niall MacCallum, Dave Brealey, Ari Ercole, Andy Jones, Simon Ashworth, Pete Watkinson, Richard Beale, Steve Brett, Duncan Young, Claire Black, Aasiyah Rashan, and supervised by Dan Martin, Steve Harris and Mervyn Singer.

Chapter 6: Physiological Morphologies in Sepsis

I designed the study outline, imported data from the CC-HIC data model into the CC-HIC research database, evaluated its quality, extracted data suitable for analysis and performed the statistical analysis. This research was supervised by Roma Klapaukh, Steve Harris and Mervyn Singer.

Chapter 7

Decoding COVID-19 (DECOVID) [2] is introduced as future work. I co-authored the DECOVID data model with Roma Klapaukh as an extension to the The Observational Health Data Sciences and Informatics (OHDSI) common data model (version 5.3.1) and I wrote the DECOVID data specification. DECOVID is a multi-centre research consortium between University Hospitals Birmingham, University of Birmingham, University College London Hospitals, UCL and The Alan Turing Institute. DECOVID comprises a large and diverse faculty, of which I am a member representing University College Hospital London, and co-leading in the informatics stream alongside Ann-Marie Mallon and Suzy Gallier. A full list of the consortium members can be found at www.decovid.org.

Abstract

Aims

To examine the strengths and limitations of a novel United Kingdom (UK) critical care data resource that repurposes routinely collected physiological data for research. Exemplar clinical research studies will be developed to explore the unique longitudinal nature of the resource.

Objectives

- To evaluate the suitability of the National Institute for Health Research (NIHR) Critical Care theme of the Health Informatics Collaborative (CC-HIC) data model as a representation of the Electronic Health Record (EHR) for secondary research use.
- To conduct a data quality evaluation of data stored within the CC-HIC research database.
- To use the CC-HIC research database to conduct two clinical research studies that make use of the longitudinal data supported by the CC-HIC:
  - The association between cumulative exposure to excess oxygen and outcomes in the critically ill.
  - The association between different morphologies of longitudinal physiology—in particular organ dysfunction—and outcomes in sepsis.

The CC-HIC

The EHR is now routinely used for the delivery of patient care throughout the United Kingdom (UK). This has presented the opportunity to learn from a large volume of routinely collected data. The CC-HIC data model represents 255 distinct clinical concepts including demographics, outcomes and granular longitudinal physiology. This model is used to harmonise EHR data of 12 contributing Intensive Care Units (ICUs). This thesis evaluates the suitability of the CC-HIC data model in this role and the quality of data within. While representing an important first step in this field, the CC-HIC data model lacks the necessary normalisation and semantic expressivity to excel in this role. The quality of the CC-HIC research database was variable between contributing sites. High levels of missing data, missing meta-data, non-standardised units and temporal drop out of submitted data are amongst the most challenging features to tackle. It is the principal finding of this thesis that the CC-HIC should transition towards implementing internationally agreed standards for interoperability.

Exemplar Clinical Studies

Two exemplar studies are presented, each designed to make use of the longitudinal data made available by the CC-HIC and address domains that are both contemporaneous and of importance to the critical care community.

Exposure to Excess Oxygen

Longitudinal data from the CC-HIC cohort were used to explore the association between the cumulative exposure to excess oxygen and outcomes in the critically ill. A small (likely less than 1% absolute risk reduction) dose-independent association was found between exposure to excess oxygen and mortality. The lack of dose-dependency challenges a causal interpretation of these findings.

Physiological Morphologies in Sepsis

The joint modelling paradigm was applied to explore the different longitudinal profiles of organ failure in sepsis, while accounting for informative censoring from patient death. The rate of change of organ failure was found to play a more significant

role in outcomes than the absolute value of organ failure at a given moment. This has important implications for how the critical care community views the evolution of physiology in sepsis.

DECOVID

The Decoding COVID-19 (DECOVID) project is presented as future work. DECOVID is a collaborative data sharing project that pools clinical data from two large NHS trusts in England. Many of the lessons learnt from the prior work with the CC-HIC fed into the development of the DECOVID data model and its quality evaluation.

Impact Statement

This thesis conducts a thorough investigation into the CC-HIC, which is one of the first attempts to share granular EHR data for research between multiple NHS trusts in the UK. The NHS is ideally placed to repurpose a vast network of routinely collected patient data making it available for secondary use research. This is now a strategic priority for the UK government, who have placed digitising the NHS, sharing of data and interoperability as tent poles of their healthcare policy [3]. The importance of data sharing has been brought into sharp focus by the COVID-19 pandemic, where the scale of the problem highlighted the benefits of extracting data from the EHR where possible, instead of using conventional data collection by hand. The findings in this thesis will be invaluable in supporting both ongoing research within the CC-HIC and other initiatives as these become more common. Having been fully characterised by this thesis, research within the CC-HIC will be catalysed, since studies can now responsibly be conducted against this important resource. The software contributions in particular are now part of the embedded workflow of clinician scientists working with the CC-HIC. This software helps remove commonly error prone parts of the data science process whilst also drawing attention to areas of concern so that they can be explicitly addressed. The next generation of the CC-HIC data resource is currently under development and follows directly from the findings and recommendations of this thesis. The DECOVID data resource has been the direct beneficiary of the learning from this thesis. The DECOVID data resource was developed *de novo* as a modularised instance of the OHDSI data model. This approach encourages the transfer of high quality data by focusing on close feedback between contributing sites and a cautious growth of the

research database over time. The DECOVID platform has the potential to help us understand numerous facets of COVID-19, including important questions relating to ventilation strategy in severe COVID-19 pneumonia, which are now ongoing. This rapid platform development was only possible in light of the deep understanding of this process that was afforded by the pioneering work at the CC-HIC.

Oxygen is a routine treatment used throughout global healthcare. The life saving properties of oxygen are well documented, though the potential harms of excess oxygen administration are in question. The findings from the clinical aspects of this thesis highlight this potential harm and have been disseminated into the critical care community through publication [4] and were presented at the Intensive Care Society State of the Art meeting in 2020. If these associations are ultimately proven to be causal through future experimental work, the proposed effect size could translate to a substantial reduction in overall mortality, given the ubiquity of oxygen administration in healthcare. Major randomised controlled trials are currently underway to test this theory. Interestingly, given their size, these studies will rely on automated data capture from the EHR to deliver findings in a cost effective manner.

Sepsis is commonly encountered in the ICU, resulting in both high mortality and morbidity. As a heterogeneous syndrome, it has been notoriously difficult to find successful therapies in sepsis. The findings presenting in this thesis provide deeper insights into the dynamics of physiology that have undergone limited prior investigation. These findings could influence how patients are recruited into clinical trials in sepsis. The findings support that future trial enrolment should be enriched under dynamic recruitment criteria. This is an exciting and novel area that could stem directly from this research.

This thesis makes the following direct research contributions:

- Appraisal of the CC-HIC data model (Chapter 3)
- Data quality evaluation of the CC-HIC research database (Chapter 4)
- Investigation into the association between the cumulative exposure to excess oxygenation and outcomes in critical care (Chapter 5 and publication [4])
- The association between different longitudinal morphologies of organ dys-

function and outcomes in sepsis (Chapter 6)

This thesis makes the following software contributions (Chapter 4):

- The wranglEHR [5] package for R; Standardised data extraction from the CC-HIC research database.
- The inspectEHR [6] package for R; Standardised data quality evaluation of the CC-HIC research database.

Acknowledgements

A PhD is as much a journey as it is a body of work. I am fortunate to have taken many friends, family and colleagues with me on this journey. The world has changed profoundly these past few years. There has been so much sorrow for so many during the pandemic. I have been fortunate. I leave this PhD with two wonderful daughters, Coral and Astrid, who fill my life with joy every day. Thank you Katrina, for standing with me and showing me endless patience.

My supervisors, Mervyn, Steve and Roma. Mervyn, thank you for your constant positivity and enthusiasm, it is infectious. And for continuing to push me to do my best. Steve, without whom I would not be here writing this at all. Thank you for the guidance and friendship, and for starting me out on the journey. Roma, who has taught me so much. I am richer for knowing you. Thank you for listening and dispensing sage advice at regular intervals.

Mum and Dad for always supporting me in every endeavour. David, my brother. There is no-one I hold in higher regard. I hope this makes you proud. Charlotte, my sister. Whos creativity, enthusiasm and talents are unmatched. I hope I have channelled you in these pages.

Mitch and Mary for welcoming me into your family with open arms (and for reading this!)

The consultants at The London Clinic, John, Geoff, Sara, Niall, Dan and Matt. It has been the greatest privilege working alongside you these past years. All the fellows at The London Clinic who shared the journey with me between academia and clinical medicine.

The friends I have made along the way; Danny Wong, Ben Post, Helen

Mckenna, and Claire Black. All the members of the Harris and Singer labs.

I would also like to thank Dimitris Rizopoulos, Giampiero Marra, Qiuju Li, Kirstie Whittacre, Wai Keong Wong, Dave Brealey and all the members of the DECOVID and CC-HIC teams.

Finally, to all those patients who contributed their data to this research, to the patients who have experienced critical illness, and to those we have lost.

"…the most elementary and valuable statement in science, the beginning of wisdom, is, 'I do not know.'"

— Lt. Cmdr. Data (2365)

Contents

List of Figures

List of Tables

Acronyms

ABG arterial blood gas.

AIC Akaike information criterion.

APACHE II Acute Physiology And Chronic Health Evaluation II.

ARDS Acute Respiratory Distress Syndrome.

AUROC area under the receiver operating characteristic.

BIC Bayesian information criterion.

CC-HIC Critical Care theme of the Health Informatics Collaborative.

CDF Cumulative Distribution Function.

CDM clinical data model, or "common data model".

CO cardiac output.

COPD chronic obstructive pulmonary disease.

COVID-19 Coronavirus disease 2019.

CPR cardiopulmonary resuscitation.

CRP C-reactive protein.

CSV Comma Separated Value.

DAMPs Danger-Associated Molecular Patterns.

DECOVID Decoding COVID-19.

DQC Data Quality Collaborative.

DSH data safehaven.

DTE differential treatment effect.

EAV Entity Attribute Value.

EHR Electronic Health Record.

ETL Extract Transform Load.

F_IO_2 fraction of inspired O_2.

FHIR Fast Healthcare Interoperability Resources.

GAM generalised additive model.

GCS Glasgow Coma Score.

HES Hospital Episode Statistics.

HIC Health Informatics Collaborative.

HTE heterogeneity of treatment effect.

i2b2 Integrating Biology and the Bedside.

ICNARC Intensive Care National Audit and Research Centre.

ICU Intensive Care Unit.

IQR Interquartile Range.

KS Kolmogorov-Smirnov.

LOCF last one carried forward.

LODS logistic organ dysfunction score.

LOESS locally estimated scatterplot smoothing.

LSHTM London School of Hygiene and Tropical Medicine.

MAR missing at random.

MCAR missing completely at random.

MCID Minimum Clinically Important Difference.

MIMIC The Medical Information Mart for Intensive Care.

MIT Massachusetts Institute of Technology.

MNAR missing not at random.

NHS National Health Service.

NIHR National Institute for Health Research.

OHDSI The Observational Health Data Sciences and Informatics.

OMOP Observational Medical Outcomes Partnership.

P_aO_2 partial pressure of arterial oxygen.

PAMPs Pathogen-Associated Molecular Patterns.

PCA patient controlled analgesia.

PCORnet Patient Centred Outcomes Research Network.

PEEP positive end expiratory pressure.

PR precision recall.

QA quality assurance.

QC quality control.

QE quality evaluation.

RCT randomised controlled trial.

RM risk magnification.

ROC receiver operator characteristic.

RRT renal replacement therapy.

SAG Scientific Advisory Group.

SNOMED systematised nomenclature of medicine.

SOFA sequential organ failure assessment.

SpO_2 peripheral oxygen saturation.

SQL Structured Query Language.

UCL University College London.

UK United Kingdom.

UML Unified Modeling Language.

US United States.

XML Extensible Markup Language.

XSD XML Schema Definition.

YAML YAML Ain't Markup Language.

Glossary

episode A continuous period of level 2/3 care, within a single physical location. This is the base unit of the CC-HIC database.

index episode the first episode for a patient seen in the database.

level 1 care ward level care.

level 2 care high dependency unit level of care, typically characterised by single organ failure.

level 3 care intensive care unit level of care, typically characterised by multiple organ failure.

spell A continuous period of level 2/3 care, regardless of physical location. Contains multiple episodes.

Chapter 1

Introduction

Modernising digital policies of the past 10 years within the United Kingdom (UK) has led to widespread adoption of the Electronic Health Record (EHR). Coupled with increasing levels of clinical device integration, routinely collected healthcare data now include a wide gamut of physiological and treatment variables from across the patient journey. The availability of these high resolution longitudinal data resources presents new opportunities for research in areas that would have only recently been impossible.

The UK National Institute for Health Research (NIHR) Critical Care theme of the Health Informatics Collaborative (CC-HIC) has taken some early steps in the UK into sharing routinely collected granular and identifiable physiological data from secondary care. The CC-HIC is the result of a call from the NIHR to operationalise routinely collected healthcare care data for research purposes. The CC-HIC has already overcome the first, and arguably most challenging element of such an endeavour; sharing granular and identifiable clinical data. The CC-HIC pools critical care data from hospitals within five UK Biomedical Research Centres. The Electronic Health Records (EHRs) contributing to the CC-HIC store and represent clinical data in diverse ways. A major technical challenge in building the research data pipeline has therefore been the harmonisation of healthcare data, semantic interoperability and clinical data modelling, while operating under the restrictions of a security hardened research environment. These challenges—and their respective solutions as applied to the CC-HIC—are reviewed in Chapter 3 (page 87).

As a previously untested resource, this thesis formally evaluates the quality of data stored in the CC-HIC research database. This evaluation, alongside its software implementation, is presented in Chapter 4 (page 107). This evaluation provides a roadmap for the inferences that follow, allowing the navigation of potential deficiencies, so that clinical research can proceed safely.

The granular longitudinal data offered by the CC-HIC is a unique feature in the UK. Research in this field has generally been limited to summary physiology from the first 24 hours of critical illness. Two exemplar studies were designed and conducted that were able to take full advantage of this unique longitudinal data: the impact of cumulative exposure to oxygen in critical illness, and the role of physiological morphologies in sepsis. These topics were chosen because they are of current importance to the critical care community and require longitudinal data to be addressed.

1.1 Cumulative Exposure to Excess Oxygen

Oxygen is an ubiquitous treatment for hypoxaemia in critical care. This ubiquity—and a perception of safety—means that patients are often exposed to quantities of oxygen far in excess of their physiological requirements for long periods of time. The harms of high levels of oxygen exposure are well described in mammalian models and are not in question. It is currently unknown as to whether or not the lower levels of exposure to oxygen that are commonly seen in clinical practice are harmful. In general, lower boundaries of oxygenation are targeted in intensive care. This results in a tendency towards over, rather than under, oxygenation. Several small randomised controlled trials have failed to provide a definitive answer to the question of potential oxygen toxicity, and much larger randomised controlled trials are now ongoing. There remains an opportunity to use the large cohort of longitudinal data available in the CC-HIC to help contribute to knowledge in this field. Chapter 5 (page 157) presents this exemplar study, focusing on the cumulative exposure of oxygen and its potential association with outcomes in critical care.

1.2 Physiological Morphologies in Sepsis

Sepsis—infection complicated by life-threatening organ dysfunction—is a global health concern. In England, Wales and Northern Ireland alone there are approximately 40,000 admissions per year for presumed sepsis to Intensive Care Units (ICUs), representing around a third of all ICU throughput. Sepsis is a highly heterogeneous syndrome, and as such has presented a challenge to find efficacious therapies. Identifying biological subgroups of this disease, so called "phenotypes" has thus been highlighted as a key research priority.

Treatment for sepsis is often framed in terms of timing; "early" treatment is considered better, while "late" treatment is potentially deleterious. However, as a prevalent—as opposed to incident—disease, "early" and "late" are terms rooted in the administrative time frames of healthcare delivery, rather than the biology of the underlying disease. Since we rarely know the true onset time of sepsis, it is important to consider the question, "early relative to what?" Despite the widespread use and acceptance of these terms, they may poorly explain the physiological heterogeneity of sepsis. In lieu of a biomarker that reliably describes sepsis on the biological time-scale, reconsidering the problem as different parameterisations of disease physiology—so called disease "morphologies"—may provide actionable insights. Is the patient improving or deteriorating? At what rate? And how does this influence survival? What is the cumulative exposure to organ dysfunction? And at what point (if any) does a patient's acute physiology stop being predictive of their outcome? Intuitively, clinicians apply many of these concepts at the bedside everyday. No blood test or physiological response is viewed in isolation, but always contextualised to the results that came before. A rigorous investigation connecting candidate biomarker morphologies to patient outcomes is currently lacking. Establishing this link is a required step if we are to consider patterns of acute patient physiology as being representative of different disease states. Chapter 6 (page 195) presents this exemplar study, focussing on different morphological representations of longitudinal physiology and their potential association with patient outcomes.

1.3 Methodological Challenges

A particular methodological challenge to overcome when modelling both excess oxygen and sepsis physiology is the endemic presence of informatively missing data in critical care. For inferences to be valid and generalisable, methods must be able to address bias encountered when analysing these cohorts. Data can be considered informatively missing when an *unmeasured* property of the patient gives rise to missing data. In critical illness these informatively missing data patterns occur when patients die, removing them from the analysis. This pattern of missing data can introduce biases to inferences made over physiological data, yet it is uncommon for this problem to be taken into account. To address this issue the joint modelling paradigm has been applied. This principled modelling approach is well suited to addressing the aforementioned bias. A recent number of methodological contributions to the field of joint models has enabled their use in the applied context in which the present research resides.

1.4 Learning Lessons

The experience of this research has directly contributed to a formal set of recommendations that have facilitated a second generation UK data sharing platform; Decoding COVID-19 (DECOVID). The DECOVID platform is discussed in Chapter 7 (page 237) as ongoing and future work. DECOVID is infused with lessons learnt from CC-HIC, removing many of the limitations that may restrict the scope of research questions that can currently be answered by the CC-HIC.

Chapter 2

Background

This thesis finds itself within the nexus of critical care medicine, healthcare data engineering and applied statistics. This chapter provides an overview of salient topics to provide the background and motivation for the data engineering and clinical research questions that follow. Sections 2.1 to 2.3 introduce clinical data models, controlled clinical terminologies and the data transfer processes of the CC-HIC. Section 2.4 introduces a nomenclature for the description of longitudinal biomarker morphologies. Section 2.5 provides the scientific background for the potentially deleterious effects of exposure to excess oxygen. Section 2.6 provides this background for sepsis and discusses the relevance of differential treatment effects. Sections 2.7 to 2.11 provide an overview of the statistical methods used in this thesis. This includes linear mixed effects models, survival models and joint models.

2.1 Clinical Data Models

Healthcare data are complex, messy and challenging to work with [7, 8, 9]. EHRs have developed organically over many years to accommodate this complexity [10]. This has resulted in competing platforms from private, public and open source origins providing varied solutions to the problem. Healthcare data—particularly across organisations—are therefore stored in myriad formats and unique database schemas. Under such conditions, collaboratively bringing data together is a challenge. A number of solutions have been proposed to allow these different systems to communicate and share data. These range from the humble disease registry to fully

fledged clinical EHR. Somewhere along this spectrum is the clinical data model, or "common data model" (CDM). The CDM is a blueprint that describes how clinical data should be standardised and represented. A number of CDMs have gained popularity to support multi-site comparative effectiveness research. Examples include:

- Intensive Care Unit (ICU) specific data models:
  - Intensive Care National Audit and Research Centre (ICNARC) [11].
  - The Medical Information Mart for Intensive Care (MIMIC) III and IV [12, 13].
  - CC-HIC [14].
- General purpose data models:
  - Integrating Biology and the Bedside (i2b2) [15, 16].
  - The Observational Health Data Sciences and Informatics (OHDSI) (pronounced "Odyssey") [17, 18][1].
  - Sentinel [19, 20].
  - Patient Centred Outcomes Research Network (PCORnet) [21].
  - The generalized data model for clinical research [22].

A CDM provides the blueprint for representing healthcare data, however when populated with data, the data model can still lack universal semantic meaning. Controlled clinical terminologies fulfil this role.

2.2 Controlled Clinical Terminologies

Controlled clinical terminologies are designed to provide the semantics necessary to define clinical concepts without ambiguity. By analogy, if the CDM is a blueprint, then a controlled clinical terminology is the language it is communicated in. Both builder and architect must understand the same language if the blueprint is to be correctly implemented. It is a recipe for failure if the architect writes in English and feet, and the builder reads in French and is accustomed to metres. Examples of commonly used terminologies are provided in table 2.1.

[1]This data model is known more commonly as the "OMOP" (Observational Medical Outcomes Partnership) data model which was the name given to an earlier version.

Vocabulary	Usage	Example code	Example description	Special features
SNOMED-CT	all healthcare concepts	53084003	bacterial pneumonia (disorder)	networked, allows "post co-ordination" modifier codes for very specific representations
LOINC	primarily laboratory findings	LA7465-3	pneumonia	
ICD-10	medical diagnosis	J15.9	bacterial pneumonia, unspecified	
ICNARC	intensive care diagnosis	2.1.4.27.1	bacterial pneumonia	hierarchical system
Read	primary care diagnosis	H22z.00	bacterial pneumonia NOS	
Dm+d	medicines (UK)	372687004	amoxicillin	
RxNorm	medicines (USA)	723	amoxicillin	
UCUM	units of measure	mg	milligram	
Athena	compendium of vocabularies	-	-	contains all above codes with unique "Athena codes". Connects different terminologies.

Table 2.1: Commonly used vocabularies are listed with examples for pneumonia (diagnosis), amoxicillin (drug) and milligrams (units) depending on the primary role of the vocabulary.

2.3 Critical Care Health Informatics Collaborative

The NIHR Health Informatics Collaborative (HIC) is a project that brings together National Health Service (NHS) trusts and partner universities to share routinely collected patient data for secondary use for research. Five clinical areas formed the founding themes of the HIC, each with a host organisation to provide leadership (table 2.2). The express aims of the HIC are:

1. "to support the establishment and maintenance of catalogued, comparable, comprehensive flows of patient data at each trust."
2. "to create a governance framework for data sharing and re-use across the trusts and partner organisations ..."
3. to conduct "a number of exemplar research studies, one in each of the established therapeutic areas." [23]

Theme subject	Host organisation
Critical care	University College London
Cardiovascular medicine	Imperial College London
Ovarian cancer	Cambridge University
Renal transplantation	King's Health Partners
Viral hepatitis	University of Oxford

Table 2.2: Founding themes of the NIHR Health Informatics Collaborative.

The CC-HIC [14], is a multi-centre research project, pooling static and time series data on critical care patients from the 12 intensive care units (ICUs) within hospitals partnered to the five biomedical research centres listed in table 2.2. The CC-HIC contains up-to hourly data on bedside monitoring, and retains identifiable data for the explicit purpose of linkage to other external data resources. 255 distinct data concepts comprise the base data model[2]. Illustrative examples include:

- patient characteristics on admission to an ICU:
  - date of birth.
  - sex.

[2]A full list of this specification with accompanying metadata is provided in appendix table A.1.

  - Acute Physiology And Chronic Health Evaluation II (APACHE II) score[3].
  - admitting diagnosis.
- longitudinal physiology (up to hourly):
  - vital signs.
  - biochemistry.
- longitudinal treatments:
  - antimicrobials (pharmacological).
  - vasoactive infusions (pharmacological).
  - respiratory support (non-pharmacological).
- patient outcomes:
  - survival at the end of an ICU episode.
  - survival at hospital discharge.

A useful UK comparison can be made with the Intensive Care National Audit and Research Centre (ICNARC) case mix programme [24]. Spanning more than 25 years, ICNARC collects a comprehensive summary of outcomes, demographics and physiological data from the first 24 hours of each admission to an ICU [25, 26, 27, 28, 29]. The ICNARC data collection is vast, covering almost all NHS adult ICUs in the UK bar Scotland. The ICNARC data collection is unequivocally of high quality, however this does come at a potentially considerable expense. Most data must be (at least partially) hand curated prior to submission, necessitating the presence of ICNARC clerks in most ICUs in the UK. This places natural restrictions on the expansion of ICNARC to cover the rich longitudinal physiology that is available within the EHR for many critically unwell patients. As a result, there is a need for a dedicated set of *automated* data extraction methods, and an accompanying CDM and research platform that can support the analysis of such data. The CC-HIC has taken significant steps to this end.

[3]The APACHE II score is an ICU risk scoring system that calculates the risk of death based upon physiology and chronic health status observed form the first 24 hours of an admission to the ICU.

2.3.1 The CC-HIC Data Pipeline

Data are extracted from each local site EHR, in many cases supplemented with data from local ICNARC files, and transformed to be represented in Extensible Markup Language (XML) format, aligning to the CC-HIC data model. XML is a mature language that underpins much of the data transfer processes that occur on the internet. XML excels at representing complex and nested data without ambiguity, and so is a natural fit for representing complex clinical data. An XML Schema Definition (XSD) provides an exacting specification to facilitate the writing of an XML document to ensure that it contains the right information, and presents it in the right way, so that it can be machine readable. These XML files are uploaded to the University College London (UCL) identifiable data safehaven (DSH). The DSH is a security hardened "walled garden" research environment that is compliant with both ISO27001:20131 certification [30] and the NHS Data Security and Protection Toolkit [31].

The version 1 pipeline[4] is illustrated in figure 2.1. This pipeline implemented the cleanEHR package for R [14]. cleanEHR served to extract data from submitted XML files and store them in a custom data object known as "ccData". Structurally, the ccData object was similar to the CC-HIC data model, and can largely be thought of as a representation of that data model in a format native to the R statistical programming language. The primary means through which analysts interacted with research data was to load the ccData object into working memory, and use the tools available as part of the cleanEHR package to reconfigure data into a rectangular format suitable for further analysis.

This approach to working with the CC-HIC data became obsolete as the size of the necessary data objects exceeded the working memory capacity that is typically available to an end user. A new XML parser and database were developed[5] to replace the cleanEHR [14] dependency, as illustrated in figure 2.2.

The new CC-HIC research database was developed in an episode centric for-

[4]This phase of the project was completed prior to my involvement and is discussed for completeness.

[5]This development phase coincided with my arrival to the CC-HIC project.

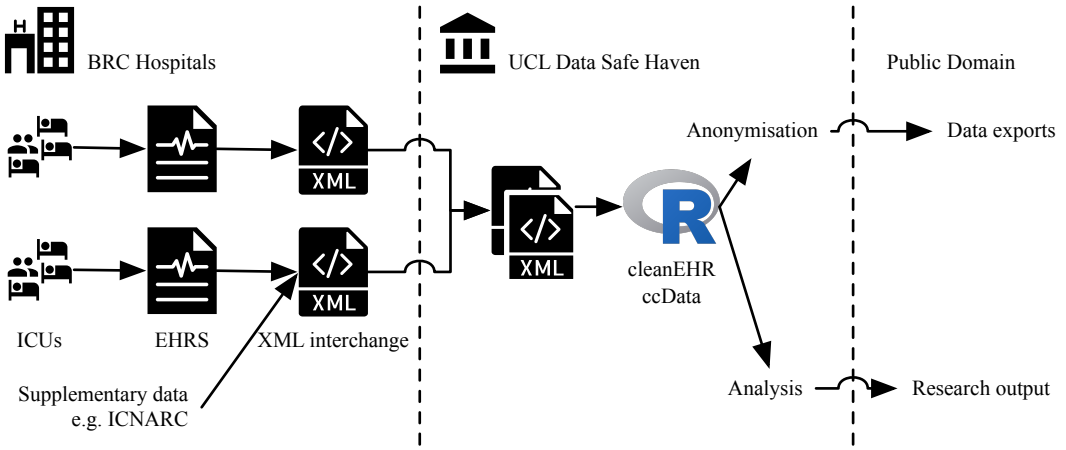


Figure 2.1: Version 1 data pipeline. Data flows from contributing sites and is warehoused in the UCL Data Safe Haven. Data flow into the analytic pipeline is via cleanEHR.

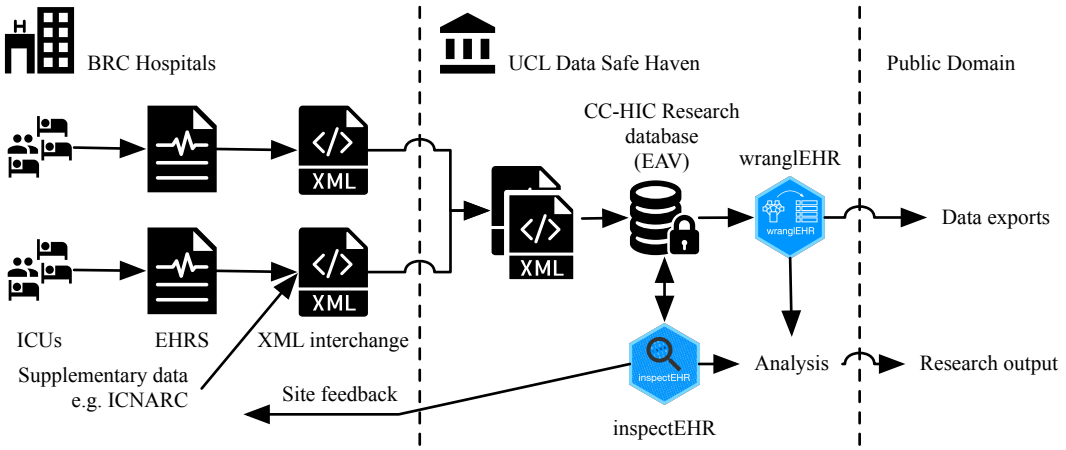


Figure 2.2: Version 2 data pipeline. Data flows from contributing sites and is warehoused in the UCL Data Safe Haven in the CC-HIC research database. Data flow into the analytic pipeline is optionally via wrangleEHR (standardised cohort extraction) which flexibly converts data from the EAV storage format to a rectangular format. Standardised data quality evaluation is performed by inspectEHR, which embeds quality metadata in the EAV database and provides feedback to each contributing site.

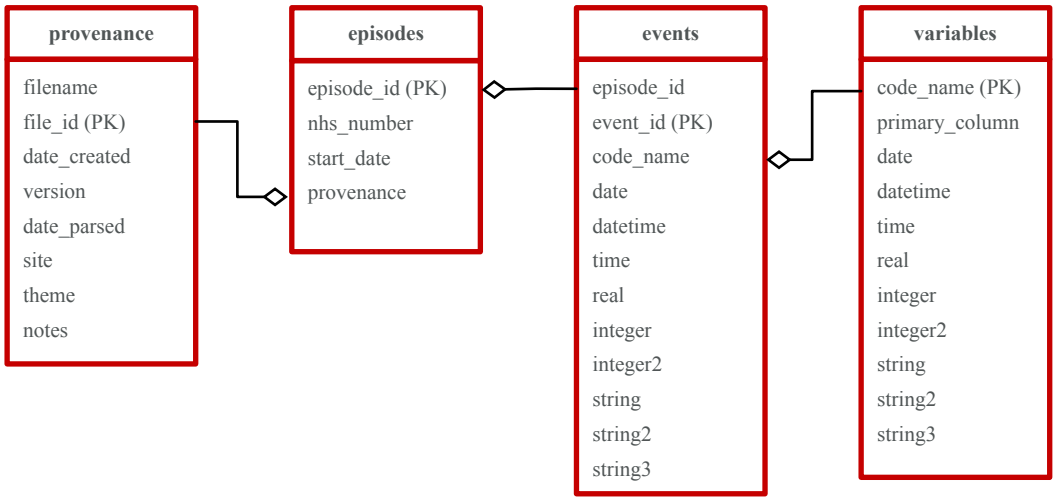


Figure 2.3: UML diagram of the CC-HIC database schema. PK = primary key.

mat, in order to accommodate the strong semantic representation of "episode" in the CC-HIC data model. This research database was designed as an Entity Attribute Value (EAV) style, as depicted in the Unified Modeling Language (UML) diagram in figure 2.3.

The hallmark of the EAV structure is a long central fact table. In the CC-HIC database this exists in the form of the "events" table. The events table accommodates the majority of the data concepts related to each episode. The transition to this research database came with many new advantages:

- analysts were free to use any general purpose approach for data extraction with which they are accustomed, without being reliant on a specific R package (or even the R language itself).
- the addition of new data concepts (although not easily supported by the underlying CC-HIC data model) would be trivial to implement at the database level.
- data best practices were promoted, as data type consistency was intrinsically enforced by database constraints.
- typical working memory limitations were removed. The database could be queried for the required cohort, rather than needing to store the entire dataset in working memory.

2.3.2 Legal & Ethical Basis for Use

A legal basis for transferring data was provided under section 251 of the National Health Service Act 2006 (Confidentiality Advisory Group reference 14/CAG/1001); this process sets aside the common duty of confidentiality in the UK. Ethics approval was granted by a Health Research Authority Research Ethics Committee (14/LO/1031). Legitimate Interest and substantial public interest provided the lawful basis for data processing under General Data Protection Regulations. The approvals listed above permit the CC-HIC to share identifiable data for the specific purpose of linkage to other data resources. This would include, for example, Hospital Episode Statistics (HES) [32]. The research agenda for this thesis was registered and approved by the CC-HIC scientific advisory group.

2.3.3 Patient & Public Involvement

A patient and public involvement session was conducted in April 2018, early in the course of the present research agenda. In this session key areas that were felt to be of particular concern or interest to patients were discussed, including:

- the principle of assumed consent.
- sharing of identifiable data.
- policies surrounding remuneration for access to data.

This session featured a video commissioned by the CC-HIC to help better inform the public on the goals and strategy behind the project [33].

An interesting outcome from the session was that many of the participants had assumed that data sharing of this nature was already taking place. On the whole, they felt that it would be important to conduct this type of data sharing. The main concerns were that data should be used responsibly to improve patient outcomes and experience. In general participants were happy for commercial and academic partners to have access to data, providing this was done securely as described during the session. It was emphasised that they would expect commercial partners to pay a commensurate fee for data access, the proceeds of which could go directly to patient care or back into the project itself to facilitate self sufficiency. Notable and

unambiguous concerns about the possibility of insurance companies gaining access to these data were raised.

2.3.4 Scientific Advisory Group

The ethics approvals that support the CC-HIC are designed to facilitate the CC-HIC as a research platform. As such, internal research applications do not, on the whole, require separate research ethics to be completed. This is supported by the Scientific Advisory Group (SAG). The role of the SAG is to evaluate research proposals with respect to:

- patient benefit.
- scientific value.
- appropriateness of methodology.
- information governance requirements.
- feasibility and workload of the CC-HIC data scientists.

The SAG comprises: one nominated person from each founding NHS Trust, one lay representative and one information governance representative. A minimum level of information governance training is required—typically NHS Data Security Awareness Level 1—for researchers to interact with the CC-HIC research database.

2.4 Longitudinal Exposures

In longitudinal studies, we may wish to understand the relationship between certain time-varying patient features and outcome. These features include those that are internal to the patient, for example a biomarker, or external, for example exposure to treatment. In both instances, analysis of the relationship is complicated by the necessity of a period of observation during which the intensity of the longitudinal feature can vary [34]. In order to provide a connection between exposure to treatment or patient biomarker, it is appropriate to construct a summary representation of both exposure duration and intensity. These summary representations are referred to as biomarker "morphologies".

2.4.1 A Nomenclature of Morphologies

In order to identify and discuss specific patterns of interest within longitudinal physiology, I shall introduce and explain key terms that define different morphologies of longitudinal patient data. These definitions are supported by figure 2.4 in which a hypothetical patient's biomarker is sampled over a 10 day period. The terms with reference to longitudinal patient physiology: severity, velocity, trajectory, cumulative exposure and weighted cumulative exposure are introduced.

The *severity* of disease is an instantaneous measure of the magnitude of disease. Severity is commonly what is actually measured on a patient to be stored in the EHR. The *velocity* of disease is the rate of change of disease severity; the "slope" or "gradient" of disease severity with respect to time. The velocity of disease is the rate at which a patient is getting better, or worse. This is unlikely to be measured directly for a given patient, but is easily determined from data. The *trajectory* of disease is the path of severity at a given moment. The trajectory represents the combination of severity and velocity. In much the same way as the trajectory of a ball being thrown through the air can be defined by a vector representing its position and direction of travel, so is the trajectory of disease. The *cumulative exposure* to disease is the area under disease severity when plotted against time. The cumulative exposure to disease includes the history of exposure up to and including the point of interest. It is a useful means to capture the full history of a biomarker, rather than

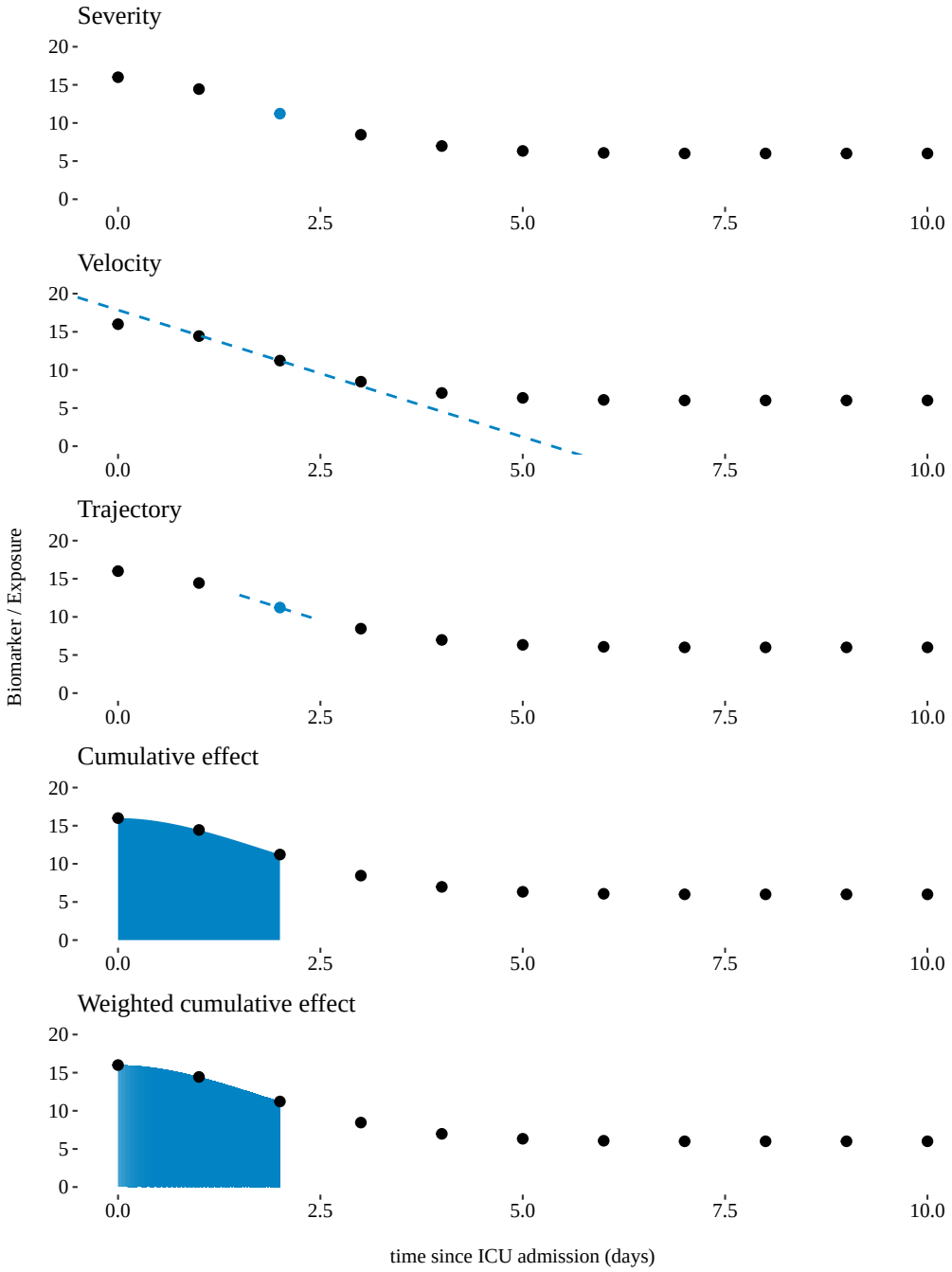


Figure 2.4: Prototypical morphologies of disease biomarkers. An illustrated patient is observed over 10 days with an evaluation on day 2. **Severity:** The severity of disease is represented by the blue observed point on day 2. **Velocity:** The velocity of disease (rate of change of disease severity) is represented by dashed line capturing the gradient of the biomarker on day 2. **Trajectory:** The disease trajectory is shown as a dashed blue line and point representing the simultaneous measure of disease severity and velocity on day 2. **Cumulative effect:** The cumulative effect of disease (area under disease severity) is represented by the blue area under the biomarker from day 0 up to day 2. **Weighted cumulative effect:** The weighted cumulative effect of disease (area under disease severity) is represented by the shaded blue area of varying intensity under the biomarker from day 0 up to day 2.

its current state. The *weighted cumulative exposure* to disease is the weighted area under disease severity when plotted against time. This is a useful morphology to consider if it is not reasonable to assume that recent exposure should carry the same weight as more distant exposure.

2.5 Exposure to Excess Oxygen

Supplementary oxygen is an ubiquitous treatment for hypoxaemia (low oxygen levels). Oxygen is so commonly used in healthcare that modern hospitals have oxygen physically piped through the building as if it were running water. Oxygen is necessary for the machinery of life to function, and so the lifesaving properties of oxygen are readily apparent. The familiarity and ease with which oxygen can be provided to a patient has led to readily apparent liberal oxygen use in critical care and healthcare in general [35, 36].

2.5.1 Historical Context

The potential harms from oxygen have long been recognised. As far back as 230 years ago, Lavoisier, while experimenting with oxygen on mammals, exposed guinea pigs to "l'air vital" (pure oxygen). He observed a high mortality in these animals that primarily arose from "une fièvre ardente" (high grade fever) and "maladie inflammatoire" (inflammatory disease) [37]. Moving two centuries forward, Haldane corroborated this finding, identifying that high levels of inspired oxygen often proved fatal [38]. Haldane recommended that we administer the lowest possible percentage of oxygen, and that its use be monitored to allow for titration to the patient's requirements [38]. The observation that high inspired fractions of oxygen lead to a profound inflammation of the lungs—so called Acute Respiratory Distress Syndrome (ARDS)—is universal in mammals [39]. In 1970, Barber and Hamilton, alternately assigned adult humans who had experienced brain death to either sustained 100% oxygen or normal air (21% oxygen)[6] [40]. Within a few days, the classical picture of ARDS arose in all those who received pure oxygen [40]. In healthy subjects, exposure to high inspired oxygen concentrations has shown to cause disruption to the alveolus (the functional anatomical unit of the lung) and release of biochemical mediators that are responsible for developing lung fibrosis [41].

[6] This study pre-dated the widespread use of randomisation as a means to identify causal mechanisms in medical science.

2.5.2 Potential Harms and Benefits of Oxygen Supplementation

The potentially harmful effects of oxygen are widespread, and include effects both systemic and localised to the lungs:

- local effects:
  - absorption atelectasis [42].
  - acute lung injury, ARDS and lung fibrosis [40].
  - inflammatory cytokine production.
- systemic effects:
  - central nervous system toxicity.
  - cerebral and coronary vasoconstriction [35].
  - haemodynamic changes [43, 44]:
    - vasoconstriction [45, 46].
    - increased peripheral vascular resistance. [47, 48, 49]
    - reduced cardiac output.
  - inflammatory changes, including the generation of reactive oxygen species [50].

Some of these potentially deleterious effects may impart benefit in the right circumstances. Vasoplegia (low blood pressure caused by a relaxation of the arterial blood vessels) predominates the clinical picture in sepsis. Vasoconstriction and increased peripheral vascular resistance have thus been speculated as mechanisms whereby high fractions of inspired oxygen may impart benefit [51].

The HYPER2S study explored this hypothesis by randomising patients with septic shock to receive either 100% inspired oxygen for a period of 24 hours or routine care [51]. This study was stopped early under the recommendation of the trial safety monitoring committee due to a larger than expected number of deaths in the high oxygen arm. While the difference in trial arms did not reach the classical boundary for statistical significance, there was a large enough discrepancy to warrant early stopping of the trial.

The ICU-ROX study investigated conservative versus liberal oxygen strategies in critical care [52]. A subsequent subgroup analysis of septic patients in this study was unable to detect a difference between the two groups [53]. These findings are unsurprising, since this randomised controlled trial (RCT) was relatively small, and applied oxygen levels in the "liberal" oxygen arm that are much lower than that used in HYPER2S and more consistent with those seen in the normal practice of critical care. In the ICU-ROX sepsis subgroup analysis, the "liberal" oxygen arm performed numerically better than the conservative arm, fuelling more speculation over this topic.

A meta-analysis of over 16,000 patients found overall evidence for harm from the use of excessive oxygen administration: "Patients treated liberally with oxygen had a dose-dependent increased risk of short-term and long-term mortality" [35].

Despite concerns raised, except for patients with type II respiratory failure[7], oxygen use remains largely unregulated in clinical practice. Prospective randomized controlled trials of oxygen therapy in patients suffering from myocardial infarction have reported either harm [54, 55] or no effect [56]. An increase in mortality risk has been suggested in patients receiving higher inspired oxygen concentrations [57, 58, 59, 60, 61, 62] in conditions such as cardiac arrest [63, 64, 65] and septic shock [51, 66, 67], and also in general critically ill populations [61, 68]. However, most of these studies lack a delineation between harm from appropriately high levels of inspired oxygen used to maintain normoxaemia, and excessive concentrations resulting in hyperoxaemia [69]. Similarly, analyses of critical care databases variably report an association [70, 71], or lack thereof [72], between hyperoxaemia and negative outcomes in the critically ill.

The varied findings between studies may be due in part to a lack of standardisation in what constitutes "excess" oxygenation criteria. Many prior approaches are limited by data availability. Often only a single measure of oxygenation is available to represent an entire healthcare encounter. There is no feasible mechanism through which such a short exposure to high oxygen levels could impart harm, and so these

[7] respiratory failure characterised by elevated arterial CO_2 levels.

findings should be met with justifiable scepticism, since they are likely to be confounded by treatment indication. This confounding has been long recognised. Osler remarked in 1898: “It is doubtful whether inhalation of oxygen in pneumonia is really beneficial. Personally, when called to consult on a case, if I see the oxygen cylinder at the bedside I feel the prognosis to be extremely grave.” [73]

The proposed evolutionary rationale for the harms of excess oxygen is straightforward. Mammalian life has been exposed to 21% oxygen for millions of years, whereas exposure to higher levels is a uniquely modern phenomenon. There has been ample time to develop evolutionary adaptations to low oxygen environments. Examples of such exposure include the in-utero development of the mammalian foetus, human communities who live at altitude, and the survivors of polytrauma and pneumonia [74].

2.6 Sepsis

Sepsis—infection complicated by life threatening organ dysfunction—is commonly encountered in critical care. The mortality of sepsis is in excess of 30% and survivors often have long-term health issues [75, 76]. Recent estimates suggest that in 2017 alone, sepsis was responsible for 11 million deaths globally [77]. Sepsis undoubtedly confers a high humanitarian and economic cost to society [78].

2.6.1 The Pathobiology of Sepsis

Over many thousands of generations, evolutionary pressures have forged a complex interaction between humans and microorganisms. The result is a tightly interconnected network of host defences, designed to mount a proportionate response to infection, so as to render it harmless [79]. Human life is maintained by a delicate balance in the activation of this cascade. So, perhaps, it should not be any surprise that this balance is occasionally perturbed [80].

Microorganisms, having circumvented the skin or mucosa, are immediately recognised by local actors of the innate immune system; granulocytes, macrophages, dendritic cells and complement proteins. These cells and proteins are activated by non-specific structural molecules expressed within the microorganism, known as Pathogen-Associated Molecular Patterns (PAMPs) [81, 82]. Local tissue damage caused by the infection triggers the release of host cellular contents, referred to in this context as Danger-Associated Molecular Patterns (DAMPs) [83, 84, 85], to which the immune system is primed to detect and respond via pattern recognition receptors. DAMPs and PAMPs activate the complement cascade, while release of chemotaxic cytokines (chemical messengers) facilitates recruitment of white blood cells and other actors of the innate and adaptive immune system to the site of infection. This is a highly conserved pathway across mammalian life, and is essential for the normal functioning of life [86, 79].

PAMPs and DAMPs are detected by leukocytes, macrophages and endothelial cells which, in turn, activate genes responsible for promoting inflammation. This is the primary means through which infection is arrested. For reasons that are unclear, in some cases, this process becomes dysregulated, undergoing positive feed-

back that results in massive amplification of the inflammatory pathways. Excessive complement activation triggers the coagulation (clot forming) and fibrinolytic (clot breakdown) pathways, leading to microvascular clot formation and an alteration in microvascular blood flow [87, 88]. This disrupts blood flow, impairing oxygen and nutrient delivery to body tissues.

While the specific causative pathways are unclear, the prevailing scientific consensus is that the energy dependent pathways governed by mitochondria *temporarily* shut down in a process known as bioenergetic failure [86, 89, 90, 91]. This results most notably in cardiac dysfunction, circulatory collapse and multi-organ failure [92].

This *maladaptive* response to infection is known as sepsis. Clinical hallmarks of sepsis are a derangement of the vital signs, in particular low blood pressure caused by a relaxation of smooth muscle found throughout the circulatory system (vasoplegia). The definition of sepsis underwent its third revision in 2016 [93, 94], emphasizing the presence of organ dysfunction as being central to the syndrome.

> "Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection"
>
> — *Singer et al* [93, 94]

Sepsis is not one single disease, but rather a collection of disease states caused by a wide range of infectious organisms targeting different host organs. Pneumococcal pneumonia is inarguably a very different disease to surgical peritonitis, but both are still called sepsis. While united by the syndromic presentation of organ dysfunction caused by infection, there is necessarily heterogeneity in the clinical presentation [79].

Sepsis is often treated in the ICU, since the primary role of the ICU is to provide organ support to failing organs. This helps to keep patients alive while targeted therapies—or the passage of time—can reverse the organ dysfunction.

2.6.2 Identifying Markers of Organ Dysfunction

Table 2.3 highlights a number of examples of common derangements observed for each organ system routinely monitored in the ICU during sepsis, and how those perturbations might be recorded in an electronic health record (EHR) as part of routine care. One important feature of note is that, with the exception of routinely sampled daily blood tests, most patient sampling occurs as a direct result of physiological status; sicker patients are sampled more frequently.

The sequential organ failure assessment (SOFA) score (shown in table 2.4) was developed to allow day-to-day tracking of organ dysfunction in the intensive care unit [95]. Developed by expert consensus, SOFA is an ordinal scale of organ dysfunction from 0 to 24. A higher score corresponding to a greater degree of organ dysfunction. Each of the six individual organ components are scored between 0 and 4 points and are summed to provide a notion of overall organ dysfunction. For the cardiovascular and respiratory systems, SOFA encapsulates treatment-physiology

Organ system	Sampling patterns	Observations	Treatments
Respiratory	↑ ABG sampling	↑ F_IO_2 ↓ SpO_2 ↓ P_aO_2/F_IO_2 ↑ respiratory rate	supplementary oxygen ventilation blood transfusion
Cardiovascular	↑ CO monitoring	↑ inotropes/vasopressors ↓ blood pressure	intravenous fluid inotropes vasopressors
Renal		↑ creatinine/urea ↓ urine output	intravenous fluid renal replacement therapy
Clotting		↓ platelets	platelet transfusion
Liver		↑ bilirubin	cause-specific no general purpose treatments
Neurological	↑ evaluations	↓ GCS	tracheal intubation and ventilation

Table 2.3: Examples of electronic health record representations of organ dysfunction, as defined in sequential organ failure assessment (SOFA). Abbreviations: arterial blood gas (ABG), cardiac output (CO), Glasgow Coma Score (GCS), peripheral oxygen saturation (SpO_2), fraction of inspired O_2 (F_IO_2), partial pressure of arterial oxygen (P_aO_2).

Organ system	Biomarker (units)	0	1	2	3	4
Respiratory	P_aO_2/F_IO_2 (kPa)	≥ 52.6	< 52.6	< 39.4	< 26.3 + RS	< 13.1 + RS
Cardiovascular	MAP (mmHg) drug dose ($mcg.kg^{-1}.min^{-1}$)	≥ 70	< 70	Dopamine ≤ 5 Dobutamine (any)	Dopamine (5-15] Noradrenaline ≤ 0.1 Adrenaline ≤ 0.1	Dopamine > 15 Noradrenaline > 0.1 Adrenaline > 0.1
CNS	GCS (points)	15	[14-13]	[10-12]	[6-9]	< 6
Hepatobiliary	Bilirubin ($\mu mol.L^{-1}$)	< 20	(20-32)	(33-101)	(102-204)	(204)
Coagulation	Platelets ($10^3.mm^{-3}$)	> 150	< 150	< 100	< 50	< 20
Renal	Creatinine (Cr) ($\mu mol.L^{-1}$) Urine output (UO) ($mL.day^{-1}$)	Cr < 110	Cr [110-170]	Cr [171-299]	Cr [300-440] or UO < 500 mL/day	Cr >440 or UO < 200 mL/day

Table 2.4: The SOFA (sequential organ failure assessment) score. MAP: mean arterial pressure, RS: respiratory support, CNS: central nervous system. GCS: Glasgow Coma Scale.

interactions. For example, a patient with "normal" blood pressure supported by drugs designed to augment the cardiovascular system would score more highly (two points) than a patient with low blood pressure but without this support (one point).

This avoids the mistaken conclusion that the latter patient has a greater degree of organ dysfunction simply because their blood pressure is lower. Addressing this treatment-physiology interaction is particularly important; if we were to base an analysis upon patient physiology alone, then the story would be incomplete.

In the ICU, physiology is manipulated by drugs and equipment. To track and understand changes in organ dysfunction, this system as a whole must be interrogated; physiology, drugs and equipment.

The respiratory system is represented in SOFA by the use of the P_aO_2/F_IO_2 ratio and ventilation status. The renal system is represented by creatinine and urine output. This is an imperfect representation of treatment-physiology interaction and is notably lacking the use of renal replacement therapy (RRT). With regards to RRT, one could conclude that, in most cases, urine output is a reasonable proxy for the renal system treatment-physiology interaction, since the patient is unlikely to be liberated from RRT without passing urine. The other organs systems in SOFA (clotting, hepatobiliary, and neurological) isolate specific biomarkers[8]: platelets, bilirubin and the Glasgow Coma Score (GCS) respectively. None of these systems are able to represent treatment-physiology interactions, and are potentially confounded by interventions aimed at correction of organ dysfunction. A mitigating argument as to why they do not capture a treatment-physiology interaction, is that perhaps there are not many therapeutics available and in wide-spread use for these organ systems. There are no general purpose therapeutics that would improve GCS or lower bilirubin. Platelets are often given to patients with severe platelet deficiency. The use of platelets is somewhat conservative, and it would not be a UK practice to correct to a value above 150 $\times$cells^9/L (the starting threshold to define impairment).

The main utility of SOFA in critical care research, is therefore its ability to

[8]though there are several features of the patient that are not biomarkers in the conventional sense, I use the term "biomarker" to refer to any measurable asset of the patient, which may include clinical markers or physiology

account for the treatment-physiology interactions—particularly cardiorespiratory—that are common in this setting.

Other scoring systems exist to achieve similar goals. The logistic organ dysfunction score (LODS), for example, is similar to SOFA, and was developed using a data-driven approach [96]. The present research will focus on SOFA, since this metric is in widespread use in UK critical care research, forms part of the clinical criteria underpinning the sepsis definition itself [93], and the raw values that go into the calculation are readily available in the CC-HIC data model.

2.6.3 An Operational Definition of Sepsis

An operational definition of sepsis is an increase of the SOFA score by two or more points from baseline in the presence of suspected or confirmed infection. Most patients will have a pre-morbid SOFA score of zero [94], and so an initial SOFA score of two is synonymous with an increase in two points from the baseline, should there be no prior information.

It is of interest to highlight that this definition already encodes an intrinsic notion of disease trajectory; a patient must demonstrate an increase in organ dysfunction to meet the sepsis definition.

SOFA is designed to monitor *acute* organ dysfunction. If a patient has prior evidence of organ dysfunction (for example, being a recipient of kidney dialysis) then the affected organ is typically excluded from the calculation [97].

Most episodes of sepsis begin outside the ICU as the patient deteriorates either at home or in a hospital ward. A convenient method to identify sepsis is to seek evidence of infection within the ICU admission diagnosis [98], and evidence of organ dysfunction via the maximum SOFA score achieved in the first 24 hours following admission. This is the same approach as implemented by ICNARC [97].

2.6.4 Sepsis Heterogeneity

Despite a thriving research community, and an abundance of promising therapeutic discoveries, no targeted therapies for sepsis have entered routine use [99], while others have been harmful [100]. Thus, the mainstay of treatment for sepsis includes

source control of the infection, antimicrobials, and organ support. A key barrier to treatment discovery is the heterogeneous nature of sepsis. Robust descriptions of sepsis heterogeneity have thus been highlighted as a key research priority [101, 102, 103].

When studying sepsis in animal models, many conditions can be meticulously controlled [91, 104]. Animals are generally of similar weight, age and sex. The breeding stock is controlled, limiting genetic diversity. The infectious inoculation is standardised, and treatment is given at a fixed time point, referenced to this insult. These experimental conditions unveil the natural history of sepsis on the timescale of the disease, often referred to as *disease time*. The fraction of mortality attributable to sepsis (the so-called "attributable fraction", or the proportion of subjects that died *of* sepsis, rather than *with* sepsis) can be guaranteed to be close to 100%. Under these enriched experimental conditions, statistical power is maximised to discover effective interventions, should they exist.

Compare this to the typical presentation of sepsis in humans. Patients are infected, often without reference to a discrete insult. The early symptoms of sepsis are non-specific and indolent, appearing over hours or days [105]. A broad range of microorganisms can inoculate many different host tissues causing subtle variations in the way in which organ dysfunction manifests. People themselves are heterogeneous in terms of age, sex, genetic background, co-morbidities and access to healthcare. After a variable amount of time, a patient deteriorates and travels to hospital. Patients are now no longer referenced to disease time, and become grounded in administrative time-frames: hours since admission. Complicating matters, not all patients will die *of* sepsis, but rather *with* sepsis. The attributable fraction of mortality in sepsis has been estimated at a surprisingly low 15% [106][9]. The net effect of these features conspires to generate a noisy signal, where effective treatment signals may be lost as they were delivered too late, or to those patients who could never have derived benefit in the first place.

To this end, many promising research avenues to explore this heterogeneity

[9]This surprisingly low figure may reflect the increased diagnostic rate for sepsis in the UK [107].

have emerged. This area of research is known as "phenotyping". A phenotype refers to "a clinical entity defined by observable characteristics that are produced by interactions of the genotype and the environment." [103] Phenotype discoveries in sepsis have been found from such diverse research methods as transcriptomics [108, 109], metabolomics and proteiomics [110]. Other approaches have targeted features present within directly observable patient physiology [111, 112, 113, 114, 115, 116, 117, 118].

It remains to be seen how these newly discovered sepsis phenotypes relate to underlying biological mechanisms of sepsis or how they could be operationalised for treatment purposes.

Heterogeneity of Treatment Effect

The modification of a treatment effect across patients is formally known as heterogeneity of treatment effect (HTE) [119]. The manifestation of HTE is a reduction in statistical power to detect a true population average treatment effect. The impact of HTE has been repeatedly demonstrated via simulation studies, which suggest that important therapeutic options now in use for known subtypes of sepsis—in particular the ARDS—might not have been discovered had those patients not been specifically isolated [120]. This argument has been extended to conclude that either the sepsis syndrome is not amenable to study under the RCT paradigm, or that mortality is not a suitable endpoint to study [120]. While I disagree that sepsis should not be studied with RCTs, this position must be taken with serious consideration. At best, it suggests that critical care trials in sepsis are systemically underpowered, and that simply increasing the size of the trial is unlikely to be a successful strategy. In view of this HTE has long been at the core of discussions of sepsis research [101].

In order to fully understand HTE, it must be broken down into two key principles: risk magnification (RM) and differential treatment effect (DTE).

Risk Magnification

RM is a mathematical artefact that occurs when considering the risk reduction of a treatment on an absolute scale [121]. This follows necessarily from the nature of risk (the probability of an event) being confined on a scale from zero to one. In the

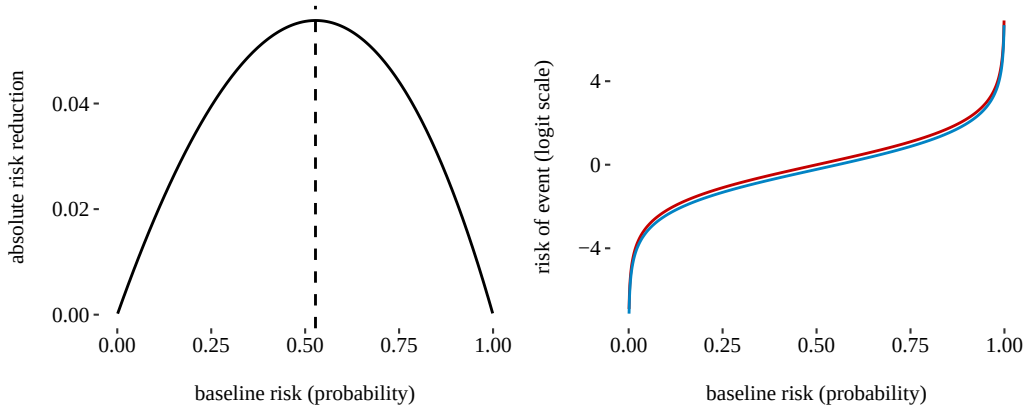


Figure 2.5: Left panel: Expected survival benefit in absolute risk reduction across the full range of baseline risk death. The odds ratio is set at 0.8. Counter-intuitively, there is an optimal point of maximum absolute risk reduction near 0.53; the effect is not symmetric around 0.5. The absolute benefit conferred decreases predictably toward both extremes of risk. **Right panel:** Effect of consistent improvement in survival benefit acting across the whole cohort mapped from the probability scale (x axis) to the logit scale (y axis).

extreme case, if the risk for death is 1 (absolute certainty of death) then there is no scope for a treatment to improve the prognosis. Similarly, if the risk for death is 0 (absolute certainty of life) then there is equally no scope for a treatment to improve the prognosis; it is already guaranteed. There lies an optimal point on this spectrum of risk, where an effective treatment results in the maximum absolute risk reduction in the probability of death.

Figure 2.5 illustrates this concept for a theoretical treatment that confers the same relative risk reduction—defined as an odds ratio for death of 0.8—across the target population. At the extremes of risk, the patient benefits from a minuscule absolute risk reduction. With low risk of a poor outcome, it takes relatively little harm (which is common in healthcare [122]) to overwhelm the signal of benefit; patients at low risk of an adverse outcome might be better served by not having the treatment.

Risk magnification itself could never confer harm directly to low risk patients. However, if there are complications that arise from treatment that are fixed across all patients[10] then low risk patients, who are equally likely to suffer complications

[10]Estimates as high as 1% have been found for life threatening complications related to placing

as high risk individuals, may ultimately come to more harm than if they had not received treatment. This effect has been elucidated in simulation, where a therapy was shown to be ineffective overall, despite being highly effective in high risk groups, whilst neutral or harmful in low risk groups [124]. Examples of this phenomenon exist in other medical fields such as neurology, cardiology and vascular medicine [125, 126].

Risk magnification does not isolate a particular mechanistic or biological process. Patients who are at higher risk of death from any cause—advanced age, frailty, or any other comorbidity—will stand to benefit from a treatment that confers the same relative risk reduction to all.

Differential Treatment Effect

A treatment can afford a truly differential effect across biological characteristics. This component of HTE is known as differential treatment effect (DTE).

To illustrate, as previously introduced, ARDS is a pulmonary syndrome that can manifest as a distinct subtype of sepsis. The clinical presentation of ARDS is characterised by profound hypoxaemia due to an increase in extra-vascular lung water, poor lung compliance and impaired gas exchange. COVID-19 and influenza are both viral causes of ARDS. Steroids have been shown to be effective in the treatment of COVID-19 ARDS [127], yet harmful in the treatment of influenza ARDS [128]. This is a contemporaneous example of differential treatment effect in action; patients with the same syndrome, responding differently to a treatment, driven by underlying biological differences.

In 2000, the ARDSnet trial—a landmark trial in critical care—demonstrated that a so-called "lung protective" ventilation strategy improved mortality in the management of ARDS [129]. Recently, Girbes *et al* have suggested [120], that had the ARDSnet trial not enriched their cohort by selecting for the distinct ARDS phenotype, then the discovery that low tidal volumes were beneficial might well have gone unnoticed.

---

a patient onto mechanical ventilation during an emergency [123]. This is a common critical care intervention in sepsis.

Heterogenious Patient Morphologies

Typical models in critical care rely on static notions of disease severity, asking "How sick is this patient upon arrival to intensive care?". It is common for heterogeneity to be explored through subgroups that are defined at the point of admission. Examples include:

- sex.
- age.
- severity at presentation.
- "early" vs. "late" disease.

A hitherto unexplored part of patient heterogeneity is the morphology of longitudinal physiology observed during the course of critical care. Clinical intuition tells us that longitudinal information is important. For example, when treating a patient, it is not enough to know only todays biomarkers. One must know yesterdays results to contextualise the findings. A patient with severe disease, who is improving, may well have a better prognosis than a patient with mild disease who is deteriorating. This kind of determination requires the interrogation of longitudinal patient data.

Quantification of the relationship between longitudinal disease morphologies and patient outcomes in critical care is non-trivial and has so far remained elusive [130].

2.6.5 Prior Evidence for Disease Morphology

Notable examples of investigating longitudinal organ dysfunction in sepsis include those by Toma *et al* [131, 132], Holder *et al* [133], Badawi *et al* [134] and, using a joint modelling approach, Deslandes and Chevret [135] and Musoro *et al* [136].

The Toma *et al* approach relies on aligning commonly observed daily temporal patterns of SOFA. In order to improve the computational properties of the approach, they categorise SOFA into: "low", "medium" and "high" [131]. They weighted more recent events are being more important in the model, and found that certain recurrent patterns of organ dysfunction were more commonly associated with a poor outcome.

Holder *et al* investigated the incremental improvement to model fit with the sequential addition of each new day of SOFA [133]. They found that the addition of SOFA beyond the fifth day of treatment on an ICU did not significantly improve model discrimination for death.

Badawi *et al* investigated the use of SOFA, APACHE II and the Discharge Readiness Score (DRS) to explore trajectories within the ICU [134]. Patients were stratified into cohorts of one, three and seven days' stay inside the ICU, with clear differences observed in the trajectories between survivors and non-survivors.

Deslandes and Chevret found, perhaps unsurprisingly, that increased SOFA severity was associated with death [135]. What is unique about their approach was the implementation of joint models to meticulously model the entire longitudinal history of SOFA and come to this conclusion in a principled manner. This finding was corroborated by Musoro *et al* using similar methods [136].

Acute physiological trajectories have previously been explored in the (SPOT)light study [130]. This study investigated the changes in physiological measurements made in the 24 hours before and after admission to an ICU. When using the post-ICU admission values as the reference group, pre-ICU trajectory of systolic blood pressure, was the only physiological marker associated with 28 day mortality. By contrast, when using the pre-ICU assessment as the reference, and looking at added value for physiological values projected into the future, many physiological markers were associated with 28 day mortality. It was therefore suggested that these patients were demonstrating the Markov property, in that the present state of the patients did not depend upon their history.

2.7 Simulated Critical Care Cohort

A simulated dataset is presented containing 200 patients in whom the daily SOFA score is recorded in the "ICU" from day 0 (the first 24 hours) up to day 30. This cohort was created with the simjm package for R [137] developed by Brilleman [138]. This dataset provides the basis from which particular features likely to feature in real data can be discussed. Three individual cases from the cohort have been selected and assigned names for ease of reference: Athena, Hermes and Zeus. The longitudinal data for these simulated patients are shown in figure 2.6:

- Zeus deteriorates, seeing an increase in his SOFA score, and dies shortly before the 10th day of being inside the ICU, removing him from the cohort.
- Hermes remains stable, and remains in the cohort alive on the 30th day of observation.
- Athena shows improvements with a reduction in her SOFA score. She is discharged alive on the 18th day of being inside the ICU and so stops contributing any further data to the study.

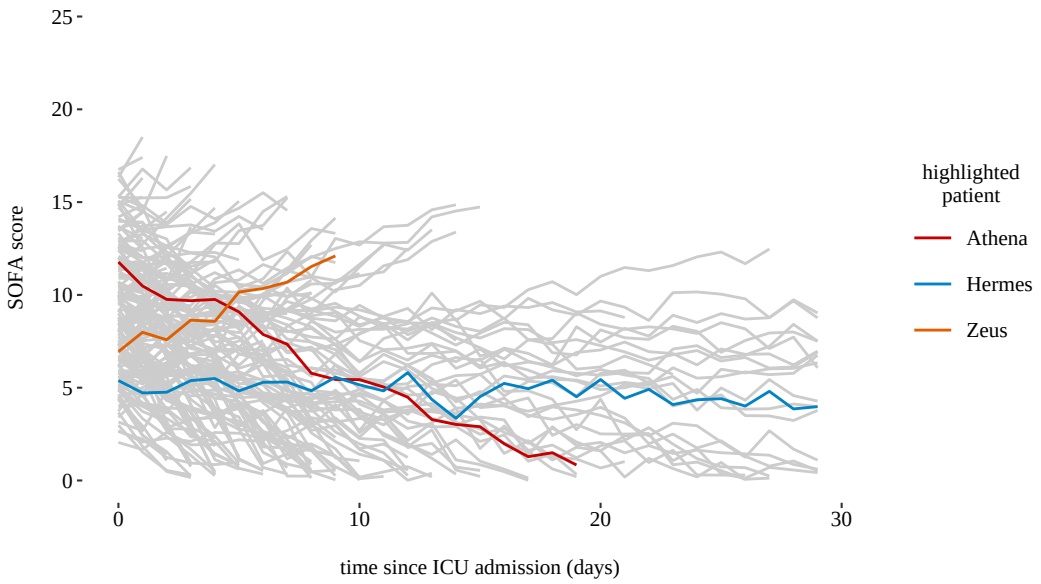


Figure 2.6: Simulated cohort to illustrate key statistical concepts. The daily SOFA scores of 200 simulated patients are displayed. Three patients have been highlighted and named to draw out specific features of the data.

2.8 Longitudinal Data Analysis

A motivation for exploring longitudinal patient biomarkers is to understand how a biomarker changes over time, as part of the natural history of a disease, or in response to a treatment or patient characteristics. The statistical hallmark of such biomarkers is that, in general and for a given patient, they are autocorrelated; i.e. where the latest result is correlated with previous results. If this were not true, and repeated samples from each patient were independent of one another, then there would never be any physiological trajectory information arising from the patient to explore. This pattern is clearly demonstrated in figure 2.6, where each patient can be seen to follow their own patient specific trajectory. The linear mixed effects model is particularly well suited to answering questions in the context of longitudinal data, and so has become virtually synonymous with such an analysis [139][11].

2.8.1 Linear Mixed Effects Model

The linear mixed effects model (equation 2.1) has three main components: fixed effects, random effects, and stochastic error.

$$\begin{aligned} y_i &= X_i\beta + Z_i b_i + \varepsilon_i \\ b_i &= \mathcal{N} \sim (0, D) \\ \varepsilon_i &= \mathcal{N} \sim (0, \sigma^2 I_{n_i}) \end{aligned} \tag{2.1}$$

The "fixed" effects—represented by $X_i\beta$ in equation 2.1—capture the population average trajectory of a biomarker. This is demonstrated in figure 2.7 where the population average trajectory has been overlaid onto individual patient trajectories. The fixed effects therefore explain all that is commonly shared in a biomarker between patients.

The "random" effects—represented by $Z_i b_i$ in equation 2.1—capture deviations from the population average trajectory necessary to describe the specific individual trajectory of each patient. Conceptually, it is helpful to think of the random effects as a "nudge" that pushes the average trajectory of the population as a whole

[11] Other names include multi-level or hierarchical model.

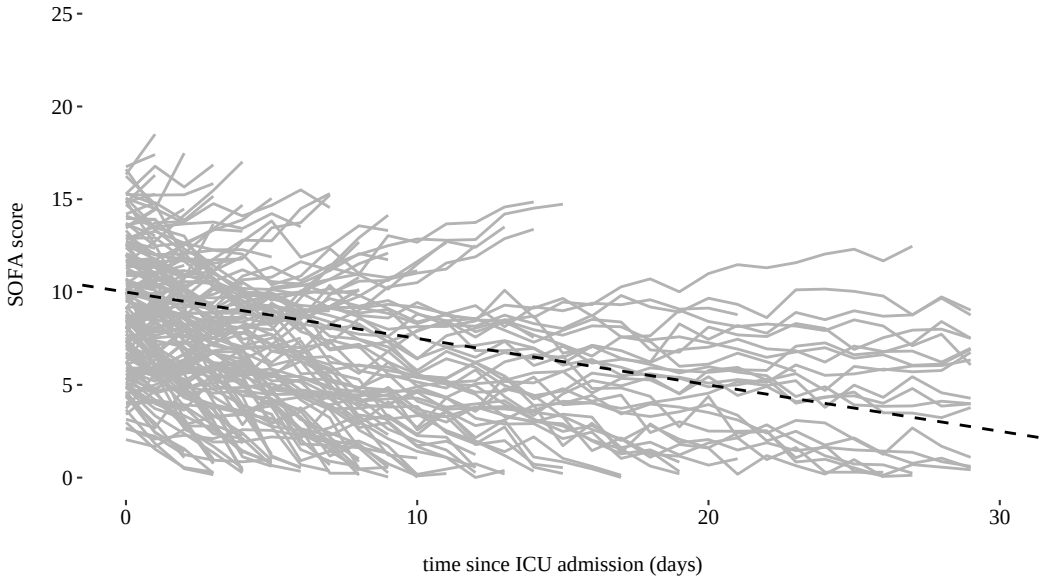


Figure 2.7: The "fixed effects" for the simulated cohort are shown as the dashed black line. In this example, this shows a gradual reduction in the biomarker over time. This is the average biomarker trajectory. The underlying individual biomarker trajectories are shown in grey behind.

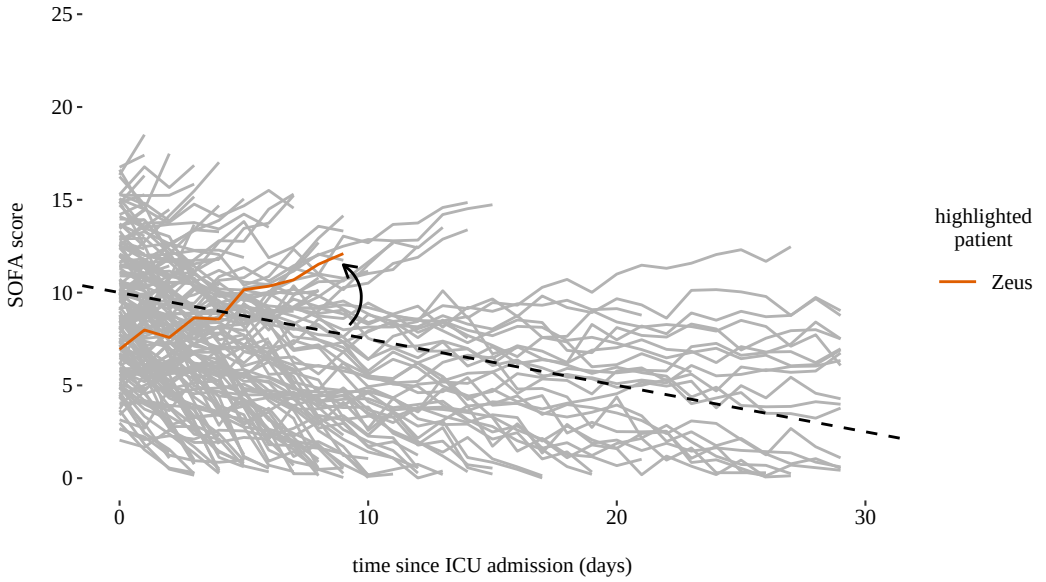


Figure 2.8: The "random effects" for the patient identified as "Zeus" are highlighted as a nudge (the arrow) transforming the population average trajectory of the biomarker ("fixed effects") into the individual specific biomarker trajectory for the patient.

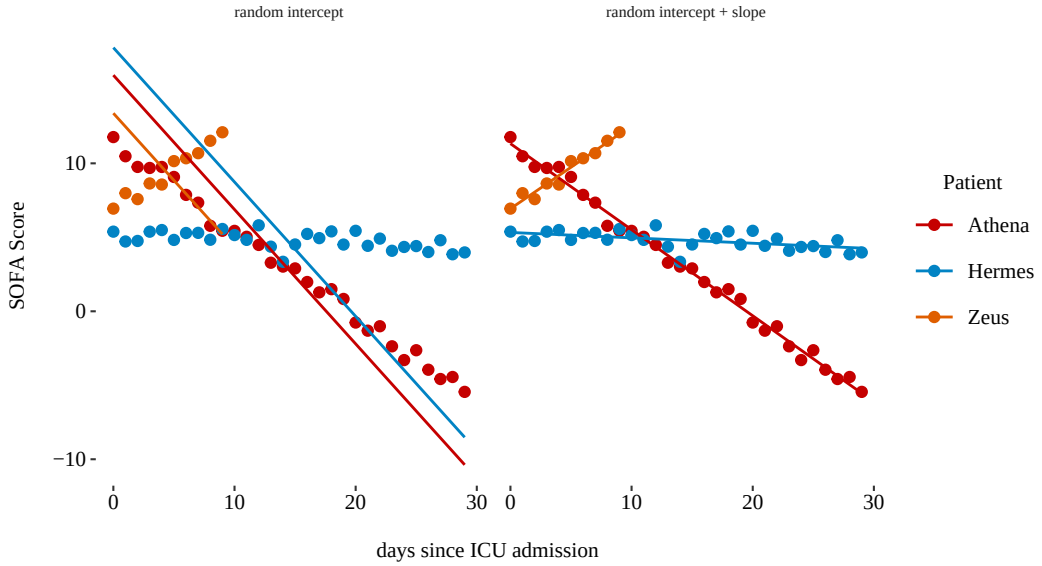


Figure 2.9: Classical implementations of the linear mixed effects model. **Left:** Random intercept model only model, permitting parallel individual biomarker trajectories. **Right:** Random intercept plus slope model, permitting non-parallel individual biomarker trajectories.

to individual patients (figure 2.8). The random effects can more broadly be thought of as latent biological and environmental factors that may explain variation between patients, but are not directly measured as specific characteristics of the patient. A direct implication is that the whole cohort shares the same fixed effects (hence the term "fixed"), but each patient has their own random effects.

After the fixed and random effects in the model have been applied—the deterministic components of the model—any discrepancy between the model prediction of a patient's biomarker, and what is actually observed is known as the residual error (ε_i). When defining the linear mixed effects model, there is often a balance to be struck between how much variance of the biomarker can be captured by the fixed and random effects, and how much should be allocated to the residual error.

Figure 2.9 highlights two common implementations of the linear mixed effects model: random intercept, and random intercept plus slope.

In the random intercept model, the individual fitted trajectory of each patient is permitted to change only by shifting along the y axis, while keeping the trajectories parallel. In this particular example, this has resulted in a poor model fit. In contrast,

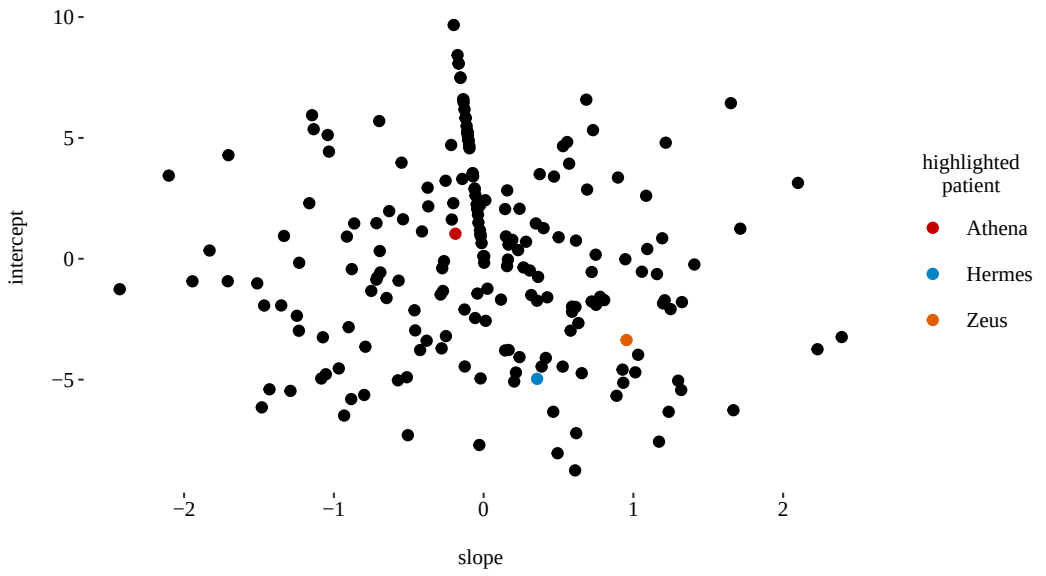


Figure 2.10: Multivariate normal distribution of random effects. Each patient's random effects are plotted demonstrating a multivariate normal distribution with little correlation between the intercept and slope. The highlighted patients are shown in this cloud.

the random intercept plus slope model permits the intercept to also vary, allowing for a much better model fit in this particular instance.

The random effects are typically defined in terms of a multivariate normal distribution with mean vector zero and covariance matrix D to be estimated from the data. This formulation suggests that more patients will more commonly act similarly to the population average trajectory, rather than take their own outlying trajectory. This multivariate normal distribution can be shown by plotting the random effects in a scatter plot (figure 2.10).

2.9 Survival Analysis

When investigating the time to survival for a patient, there are two common features of the data that must be addressed. First, as the distribution of survival times are strictly positive (events are in the future), we can expect this target distribution to be positively skewed. Second, censoring of death is common. Censoring occurs because a study has a finite follow-up time and there will usually be patients who are still alive at the end of the study period, who have yet to experience the event

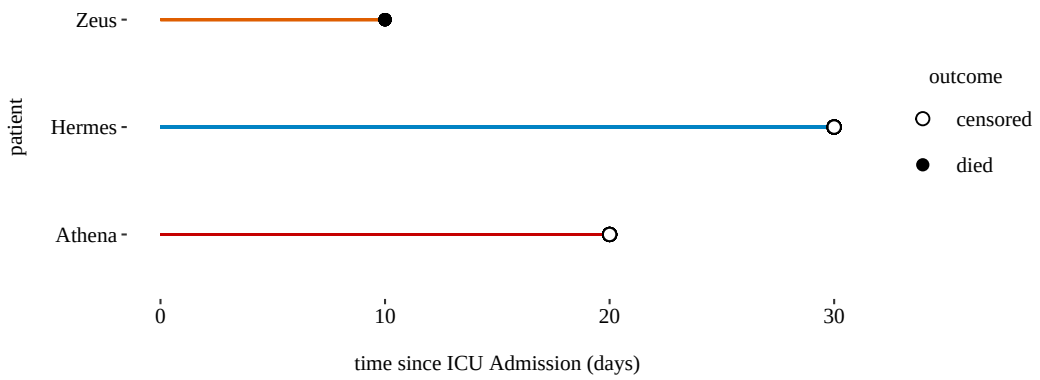


Figure 2.11: Censoring of outcomes as a common feauture of survival analysis

under investigation.

Figure 2.11 depicts the three exemplar patients from the synthetic cohort. Here, their time to outcome is displayed. All patients arrive in the ICU at time zero and are "enrolled" into the study. Zeus dies inside the ICU on day 10 (closed circle). Hermes is alive and inside the ICU on day 30. Athena is discharged alive from the ICU on day 20. Both Hermes and Athena have unknown outcomes, but in both cases it is known that they survived at least a certain amount of time.

Survival models are optimised to address this censoring as despite not experiencing the survival time explicitly, it remains useful to know that a patient did not experience an event up to a particular time point.

The distribution of event times can be re-expressed as the survival, the hazard and the cumulative hazard functions (figure 2.12). These functions are used in the modelling of event times, and important to understand the implications of survival modelling.

The survival function (equation 2.2) describes the probability of surviving to some time (T), conditional on having survived till now (t). As such the survival function is the compliment of the Cumulative Distribution Function (CDF) of the event times.

$$S(t) = Pr\{T > t\} = 1 - F(t) \tag{2.2}$$

The survival function can be reformulated as the hazard function (equation 2.3)

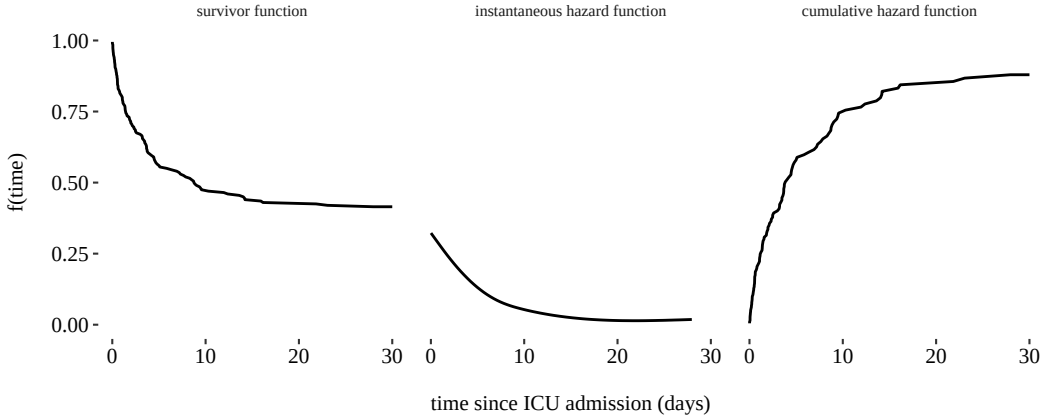


Figure 2.12: Three functions of time-to-event type data are shown. **Left panel:** The survival function. This shows the probability of survival from the start of the study to a given time into the future, it is the compliment of the Cumulative Distribution Function (CDF). This plot is anchored at 1 at time 0 since all patients are (typically) event free at the beginning of the study. **Middle panel:** The hazard function. This shows the instantaneous rate of events at a given time, conditional on having survived to that time. **Right panel:** The cumulative hazard function. This shows the integral of the instantaneous hazard function, and as such represents the cumulative risk for the event.

to express the hazard of an event at an instantaneous moment in time. The hazard function considers the risk of the event over some very small interval of time (dt). As this interval approaches zero, the solution for the instantaneous hazard at time t can be found. The hazard function is particularly useful in the communication of risk, since it is a conditional probability based on survival to the time of interest.

$$\lambda(t) = \lim_{dt \to 0} \frac{p(t \leq T < t + dt) \mid T \geq t}{dt}, t > 0 \tag{2.3}$$

The instantaneous hazard function can be re-expressed as the cumulative hazard function by integrating from time zero to the time of interest (equation 2.4).

$$H(t) = \int_{t=0}^{t} \lambda(x) dx \tag{2.4}$$

All three functions are connected, and any one can be used to derive the other two.

2.9.1 Cox's Proportional Hazard Model

Cox's semi-parametric proportional hazard model, or its parametric cousin the Weibull survival model, are both commonly employed in this task of survival modelling [140, 141]. The general formulation of Cox's proportional hazard model (equation 2.5) is given in terms of the instantaneous hazard function.

$$\lambda(t|X_i) = \lambda(t)\exp(X_i\beta) \tag{2.5}$$

Just as in the linear mixed effects model, Cox's model contains fixed effects ($X_i\beta$) which in this context are the population level modifiers of the baseline hazard function ($\lambda(t)$). Cox's model is semi-parametric in that the baseline hazard function does not need to be specified. This is contrasted with the weibul model, where the baseline hazard function is parametrised by the weibul distribution. Differing from the linear mixed effects model, the fixed effects of Cox's model act multiplicatively (rather than additively) on the baseline hazard, owing to their position in the exponent of the baseline hazard function.

Cox's model is orientated toward the analysis of how measures taken at baseline—such as patient characteristics upon arrival to the ICU—affect survival. While there are extensions of Cox's model to include time varying data, their implementation has been shown to introduce bias secondary to the implicit assumption that biomarkers are static between samples; which is often an unreasonable assumption in critical care when patients are by definition physiologically unstable. Further, the time varying extension to Cox's model is only valid when biomarkers are of an exogenous nature (i.e. they are not generated by the patient themselves). In this instance, the biomarkers of interest are endogenous; they are created by the patient and cease to exist after patient death. When the desire is to investigate specific morphologies of biomarkers, a different modelling approach will be required. Joint models are such an approach, and these will be the focus of the remaining discussion. It will first be useful to discuss the different patterns of missing data that are likely to be encountered, as these are central to the application of joint models.

2.10 Patterns of Missing Data

For any given observation window, patient data can be missing from the electronic health record. It is generally true that patients seek treatment at times of ill-health, and so in the first instance the electronic health record is a sampling of patients during these times. As a result, there will be an over representation of patients who are acutely unwell.

The most striking missing data pattern, which can be clearly seen in the simulated dataset (figure 2.6 on page 68), is that patients who experience more extreme physiology are more likely to die and be removed from the cohort. Patients who die no longer produce any physiological data, and so this can be regarded as a form of missing data.

In order to formalise this intuition of missing healthcare data it has proven useful to categorise missing data as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [142]. Each pattern of missing data can be visualised with the aid of the causal diagrams (also known as a Directed Acyclic Graph) shown in figure 2.13 [143, 144]. Three scenarios are presented to support the discussion of missing data patterns. In each scenario, we are interested in the effect of any particular patient characteristic (Pt Chr) on their longitudinal outcome (Y), as indicated by the blue arrows in figure 2.13. This could be any biomarker, but to help make the scenarios concrete, we shall take Y to be the daily SOFA score. In each scenario we are only able to observe the subset of Y that does not contain any missing data; Y_O.

2.10.1 Missing Completely at Random (MCAR)

In the first scenario (left most panel of figure 2.13) corruption of the EHR database has led to the random loss of some patient data (Error). As such, we do not know the SOFA score for every patient at every time point, and instead have a random subsample of these data (Y_o). The mechanism causing the missing data to occur is a random process with respect to the patient and biomarker, and so we can say that Y_o is a random sample of Y and thus any inferences we wish to make over the pathway of interest will remain valid and unbiased if we choose to study the

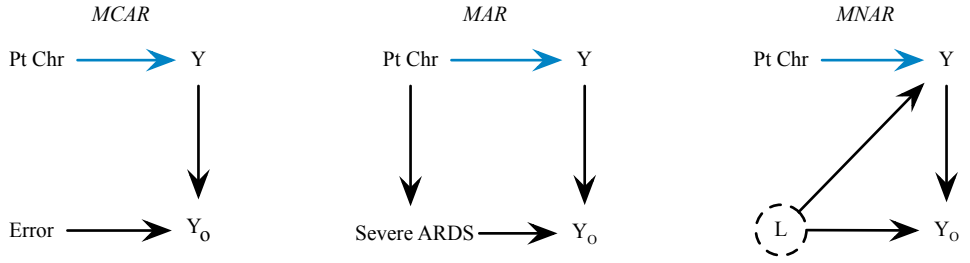


Figure 2.13: Classification of missing data. **Left:** missing completely at random (MCAR). **Middle:** missing at random (MAR). **Right:** missing not at random (MNAR). A full description of the causal diagrams are provided in the accompanying text.

available data (Y_o). This corresponds to the MCAR data pattern. Formally, MCAR corresponds to a missing data distribution that is independent of both observed and missing data. The consequence of such missingness is a reduction in statistical power to detect an effect, but estimates of the effect should remain consistent and unbiased on the whole.

2.10.2 Missing at Random (MAR)

In the next scenario (middle panel of figure 2.13), patients are transferred to a specialist ICU when their respiratory function reaches a pre-defined threshold. In this case, naive analysis of the observed SOFA (Y_o) will produce biased estimates, as patients with a certain degree of respiratory impairment will be absent from the cohort. Since we know the observable patient features that give rise to the missing data, this can be controlled for. Formally, MAR corresponds to a missing data distribution that is independent of the missing data, conditional on the observed data. Thus, by conditioning on the observed data, we can recover the true distribution of complete data.

2.10.3 Missing Not at Random (MNAR)

In the final scenario (right panel of figure 2.13), a latent process (indicated by the L with a dotted circle) is responsible for generating both the distribution of SOFA, and its missingness. This latent process could be a combination of the environmental and physiological factors that give rise to patient deterioration and death. In

this instance, sicker patients produce higher SOFA scores, and are ultimately eliminated from the cohort as a direct result of their physiology. This pattern of missing data is known as missing not at random (MNAR). Formally, MNAR corresponds to a missing data distribution without the conditional dependences defined by the MCAR or MAR patterns. That is, MNAR describes a missing data distribution that is dependent on both observable and missing data.

Models that try to estimate a parameter from data that exhibits a MNAR pattern *can* be biased. The intuition for this bias is that sicker patients die and stop contributing data to the study. Those that are left, are inherently more stable, and so the model has a bias toward these less severe patients.

This final scenario is realistically encountered in critical care, and is the type of missingness that is endemic when interested in patterns of longitudinal physiology, where mortality is prevalent. In these circumstances it is necessary to propose a model for the missing data mechanism, and model this jointly with the distribution for the biomarker under investigation. The joint modelling paradigm provides such a framework through which unbiased estimates for longitudinal variables of interest can be identified, in the presence of such missing data patterns. There are many other merits to the joint modelling approach, and these will be discussed in the following section.

2.11 Joint Analysis

The explanation that follows is influenced by the excellent introduction to the topic of joint models by Rizopolous [145], including the academic training program on joint models conducted at Erasmus University [146], on which a substantial portion of the applied research in this thesis is based.

The standard joint model is a shared parameter model that simultaneously models a biomarker (or other endogenously generated longitudinal outcome), and an event time, such as death [147, 148, 145]. The standard joint model combines a linear mixed effects sub-model for the longitudinal outcome and a proportional hazards sub-model (such as Cox's model) for the survival outcome. These sub-models

are connected through their random effects—that is, latent patient characteristics—that are calculated simultaneously for both sub-models. Equation 2.6 shows the formulation for the standard joint model.

$$\begin{aligned} y_i(t) &= X_i(t)\beta + Z_i(t)b_i + \varepsilon_i(t) \\ m_i(t) &= X_i(t)\beta + Z_i(t)b_i \qquad \text{without error} \\ h_i(t) &= h_0(t)\exp\left\{\omega_i\gamma + \alpha\{m_i(t)\}\right\} \end{aligned} \tag{2.6}$$

Where α is the association parameter that quantifies the connection between the two sub models. The other parameters are as described previously for each sub model. From the formulation in equation 2.6 we can see how the models are connected through their random effects. Not only are the random effects used to connect both sub models, but the error free component of the longitudinal submodel (m_i) is incorporated into the survival sub-model so as to account for the error prone measurements in the longitudinal biomarker. The advantages of the joint modelling paradigm include:

1. Improved statistical efficiency as survival and longitudinal outcomes are modelled simultaneously, making full use of all information available.
2. Easy exploration of a number of morphological patterns that related the longitudinal biomarker to survival (as described in section 2.4.1 on page 49).
3. Parameter estimates are less prone to bias in the presence of informatively (MNAR) missing data patterns.
4. Parameters estimates given an endogenous process of the longitudinal biomarker are less biased.
5. An account is made for the fact that the longitudinal biomarker is often measured with error.

To offset these numerous advantages, there is a significant disadvantage that joint models are computationally expensive. Simplifications of this approach have been suggested that have generated computational improvements. For example, the two-stage approach whereby one fits the linear mixed effects model first and then

uses this to provide fitted biomarker values to the event sub-model. This approach has generally been shown to yield biased estimates and so is not presently recommended. However, this remains an active area of development, with recent research suggesting that certain implementations of the two stage approach, do have the potential to yield unbiased estimates [149].

Owing to these properties, joint models have proven useful in branches of medicine where biomarker trajectory plays a strong mechanistic role in death; i.e. death represents an appropriate event to model the missing data process. Specific examples include the trajectory of CD4 T-cell count in HIV [150], the rate of change in aortic valve area in aortic stenosis, deterioration in renal function in renal transplant [151] and the change in intraocular pressure in glaucoma [152]. Joint models have only rarely been used in critical care research, with some notable recent exceptions [135, 136].

A major challenge of the joint modelling paradigm has been in developing the applied software capability to build such models. Fortunately, a rich ecosystem of software has now been developed that allows for the application of joint models within the R [137] statistical computing language. These packages (non-exhaustively) include JM [153], Merlin [154] and joineRML [155] with a frequentist approach or JMBayes [156] and RStanarm [157, 158] with a bayesian approach.

2.11.1 Matching Model to Morphology: Association Structures

A number of reparameterisations of the association parameter have been described to capture different kinds of biological mechanisms. These are aligned to the morphologies described in section 2.4.1 on page 49. Common implementations include the value, value and slope, and cumulative effect parametrisations, which are synonymous with severity, severity plus velocity (trajectory) and cumulative effect morphologies as previously defined.

2.11.2 Severity (Value) Association

The severity (or value) association is the basis for the standard joint model as demonstrated in equation 2.6. This formulation connects the longitudinal and sur-

vival outcomes based upon the current value of the biomarker, measured without error.

2.11.3 Severity (Value) and Velocity (Slope) Association

The severity and velocity association (or value and slope), which together represent disease trajectory, evaluates the relationship between the current value of a biomarker and its slope, and the event outcome. This formulation is shown in equation 2.7.

$$\begin{aligned} y_i(t) &= X_i(t)\beta + Z_i(t)b_i + \varepsilon_i(t) \\ m_i(t) &= X_i(t)\beta + Z_i(t)b_i \\ h_i(t) &= h_0(t)\exp\left\{\omega_i\gamma + \alpha_1\{m_i(t)\} + \alpha_2\{m_i'(t)\}\right\} \end{aligned} \tag{2.7}$$

α_1 carries the same interpretation as for the value only formulation, with α_2 being the association parameter for the slope of the biomarker represented by $m_i'(t)$. Generally, the slope parameter is not used in isolation, as there is rarely a clinical situation where the current value of the biomarker is irrelevant, with only the slope being useful.

2.11.4 Cumulative Effect Association

The cumulative association evaluates the relationship between the integral of a biomarker and the event outcome. Typically, this is implemented as the area under the biomarker time curve, from time zero to the current time (equation 2.8). This parametrisation is particularly useful to explore if the individual history of a biomarker has an impact on the event outcome.

$$\begin{aligned} y_i(t) &= X_i(t)\beta + Z_i(t)b_i + \varepsilon_i(t) \\ m_i(t) &= X_i(t)\beta + Z_i(t)b_i \\ h_i(t) &= h_0(t)\exp\left\{\omega_i\gamma + \alpha\int_0^t m_i(s)ds\right\} \end{aligned} \tag{2.8}$$

2.11.5 Joint Models and Missing Not at Random

As previously defined, MNAR describes a missing data distribution that is dependent on both observable and missing data. Under such circumstances is it necessary to jointly model the missing data mechanism, and the longitudinal outcome, to attain unbiased estimates for the latter [159, 145]. The intuition behind this is that patients who experience more extreme deviations in their longitudinal biomarker from the population average, will be more likely to die and thus be removed from the cohort. *If* these deviations can be captured by the shared random effects, then it is possible to remove or reduce bias that has been introduced by the death process. This is governed by the conditional independence assumption, which stipulates that conditional on the random effects, the longitudinal outcomes are independent of the death process, and that repeated measures of the longitudinal outcome are independent of each other. A further implication of this assumption is that the joint model will necessarily make explicit assumptions about the full path of the longitudinal biomarker, regardless of whether or not its observation is interrupted by death. In practice, these are un-testable assumptions, and so we must rely on domain knowledge of the subject to provide the evidence base to support this modelling approach.

To summarise, by modelling the survival process simultaneously with the longitudinal biomarker, an account can be made for the bias that is introduced by the death process.

When interpreting the joint model, if the association parameter (α) is zero, then it is implied that both sub-models are independent. If this is the case, one could also reasonably use this to infer that since the models are independent, then informative missingness is in fact not a feature of the data.

2.11.6 Evaulation of Joint Models

It will be necessary to compare joint models with different association structures to seek evidence for which morphology is best representative of outcomes. As with other statistical models, joint models can be compared in terms of Akaike information criterion (AIC), Bayesian information criterion (BIC) and the Log-Likelihood. Within a set of models, this provides a gauge as to which model per-

forms best. Where models can be nested—as is the case with the severity and trajectory models—the likelihood ratio can be used to provide a formal evaluation of improved model fit.

Other performance measures are directed at evaluating model calibration and discrimination, though these concepts will need extending to the longitudinal domain [160].

In static models, receiver operator characteristic (ROC) curves and the mean square error (Brier score) can be used for determining discrimination and calibration respectively. The sensitivity and specificity of a biomarker are interrogated at varying thresholds to draw the ROC curve. The ROC curve plots the sensitivity (true positives) against 1 - specificity (1 - false positives) which are defined in equation 2.9.

$$\begin{aligned} TP(c) &= Pr\{y_i > c | d_i = 1\} \\ 1 - FP(c) &= Pr\{y_i \leq c | d_i = 0\} \end{aligned} \tag{2.9}$$

Where y_i is a biomarker of interest, c is the range of biomarker thresholds and d_i is an indicator for the true underlying state of the patient; 1 if the patient has experienced the event of interest, and 0 otherwise. The area under the receiver operating characteristic (AUROC) curve, shown in equation 2.10, provides a measure of discrimination over the full range of the biomarker.

$$AUROC = Pr\{y_i > y_j | d_i = 1, d_j = 0\} \tag{2.10}$$

Equation 2.10 reveals the direct intuition for the value of the AUROC. If one were to compare any two randomly selected patients (patients i and j), the AUROC corresponds to the probability that the biomarker would correctly order these patients in terms of their probability of experiencing an event. This provides a value between 0 and 1, where 1 is perfect discrimination, and 0 is perfectly incorrect discrimination[12]. 0.5 corresponds to random chance discrimination, and would be if we were

[12]In reality the scale is from 0.5 to 1, since all one needs to do to have improved discrimination if the AUROC is < 0.5 would be to reverse the labels

to flip a coin independent of the biomarker to determine patient status.

These concepts can be extended into the time-series domain, by defining a clinically meaningful time horizon $(t, t+\Delta t]$ over which this evaluation will take place. In particular, we are interested in the ability for a model to correctly rank the mortality predictions over Δt for all pairs of patients based upon their baseline characteristics and their observed longitudinal biomarker up to the current moment in time (t), as shown in equation 2.11.

$$\begin{aligned} TP_t^{\Delta t}(c) &= Pr\{y_i > c | T_i \leq t\} \\ 1 - FP_t^{\Delta t}(c) &= Pr\{y_i \leq c | T_i > t\} \end{aligned} \tag{2.11}$$

An important consideration for such discrimination measures is that they now depend on the starting time of interest (t), since this will effect the number of patients at risk of the event, and the time horizon of interest (Δt). So while it is possible to collapse this measure into a single number of overall discrimination, as with AUROC, it may be more useful to interrogate the model over a range of different starting times and time horizons to understand the strengths and weaknesses of the model over time.

Part I

Data & Software Resources

Chapter 3

Data Model Evaluation

A good clinical data model, or "common data model" (CDM) is vital to a successful EHR driven data collaboration. In 2003 Moody and Shanks wrote their seminal paper on an empirical evaluation framework for data models [161]. Their work highlighted the large inefficiencies with which data models are most commonly developed; often relegating any formal evaluations until after data have started to flow between organisations. The Moody and Shanks framework was extended to the comparative effectiveness research area by Kahn *et al* in 2012 [162] who developed an evaluation framework through which any CDM should be scrutinised. This approach considers eight domains: completeness, integrity, flexibility, understandability, correctness, simplicity, integration and implementation. These domains (defined with specific examples in table 3.1) form the basis of the data model evaluation undertaken in this chapter. The scope of this evaluation is intentionally limited to the CC-HIC CDM itself, rather than data populated in the model, which is evaluated in Chapter 4 (page 107).

In many respects, the CC-HIC data model can be thought of as an extension to the third version of the ICNARC data model. The ICNARC data model focuses of storing demographic and physiological data at the point of admission to an ICU. The CC-HIC data model extends these concepts into the longitudinal domain, adding new concepts to cover the entire period under critical care. The CC-HIC data model is specified in its entirety by an XML Schema Definition (XSD); the blueprint against which Extensible Markup Language (XML) can be written.

The CC-HIC data model is an episode centric model[1], in that an episode has a strong structural representation, while all other concepts are considered attributes of an episode. In this way, a concept *cannot* exist in the CC-HIC data model, without being a component of an episode. Concepts within the CC-HIC data model are connected through a simple nested hierarchy; concepts are related via parent and child relationships.

[1] here and throughout, when referring to "episode" this is *always* shorthand for "an ICU episode"; a continuous period of critical care, within the same physical location.

Term	Description	Example
Completeness	Does the data model meet all research requirements?	The data model can express all concepts required by a prototypical study question
Integrity	Does the data model enforce relationships and constraints so that the original use of the data is represented without significant data loss?	The data model can express the relationships that are common in healthcare data without the need for excessive deduplication or deconfliction of data. Examples include the relationship between: samples and results, patient and episodes, hospital locations and patients
Correctness	Does the data model conform to good data modelling practices?	The data model is normalised to a degree that facilitates high quality research
Flexibility	Can new concepts be added, old concepts removed, or data representations changed to meet ongoing needs?	a previously undefined concept (like COVID-19 status) can be included without structural change to the data model, or significant resource requirements
Understandability	Is the data model and its contents understandable to an inclusive range of stakeholders (clinicians, scientists, statisticians, data engineers)?	Domain experts with training that is typical in the field can interpret the data model, without requiring significant additional training
Simplicity	Does the model add unnecessary complexity that could be removed?	The data model can be interrogated and data extracted by an appropriately trained researcher
Integration	Is the model consistent with the models employed by external collaborators or elsewhere in the organisation?	The data model consistent across HIC themes facilitating multi-domain collaborations. The data model uses international standards (such as OHDSI and SNOMED.)
Implementation	Is the data model implementable with current resource restrictions? Are there reasonable modifications of the model that would make it more readily implemented under the available resources?	The data model is written in a language that is straightforward to use in the secure restricted environment of the UCL DSH

Table 3.1: Summary definitions and examples from the data model evaluation framework developed by Kahn *et al* [162]

3.1 Completeness

The CC-HIC data model can be considered "complete" if it meets all the requirements to support the proposed research use. Primary cohort statements for the sepsis and hyperoxaemia studies are therefore provided against which the data model can be evaluated for potential utility.

3.1.1 Cohort Definition: Sepsis

The proposed sepsis cohort includes all index episodes for adult patients within the CC-HIC network with a diagnosis of sepsis within the first 24 hours following admission. This cohort definition allows maximum exploration of biomarker morphologies following the onset of sepsis. Patients with sepsis can be identified from an EHR by the triangulation of a subset of the following elements of the CC-HIC data model:

1. diagnostic codes indicating sepsis (or infection).
2. evidence for the onset of new organ dysfunction (which by definition requires consideration to be made for patients with pre-existing organ dysfunction, for example, recipients of dialysis).
3. treatment patterns suggesting deterioration in response to infection. For example, starting or escalating antibiotics.

Microbiological evidence of infection could also be used in this respect, however there are some flaws in how this concept is represented in the CC-HIC data model that likely preclude its use. Discussion of this will appear in greater detail in Section 3.2 (page 97).

The CC-HIC model is episode centric, and each episode is tagged with an attribute for the primary diagnostic code for the episode using ICNARC codes; a controlled terminology. This is an established standard used by most ICUs in the UK. These codes are specific to the ICNARC data submission process and are unlikely to be persisted within any EHR. The use of ICNARC codes may therefore incur an implementation penalty, as it seems likely that the CC-HIC data submission would require supplementation from an existing ICNARC data pipeline. Most

UK hospitals collate diagnostic information aligned to SNOMED and ICD-10. The CC-HIC data model could possibly have greater generalisability by using codes from these ontologies in addition to (or instead of) the ICNARC codes. Reliance on an ICNARC dependency may have the unintended consequence of restricting the research scope of the CC-HIC data model.

In order to identify organ dysfunction, the query would require the co-ordination of a number of concepts that are needed to calculate the SOFA score. The CC-HIC data model expresses most of these concepts straightforwardly, with the exception of ventilation which warrants further discussion. A number of concepts that are helpful in accurately describing ventilation are notably absent from the CC-HIC data model. This includes:

- the patient airway status, which is currently limited to the use of endotracheal or tracheostomy tubes only.
- the method of oxygen delivery is missing in its entirety. Examples include: facemasks, endotracheal tube, nasal cannulae etc.
- certain specific ventilator settings, most pertinently the different modes of ventilation.
- end tidal CO_2 monitoring, which is ubiquitous when delivering ventilation in the UK.
- certain intermediary methods of oxygen delivery are missing, such as high flow nasal cannulae. These are not strictly "ventilation" per se but form part of the gamut of respiratory therapies that are used in co-ordination with ventilation.

Further, concepts that are used to represent ventilation in the CC-HIC data model may not exist verbatim in the source EHR. For example, "ventilation" itself is listed as a concept within the CC-HIC data model with the permitted options: "invasive" and "non-invasive". Generally, ventilation is not documented in this way in a source EHR. Patient documentation is typically based upon directly observed features of the patient. It would have been preferable to have represented all these directly observed raw concepts that *are* typically present in the source EHR, over which

Figure 3.1: Episode centric (left) vs patient centric (right) data model views. In the episode centric approach, the patient becomes an attribute of each new episode, necessitating duplication of this information across episodes. Should an error occur in some data used for linking (for example, missing information as indicated by "XXX") the likelihood of a successful patient match is reduced. In the patient centric view, patient level information is abstracted away from the episode, and a direct relationship is instead maintained

a "ventilation phenotype" could be defined and validated against a subset of data. As a result, some misclassification bias may be introduced into any analysis when studying ventilation, as a base limitation to the CC-HIC data model.

Pre-existing organ dysfunction is well defined in the CC-HIC data model, and is a required component of basic data capture on all intensive care patients in the UK. Excluding the components of SOFA that show pre-existing organ dysfunction is well supported by the CC-HIC data model and should be trivial.

To focus on index admissions, readmissions must be excluded. This is dependent upon correctly identifying patients across repeat episodes. Since the patient is an attribute of an episode, patient details are duplicated across readmissions, as illustrated in figure 3.1. This increases the likelihood that redundant patient information would misalign between episodes. The patient specific attributes in the CC-HIC data model against which linking is possible include: NHS number, date of birth, hospital number and post code. All these elements are fallible as a method of linking, because they could be incorrect or missing between episodes; problems that are well known to feature in healthcare data [163, 164, 165, 166]. Patients may move house, or be assigned a new NHS number or hospital number as part of standard administrative reconciliation. As a result, readmissions can be identified, but the existing data model is susceptible to misalignments.

3.1.2 Cohort Definition: Hyperoxaemia

The proposed hyperoxaemia cohort includes all index episodes for adult patients within the CC-HIC network who can provide at least one arterial blood gas sample. This cohort definition provides a pragmatic cohort for the exploration of the potential association between the cumulative exposure to hyperoxaemia and outcomes.

The CC-HIC data model represents the components of a blood gas sample as separate (non-linking) concepts. Results from a single blood gas sample must therefore be connected post-hoc based upon their sample times. For each blood gas component, a meta-data label is used to assign the anatomical source as either arterial or venous. This approach leads to the duplication of anatomical information across many linked concepts, similar to the duplication of patient information across multiple episodes seen previously. This is a missed opportunity, since concepts from the same blood gas are invariably stored as a linked panel within the source EHR. This data representation places a high burden of responsibility on each contributing site to manually curate blood gas data. Downstream data quality evaluations will be necessary to ensure data integrity prior to use. Restricting the use of anatomical labels to only arterial or venous may also prove problematic, since capillary and extra-corporeal[2] sources are also common potential sources for these samples in UK ICUs.

3.1.3 Time Cadence Specification

The CC-HIC data model provides an unenforced requirement[3] that longitudinal data concepts are delivered hourly. This may be problematic for concepts that are naturally recorded at a higher frequency than hourly, for example, vital signs. This forces data engineers to make a decision to either ignore the requirement, or to down sample raw data to the hourly cadence. This encourages both divergence in how data is to be represented, and data loss between the source EHR and the data that analysts will see. There appears to be little advantage to this approach since data storage limits are not a research restriction.

[2] A sample taken from blood that is circulating through a device external to the body.

[3] In that no specific XML validation has been implemented to enforce this requirement.

3.1.4 Medicines Administration

The route and method of administration for a medication in the CC-HIC data model are expressed via their schema hierarchy. For example, propofol, a drug given intravenously as a continuous infusion is encoded as follows:

1. drugs.
2. CNS.
3. sedatives continuous infusion.
4. induction agent.
5. propofol.

Since a short infusion is synonymous with a bolus drug administration, and units are not specified in the data model, there is no way to identify if the numerical value supplied refers to either of the following patterns:

1. a bolus of a drug, where the numeric value indicates the total dose administered in mass of drug (e.g. micrograms).
2. an infusion of a drug over a defined period of time, where the numeric value indicates either the total dose infused or the rate of drug administration over a defined period.

This problem is illustrated in figure 3.2 where the same data is illustrated under these two different interpretations. The lack of units within the data model renders either interpretation challenging to implement in practice. In ICU, where many drugs are given as continuous infusions, but for different reasons, this approach renders the drug components of the CC-HIC data model at risk of misalignment between contributing sites, and misinterpretation by analysts. A preferred approach would be to reduce ambiguity by adding units and route of administration as explicit attributes of each medication, allowing these concepts to be represented in the CC-HIC data model without transformation or ambiguity. Additional scrutiny of drug administrations will likely be required before an analysis uses these data.

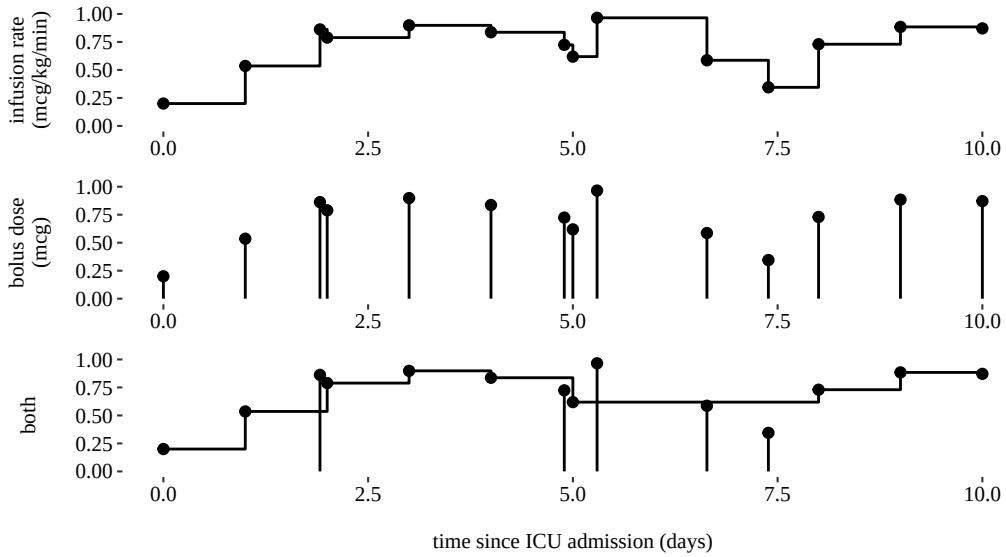


Figure 3.2: Theoretical example of drug infusion representation in the CC-HIC data model. Propofol is used as a motivating example, since it can be given as both bolus and continuous infusion. Top panel: values have been interpreted as a continuous infusion of drug, with each new value translating to a infusion rate change. Middle panel: values have been interpreted as bolus administrations of the drug. Bottom panel: original underlying data displaying a mix of administrations. In all cases, units remain unknown.

3.1.5 Representation of Missingness

One of the main challenges in EHR driven research is the presence of informatively missing data patterns [167]. Data in EHRs can be missing because:

- patients are predominately sampled during hospitalisation, coinciding with a period of acute ill health.
- sampling occurs more frequently in sicker patients.
- patients can move out of area, or seek treatment for a particular aspect of their health at different institution, basing that decision to move partly or wholly on their health.
- salient negatives are not typically captured in a *structured* way in an EHR. Examples would include:
  - past medical history conditions are recorded if they are present, not if they are absent.
  - a useful counter example is smoking status, which *is* routinely recorded

if the patient is a non-smoker. This is atypical and occurs in this instance because there are financial incentives available for UK hospitals to document smoking cessation.

A key difference in requirement between a data model to support EHR driven research, and a data model to support a disease registry is how missingness is represented. In a disease registry, where data is collected by hand, salient negatives like "patient does not have asthma" can be collected and made explicit. When moving to a data model to support larger scale EHR research, implicit missingness becomes the norm, and one must generally assume that missing data is synonymous with a negative. This approach can have deleterious consequences to inferences [9].

Regardless of the interpretation of missing data, it is preferable to have some consistency within a data model as to how missing data should be handled. The CC-HIC data model does not represent missingness in a consistent way. Some comorbidities are listed explicitly, with both positive and negative assertions. It is likely that in many cases, the negative assertions will not be possible to complete (since they are not stored in the source EHR). This may cause confusion during analysis over what constitutes a confirmed negative, and what constitutes missing data.

3.1.6 Specificity to Intensive Care

The CC-HIC data model is tailored specifically to an application in intensive care. Examples of this specificity include:

- the use of ICNARC diagnostic codes.
- key events are linked to the ICNARC temporal schema.
- comorbidities are limited to those of the ACPAHE-II score.
- the base unit of the data model is an ICU episode, with any other structural hospital elements (like hospital admission time) expressed in relation to that episode.

This is a side effect of the CC-HIC data model being developed as an extension to the ICNARC data model. This focus on the episode centric view limits the

potential usefulness of the data model, including for intensive care research. In both sepsis and hyperoxaemia studies to follow, it would have been useful to have seen the hospital events that led up to ICU admission. If sepsis is the reason for admission to the ICU, then its onset will likely have physiological markers that exist prior to admission to the ICU. Patients will likely receive a non-ignorable amount of oxygen in the lead up to their ICU admission. At present, the CC-HIC data model is aligned to the CC-HIC project governance, in that only data pertaining to intensive care is to be collected. There would be numerous advantages to broadening the data model in this respect, but we should acknowledge the non-trivial changes to project governance and data sharing agreements that this would require.

3.2 Integrity & Correctness

Integrity and Correctness are reviewed jointly as they are overlapping terms as applied to the present scenario. Integrity reviews the relationships and constraints that are enforced by the data model. Correctness reviews whether or not the data model conforms to good data modelling practices including appropriate normalisation. Is it possible for the data model to faithfully represent the natural biological relationships that exist in the target data, without information loss?

3.2.1 Model Normality

The CC-HIC data model is an episode centric model that conforms to Boyce-Codd unnormalised form (UNF) [168]. Data normalisation is a set of principles used to improve the quality of data management. Specifically, the goals of data normalisation are to:

1. reduce data duplication.
2. prevent data anomalies.
3. ensure referential integrity.
4. simplify data management.

The Boyce-Codd categorisation is used to define levels of normalisation from UNF (unnormalised form) to level 6NF (6th normalised form). Each additional

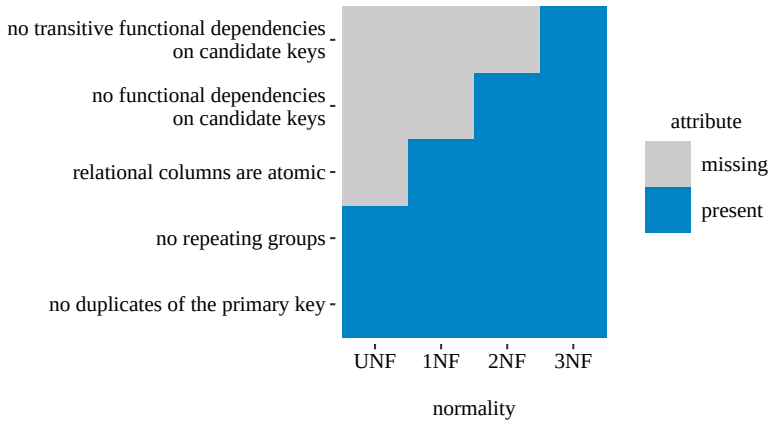


Figure 3.3: Boyce-Codd levels of normalisation. Each additional level of normalisation adds a new attribute to faciliate the primary objectives of data normalisation.

level of data normalisation automatically includes the features of the previous level, while adding additional requirements. 3NF is generally considered a robust grade of implementation, and is the standard expressed by the OHDSI common data model [17]. The attributes associated with each level of normalisation from unnormalised to third normalised form are shown in figure 3.3. With respect to the cohort definitions previously provided, and the likely more general case uses of the CC-HIC platform, the UNF representation of the CC-HIC data model is suboptimal.

In adopting UNF, the CC-HIC data model expresses some relationships that do not exist, while lacking appropriate referential integrity for important relationships that do. As an illustrative example, episodes contain the attribute of:

- an NHS number.
- a hospital death.
- a GP code.
- physical characteristics such as sex, height and weight.

Naturally, these concepts belong to a person and not an episode. When patients are re-admitted, move between physical ICU locations, or transition between levels of care (events that are all commonplace), duplication of information occurs, and therefore data reconciliation must take place. A primary goal of data normalisation is to make this form of reconciliation unnecessary.

Similarly, episodes are forced to carry semantic representations of systematic

healthcare components. For example, the bed configuration of an ICU is attached to each and every episode. In reality there is no meaningful representation of this concept at the episode level, and so in practice it is left incomplete by contributing sites.

In a best case scenario, UNF is an inefficient method to store these data. In practice, it encourages conflicting data patterns, necessitating manual data reconciliation often without enough local knowledge to make appropriate decisions.

Microbiology & Sample - Result relationships

A commonly encountered relationship in healthcare data is the "sample" to "result" relationship. This occurs when a patient is sampled (blood, tissue, vital signs etc.) and a time delayed result is reported back. The CC-HIC data model contains no support for these common relationships, and so only a single time is afforded to any time varying data concept.

This "sample" to "result" relationship is overtly problematic for a number of areas. The problems associated with blood gas data have previously been discussed. The problem is critical for microbiology data. A lag of several days commonly exists between a microbiological sample being taken and the result being made available. Other data models that do express the "sample" to "result" relationship readily demonstrate the discrepancy in the timings of these events. This is shown in figure 3.4 for a sub-sample of microbiological events from the MIMIC IV data model.

Microbiological data concepts are expressed within the CC-HIC data model as four elements, connected through a one-to-one relationship:

1. organism.
2. antimicrobial sensitivity.
3. anatomical site.
4. date and time of microbiology culture measurement.

Since the relationship between microbiological data concepts in HIC is one-to-one, there is no reasonable way to represent the following characteristics of these

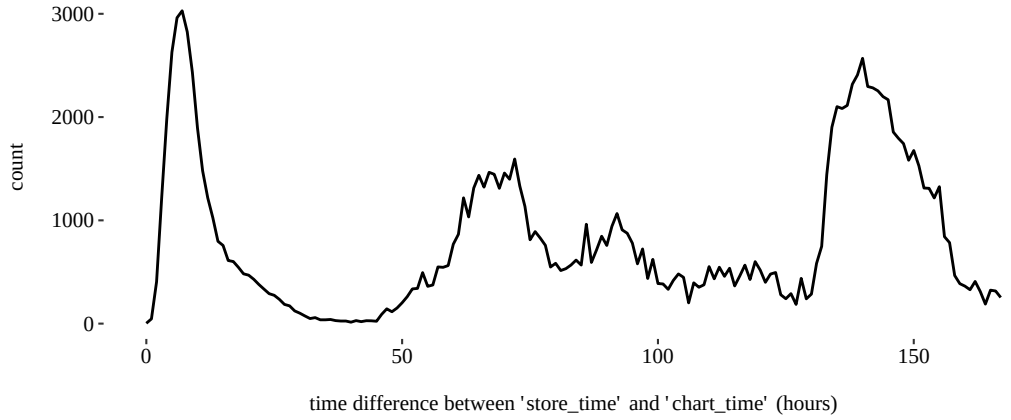


Figure 3.4: Time difference between "sample" to "result" as found in the MIMIC IV database. Time differences are from the "store_time" and "chart_time" for values which represent when concepts were first entered into the EHR, and then subsequently clinically validated when cultures became available.[4]

data:

1. the difference between sampling time and reporting time.
2. organisms with multiple sensitivities to different antibiotics.
3. samples that have been sent but are yet to be reported.
4. samples that have been sent but have yielded negative results (which is typical). This final point is possible to represent, but encourages missing data patterns, with an inability to distinguish the origins of the missing data.

This example in particular draws attention to the limitations of the CC-HIC model for concepts that have non-trivial relationships; highlighting the challenges faced when translating complex clinical data into a CDM.

3.2.2 Datetime attributes

The CC-HIC data model makes a semantic distinction between time varying and time invariant data. For example, height and weight are given time invariant qualities (they cannot change within the model). This is not directly problematic unless an analysis is required in which these items are anticipated to change. Because these elements have been hard coded into the data model, any changes to represent time

[4]With thanks to Lingyi Yang from the Oxford DataSig research group for identifying and contributing this extract from MIMIC IV.

invariant data concepts as time varying would require a modification to the schema, likely leading to backward compatibility issues. An analysis that requires time invariant concepts to be captured in time series would not be able to proceed under the current specification. The most likely candidate data element for which this is a concern is patient weight, which will change dynamically depending upon the fluid status of the patient. This is a routine measurement in many medical disciplines, particularly where the prognosis of dialysis patients is understood to be worse in patients who experience rapid changes to their weight between filtration sessions [169].

3.2.3 Non-standard Data Representation

There are several areas in the CC-HIC data model where data concepts are given non-standard representations. An example is the representation of binary concepts, which are variously encoded with the character representation of "0, 1", "N, Y" or even "1, 2". Table 3.2 tabulates the occurrence of binary representations within the CC-HIC data model.

Binary representation	Count
0, 1	31
Y, N	2
D, T	1
F, M	1
1, 2	2

Table 3.2: Binary representations implemented in the CC-HIC data model. The broad range on offer imposes cognitive burden for all users of the data model, and potentially increases the likelihood of errors during analysis.

The problems with the lack of standardisation are evidenced by the schema violations that are common with respect to binary data in data submissions. Writing queries against these data are cumbersome, as one must learn the exact representation of each and every concept in the data model.

3.3 Flexibility

Flexibility reviews whether new concepts be added, old concepts removed, or data representations changed to meet continuing project needs.

All data concepts in the CC-HIC data model are hard coded into the XSD. As such, there is no capacity for the data model to represent a data concept that has not been explicitly written into the data model. For brevity, I will refer to this style of data specification as "strongly specified". Strongly specified in this context means a data model that requests a specific data concept, describes exactly the representation it can take, and does not allow for any data concepts outside this representation. This is counter to a "weakly specified" model, that may make a more general request for "all patient physiology" so long as it is expressed in an appropriate standard. A strong specification can be advantageous, especially when it is important to represent missingness in an explicit manner. In the context of secondary use research, this approach does not scale well for two important reasons. First, as previously highlighted, EHRs do not generally record negative assertions of clinical details in a structured way. Second, the potential pool of candidate data concepts in a modern EHR number in the hundreds of thousands. Specifying each one by hand is likely an impossible task, given even the most generous resources.

Adding new concepts to the CC-HIC data model has proven to be impossible with currently available resources. Any new data concepts would necessitate a change to the XSD, and most likely require a re-write of the data pipeline from each contributing site. During the COVID-19 pandemic, this issue was directly encountered. Since there was no means through which a new concept could be easily added to the data model, there was no way to define which patients were infected with SARS-CoV-2[5]. Further, new temporary "surge" wards admitting level 2 care and level 3 care patients were being created at the source hospitals in a dynamic fashion. Anaesthetic recovery units, operating theatres, and wards were being converted to provide higher level care to patients with COVID-19. Patients on level 1 care wards also started to receive therapies that would normally have been administered within a level 2 care or 3 unit. These structural elements could not be added to the data model, and so a selection bias had therefore been introduced into data contributed to the CC-HIC, through the exclusion of patients attending these new

[5]The ICNARC coding method does allow for suspected or confirmed "pandemic influenza" which could have been co-opted for use.

surge wards.

Ultimately, modification of the CC-HIC data model proved too great a challenge given the available resources, and no modifications were implemented. The importance of considering the flexibility of a data model is brought sharply into focus by these events. Although data sharing permissions for the CC-HIC were in place at the time of the COVID-19 pandemic, the inflexibility of the data model prevented any access to these data.

3.4 Understandibility

Understandability reviews whether the data model and its contents can be understood by an inclusive range of stakeholders, for example clinicians, scientists, statisticians and data engineers.

The XML representation of the CC-HIC data model is complex. In particular, there are many superfluous nested levels, while tightly connected concepts (for example, a heart rate, and the time that the heart rate was measured) are stored in logically discrete compartments. This makes reviewing the XML challenging, particularly within the UCL data safehaven (DSH) where specialist tools for reviewing XML are not available. This presents a barrier to analysts and clinicians who typically have vastly more experience working with flat files or relational databases, than with complex nested data structures. This barrier set the early project goal to develop the cleanEHR package, the primary function of which was to remove data from the XML data model so that it could be analysed.

3.5 Simplicity

Simplicity reviews whether the data model adds unnecessary complexity that could be removed, or lacks complexity where it is necessary.

The previous arguments highlight the inherent lack of balance in simplicity for the CC-HIC data model. In an ideal situation, a data model needs to be only as complex as is strictly necessary to support the scope of research. Fundamentally, the balance of complexity for the CC-HIC data model is not well calibrated to its research goals. Examples of this mis-balance are highlighted in table 3.3

Too simple	Too complex
Unnormalised form	Disaggregation of date time data concepts
No support for sample-results	Mixed representations of basic data elements e.g. binary data representations
	Bespoke non-standard data hierarchies

Table 3.3: The balance between design simplicity and complexity. For many important aspects of the CC-HIC data model, the balance is suboptimal to support research. Many areas are over engineered without any return for the researcher, while important structural elements are missing.

3.6 Integration

Integration reviews the alignment of the data model to other models and domains within the organisation.

3.6.1 Semantic Interoperability

Controlled clinical terminologies exist to help define healthcare concepts in precise terms. The most widely implemented medical ontology is systematised nomenclature of medicine (SNOMED) [170, 171]. In healthcare research, alignment to semantic standards or "semantic interoperability" is vital to ensure that two organisations are referring to the same data concept when intending to do so. If ambiguously specified, multiple sites may contribute a data concept in different ways and systematic differences could be introduced to the data as a result.

To illustrate this issue, a common biomarker—C-reactive protein (CRP)—is represented across four themes of the Health Informatics Collaborative in table 3.4. Two themes do not contribute the data concept. The remaining two contribute the data concept in very different ways. The differences are as follows:

- CRP is identified by a different code at each site.
- date time information is represented differently.
- units are defined for one site, and not for the other.
- lab reference ranges can be submitted for one and not for the other.
- one submits CRP as a numerical value, the other as a string.
- neither makes reference to an external controlled. terminology

The differences on display for a single common biomarker are extensive. Mul-

CRP Attribute	Critical Care		Acute Coronary Syndrome	
	Reference	Data Class	Reference	Data Class
Value	NIHR_HIC_ICU_0557	Numeric	NHIC_ACS_91Crp Result	String
Units	-	-	NHIC_ACS_91Crp Unit	String
Date	NIHR_HIC_ICU_0800	DataClass	NHIC_ACS_91Crp Collected date	String
Time	NIHR_HIC_ICU_0800	DataClass	NHIC_ACS_91Crp Collected time	String
Other	-	-	Upper and lower bounds of the test assay also recorded	

Table 3.4: Examples of how CRP is representated across different themes of the HIC.

tiplied across the whole data model, this imposes excessive resource requirements on each contributing site, as each data extract must be bespoke to every theme of the HIC. Any collaboration between themes would require an extensive data mapping exercise, following which an agreed common data representation would *still* be required for research to take place. This approach is quite distinct to the goals of semantic interoperability.

3.7 Implementation

Implementation reviews if the data model is implementable with current resources, or if any reasonable modifications could be made to facilitate implementation. Since the Kahn evaluation was orientated toward a model appraisal prior to sharing data, this section is perhaps less relevant after the fact.

Following implementation of the XML pipeline, many sites were unable to provide continued support as errors were found. Instead, bespoke data patches as CSV files were favoured that could be integrated into the CC-HIC research database centrally. From a research provenance perspective this was not ideal, but was necessary for the research goals to be viable. This highlights the high resource cost associated with implementing the XML pathway within NHS organisations.

3.8 Conclusions

Overall, the CC-HIC data model evaluates suboptimally against the Kahn data model evaluation framework. As a positive, the data model was developed with reference to an existing gold standard data model (the ICNARC data model) that itself has supported more than a decade of high quality research in intensive care. It

would have been reasonable to expect that the CC-HIC data model—as a pragmatic extension to the ICNARC model—would have evoked a more positive evaluation. However, many of the features that made the ICNARC data model perform well, have not translated to support the representation of a critical care EHR in a research ready format. In particular, the CC-HIC data model was developed without reference to international standards in data modelling or controlled ontologies, both of which were in widespread use at the time of its creation.

Owing to the constraints imposed by the data model, caution needs to be exercised where clinical research is conducted with the CC-HIC data model. Rather than enable or facilitate research, the CC-HIC data model could impede such endeavours by failing to protect against erroneous data representations. A formal evaluation of data quality is therefore essential, which follows in Chapter 4.

Implementation of the CC-HIC data model in XML was a laudable goal. In practice, this created more problems for implementation than it has solved. XML adds a layer of complexity, and lacks familiarity amongst the NHS data engineers who are required to support the data pipeline. This layer of complexity is difficult to justify, since many of the special features of XML (for example, complex validation rules) have not been implemented.

In view of the changes that have been observed to the delivery of critical care during the COVID-19 pandemic, it is likely that new features will need to be supported by the CC-HIC data model. This includes adding new data concepts and relationships. It would be prudent to review whether or not extending the CC-HIC data model is a worthwhile endeavour, given its shortcomings. It is the recommendation therefore of this evaluation, that the CC-HIC data model should be scheduled for discontinuation, in favour of implementing an established international standard, for example the OHDSI common data model.

Chapter 4

Data Quality Evaluation & Extraction

It is paramount that prior to conducting any clinical research on secondary use data, its quality must be systematically evaluated. Researchers are often distant to the point of data entry and may not posses full knowledge of the clinical origins of healthcare data or the peculiarities of the research data pipeline. As a result, they may be content to perform superficial checks of summary statistics without a systematic approach or underlying theoretical basis to support the discovery of common errors that plague routinely collected healthcare data. Sections 4.1-4.2 provide the theoretical background for data quality evaluation from the motivation of systematic error discovery. inspectEHR [6] is introduced in Section 4.3 as software developed as a contribution to this thesis for the practical implementation of this theory. wranglEHR [5] is introduced in Section 4.4 as software developed as a contribution to this thesis for the standardised extraction of data from the CC-HIC research database. Principal findings from the data quality evaluation are provided in Section 4.5, with motivating examples provided to illustrate errors that are common, important, or have particular relevance to the clinical studies that follow. Section 4.6 provides a discussion of the quality of data within the CC-HIC research database, and contrasts the CC-HIC as a research platform with other similar research projects. Concluding remarks and a set of recommendations for future improvements are provided in Section 4.7.

4.1 Background

The terms "quality control (QC)" and "quality assurance (QA)" are frequently applied, particularly in the adjacent field of healthcare data engineering [172]. I have chosen to avoid this terminology. Problematically, one cannot "control" the quality of data outside the source EHR if the error exists within the EHR itself. Control in this sense implies cleaning or removal of erroneous data patterns, which may introduce bias into the cohort [173, 174]. Decisions over which data to exclude from analysis are best left to the analysis stage itself. In a best case scenario, the research database is a *perfect* reflection of the source EHR, including all the errors that the source EHR contains. Since the underlying source EHR is both dynamic and complex, it is inevitable that new errors will manifest, possibly even in areas thought to be previously of "high" quality. "Assurances" of data quality are therefore unlikely to be permanent. Instead, I favour the term "quality evaluation (QE)", since it is a more accurate description of the process that follows. In this spirit, the goal of QE is *not* to:

- facilitate the production of a "clean" or "cleansed" dataset.
- provide assurances that the source data are of high quality.
- control the quality of data.

Rather, the specific goal of QE as presented is to systematically identify and label data with their potential sources of error. This provides a common benchmark—based on expert clinical domain knowledge—against which analysts can make decisions on what data require modification, what to exclude, and what the consequences of such actions might be.

Figure 4.1 shows the current CC-HIC data pipeline, with the points where current QE takes place. This includes:

1. source validation. Confirmation by local sites that the extracted XML meets the data model specification and that data are an accurate reflection of the EHR.
2. central XSD validation. Technical validation that contributed XML conforms to the specification of the CC-HIC data model XSD.

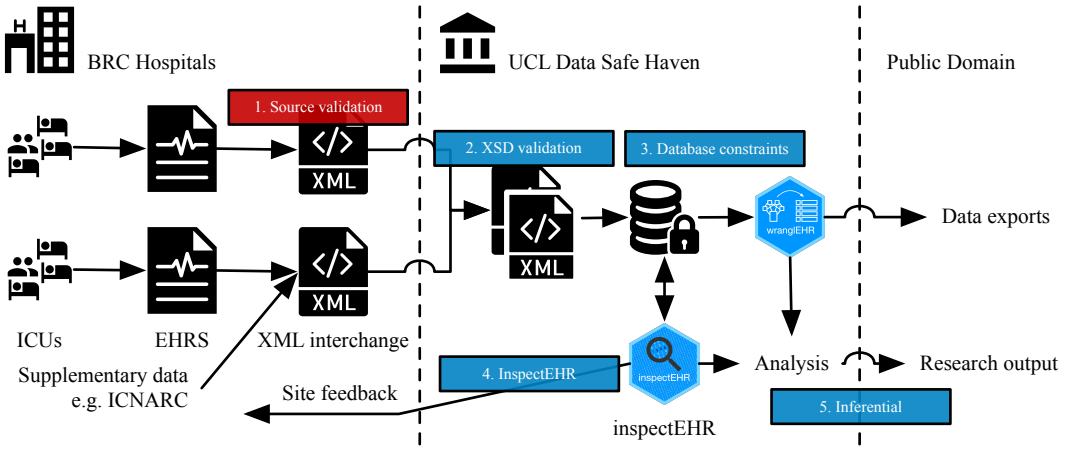


Figure 4.1: The current CC-HIC QE pipeline. Areas where data QE take place are shown in blue boxes. Places where there is missed opportunity for data QE are shown in red.

3. application of database constraints. XML data can be read into the research database without producing errors related to data type conformity or referential integrity.
4. central systematic QE. Implemented by inspectEHR [6] to be discussed within this chapter.
5. inferential. Ongoing (and often informal) QE performed when analysts investigate the multivariate properties of data through modelling. For example, the reporting of spurious findings.

inspectEHR [6] is a data QE software package written as a software contribution to this thesis. The primary objectives of inspectEHR [6] are to:

1. provide a comprehensive and transparent evaluation of data quality within the CC-HIC research database.
2. non-destructively append research data with QE meta-data labels thus facilitating:
    - a consistent approach for interpreting data quality.
    - standardisation of analyst work flows within an environment where erroneous data are common.
    - persistence of data quality labels for reusability and reproducibility.

4.2 Methods

The theoretical basis for the QE implemented by inspectEHR [6] is drawn from the standardised framework proposed by Kahn *et al* [175][1]. This framework consolidates much of the prior research in this field from the Data Quality Collaborative (DQC) [176], as well as the The Observational Health Data Sciences and Informatics (OHDSI) data validation framework, known as Achilles [177]. The design principles of inspectEHR [6] itself have been heavily influenced by The Turing Way [178]. The Turing Way is an open source set of guidelines designed to facilitate high quality reproducible data science. An overview of how the end user is expected to interact with inspectEHR [6] is provided in the software vignette in appendix Section C.1 (page 275).

4.2.1 Kahn Evaluation Framework

The Kahn evaluation framework encompasses three fundamental domains:

1. *conformance*: does data adhere to appropriate standards and formats?
2. *completeness*: are data present as expected?
3. *plausibility*: are data believable within their context?

These are evaluated through two processes:

1. *validation*: the use of an external gold standard data source that can be used for corroboration.
2. *verification*: internal checks of data so that they meet a particular standard or expectation.

Once data have been extracted and removed from source, the ability to perform validation is limited. It follows that local source validation is a vital stage in developing a high quality research database. However, this, by definition, is difficult to implement in a central capacity. In lieu of a central gold standard resource or access to local data, the multi-centre nature of the CC-HIC can be used to implement statistical validation. In this way, each site acts as a control for the others,

[1] This is the same research team responsible for creating the CDM evaluation framework reported in the previous chapter.

and various properties of data can be compared. Any differences will logically either be attributable to case mix variation or error[2]. For continuous data, this can be performed by applying the Kolmogorov-Smirnov (KS) test, as shown in equation 4.1.

$$D_n = \sup_x |F_n(x) - F(x)| \tag{4.1}$$

Where D_n is the KS test statistic, and $F_n(x)$ and $F(x)$ are two continuous empirical cumulative distribution functions to be compared. The KS test statistic is a non-parametric measure of the maximal distance between two empirical cumulative distribution functions. It returns a value between 0 (identical distributions) and 1 (maximally different distributions).

4.2.2 Implementation and Error Classification

The Kahn evaluation framework is comprehensive, though it lacks a formal system for implementation. In order to implement the Kahn framework, it has been necessary to develop a classification system for errors. In the present context, two orthogonal systems of classification can be used to describe errors:

- *by origin*: do the errors *exist* in the source EHR, or are they created within the research pipeline by the Extract Transform Load (ETL) process, or
- *by existence*: is the error caused by data that are missing, or data that are present but incorrect.

This proposed system of classification (presented in table 4.1) is, to my knowledge, unique (although derived from other similar data QE processes [172]). The classification is a practical one because there are competing requirements between how captured errors need to be stored within the research database as meta-data, and what actions analysts will want to take based upon the errors.

The error *origin* classification system categorises errors by how they arise in the EHR research pipeline into "source errors" and "transcription errors". Source errors occur when research data *are* a true reflection of the source EHR. Source

[2]I use error here in the broadest possible sense. This may include for example, differences in the source EHRs such that one system is able to capture information where another might not.

		Error existence	
		Missing	Present
Error origin	Transcription	Data omitted from the ETL (e.g. project time constraints or accidental omission)	Data modified on ETL (e.g. transformation error)
		Data forcibly omitted (e.g. not supported by data model)	Data forcibly modified (e.g. data partially supported by model requiring work-around)
	Source	Data not captured on patients (e.g. local capture of RRT is on paper, not EHR)	Data stored in EHR incorrectly (e.g. heights semantically listed in cm with range from 0-1.8)
		EHR build error (e.g. drop down menu not presented to user at correct time)	EHR build error (e.g. weights of neonates added to maternal record)

Table 4.1: Proposed classification of secondary use errors. RRT = Renal Replacement Therapy

errors reflect real but erroneous occurrences in the source EHR. In many instances, this may reflect a systematic error in the source EHR that may need to be corrected and propagated through the research pipeline[3]. While the errors continue to exist in the source EHR, they should be similarly preserved in the research pipeline and appended with meta-data labels to highlight their known error state. Transcription errors occur when research data *are not* a true reflection of the source EHR. Transcription errors form as part of the research ETL somewhere between the source EHR and research database. Since these errors are part of the data extraction process itself, they risk introducing bias into downstream research that could result in invalidation of research findings. Transcription errors are therefore *critical* to identify and correct as a matter of priority. Transcription errors are also to be found where data loss has occurred as a result of the limitations of the CC-HIC data model itself, as previously outlined. In this case, some data loss is inevitable, as the CC-HIC data model cannot fully express all features of clinical data as they exist in reality.

In the orthogonal error *existence* classification system, errors can be identified because data are suspected or confirmed to be missing, or data that are present but

[3]Indeed, the discovery of such source errors may also require that corrections are propagated into the live production limb of the EHR.

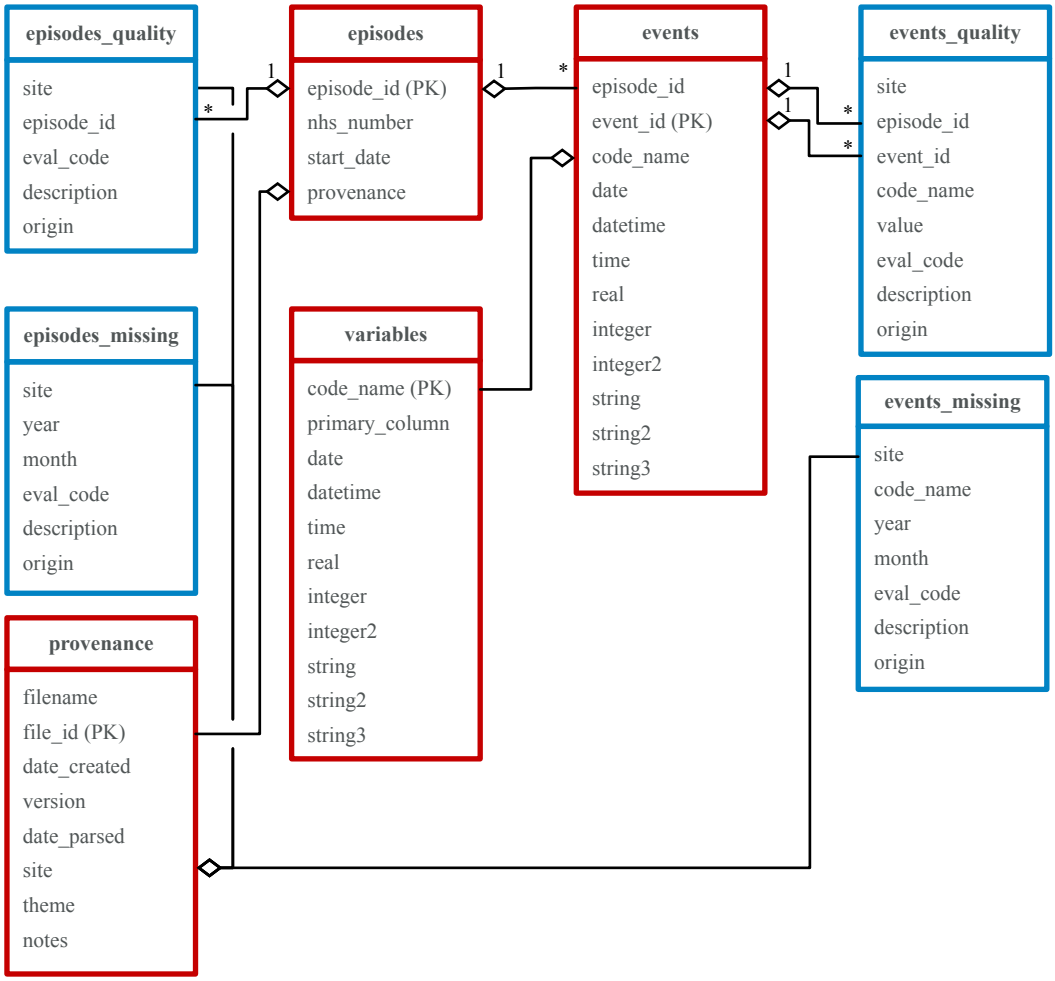


Figure 4.2: Extension to the CC-HIC schema for persistent storage of data QE meta-data. Some connections are omitted to improve readability. Each data quality table contains a column for "origin" to represent errors as transcription or source.

erroneous. This is a practical distinction, for the purpose of persisting error codes as meta-data labels in the research database. Since by definition, missing data will have no linking row within the research database, these missing data require a different representation in the research database. This alternate view is based upon observed time periods where data are suspected to be missing. I have extended the CC-HIC data model schema to allow persistence of data quality meta-data as shown in figure 4.2.

Four new tables are added to the database schema, based on whether or not data are missing, or present but in error:

1. **episodes_quality**. Data quality meta-data labels for episodes.
2. **episodes_missing**. Data quality meta-data labels for the temporal regions where it is likely that episodes are missing, based on prior contribution patterns.
3. **events_quality**. Data quality meta-data labels for events.
4. **events_missing**. Data quality meta-data labels for the temporal regions where it is likely that events are missing, based on prior contribution patterns.

Tables with the **quality** suffix make reference to data concepts that are contributed with error. Tables with the **missing** suffix attempt to qualify the time periods over which data are suspected to be missing.

Following the creation of these tables it is straightforward, via a table join, for analysts to take actions based on each of the data quality codes that they store. Since all data evaluation codes are a positive assertion that a quality concern has been raised, any research data that does not have a companion row in a data quality table can be safely assumed to have passed through all procedures without issue; no news is good news.

While it is not yet implemented, these data quality tables reserve the "origin" column for whether errors are thought to be transcription or source related. This is not yet implemented as it would require tight coordination with contributing sites and addition to the CC-HIC data model. This modification would be important in the long run so that analysts could identify where erroneous data patterns represent real data in the EHR, without needing to communicate with each contributing site.

4.3 inspectEHR

inspectEHR [6] is a software package written in the R statistical computing language as a contribution to this thesis. It is written in the tidy style, follows best practices from the R Studio development team where possible, and implements the modern paradigm of tidy evaluation [179, 180]. inspectEHR [6] uses the S3 class

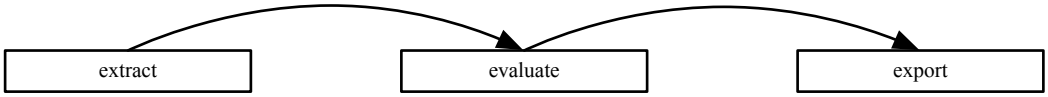


Figure 4.3: A simplified overview of the extract, evaluate and export paradigm followed by inspectEHR [6]. Data are extracted from the research database and standardised. Evaluation takes place by R in memory. Results are standardised and exported as meta-data labels to be appended to either a **quality** or **missing** table depending upon their context.

system of method dispatch to provide a clean user interface while performing data quality evaluation.

inspectEHR [6] follows an extract, evaluate and export paradigm as depicted in figure 4.3. Data are extracted from the research database and standardised so that all primary data and meta-data (including any salient date time information) are transformed from the sparse EAV structure of the research database and into a dense rectangular data frame suitable for further analysis. Other salient information required for downstream processing is appended to the extracted data frame as attributes. This includes the designation of the extracted data item and its class as assigned by inspectEHR [6]. Classes are implemented within inspectEHR [6] to allow S3 method dispatch to take appropriate action based on the fundamental properties of an extract data item, while abstracting away the complexities of the process to help ensure that the codebase of the package is straightforward to maintain (table 4.2).

The standardised extracted data frame is passed to a number of evaluative functions, each designed to evaluate a specific facet of the Kahn data quality evaluation

Data type	Temporal component	inspectEHR class	Example
Integer	Static	integer-1d	Recent steroid use
Real	Static	real-1d	Height
String	Static	string-1d	Ethnicity
Date	Static	date-1d	Date of death
Time	Static	time-1d	Time of death
Datetime	Static	datetime-1d	ICU Admission
Integer	Time varying	integer-2d	Heart rate
Real	Time varying	real-2d	Central venous pressure
String	Time varying	string-2d	Organism

Table 4.2: inspectEHR [6] class system

framework. Each evaluative function takes as an input the standardised data frame produced from the data extraction previously described, and produces as an output a standardised data frame that can be written back into the research database as either a **quality** or **missing** table. If an evaluative function finds no errors then an empty table with the correct column headers and class labels are returned for consistency. Each found error is assigned a code that corresponds to the Kahn evaluation framework. These codes follow the format "process-domain-number" (PP-DD-##) where:

- process (PP) = either VA (validation) or VE (verification).
- domain (DD) = a two letter short code for the specific Kahn evaluation domain being evaluated.
- number (##) = an unique code for the specific evaluation within inspectEHR [6].

The **quality** or **missing** tables produced at the end of the quality evaluation process have their own validating functions that are evoked prior to writing out to the research database. A schematic overview of this process is outlined in figure 4.4. During each data extraction, a series of diagnostic plots are produced as side effects and captured external to the research database in a user defined location. Examples of these plots are contained throughout the rest of this chapter, each of which was produced by inspectEHR [6] during a live evaluation.

4.3.1 Implementation of the Kahn framework

A detailed breakdown of all the QE procedures currently implemented against the CC-HIC research database are included in tables 4.3 and 4.4 representing the validation and verification procedures respectively as previously described. These tables list all the areas that are covered by the Kahn framework, including those that are *not* presently implemented in inspectEHR [6]. Procedures that have been implemented are accompanied with a unique implementation code and indicated in the tables.

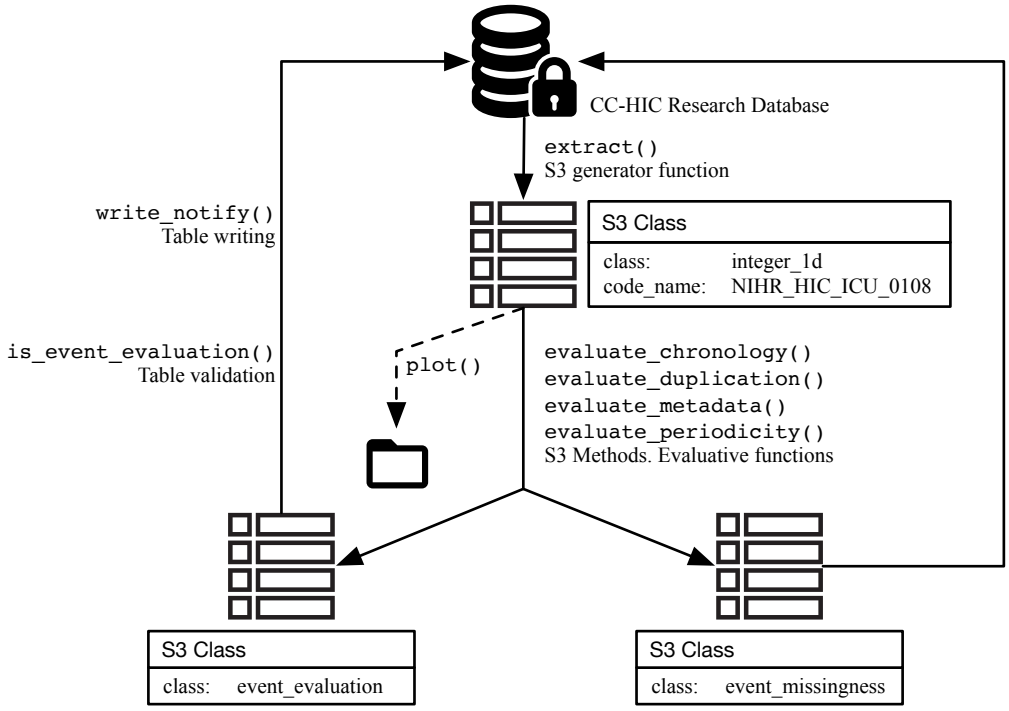


Figure 4.4: Schematic overview of inspectEHR [6]. Data is extracted from the research database using the extract S3 generic. This creates an object with an S3 class corresponding to one of 7 prototypical classes supported by inspectEHR [6] (see table 4.2 for details). The extracted data object is passed to a series of evaluative methods. Each evaluative method performs a specific data quality check and must output a data frame of class "event_missingness" or "event_evaluation". These data frames are then validated and exported back into the research database where they persist for re-use. This entire pipeline is automated from a single function call exposed to the end user. Further details of how the end user is expected to interact with inspectEHR [6] can be found in appendix Section C.1 (page 275).

Classification	Definition	Implementation
Value Conformance (VA_VC_01)	Data values conform to representational constraints based on external standards.	Values that have a specific external stanard are validated against that standard. This includes: NHS number, Post Code, ICNARC diagnostic codes
Relational conformance (VA_RC_01)	Data values conform to relational constraints based on external standards.	Structural data missingness conform to agreed schema
Computational conformance	Computed results based on published algorithms yield values that match validation values provided by external source.	
Completeness (VA_CP_01)	The absence of data values at a single moment in time agrees with trusted reference standards or external knowledge	Missingness of a particular item from the requisite schema at site level
Completeness	The absence of data values measured over time agrees with trusted reference standards or external knowledge.	
Uniqueness Plausibility	Data values that identify a single object in an external source are not duplicated.	
Atemporal plausibility (VA_AP_01)	Data values and distributions (including subgroup distributions) agree with trusted reference standards or external knowledge.	Values share similar distributions compared between sites
Atemporal plausibility	Similar values for identical measurements are obtained from two independent databases representing the same observations with equal credibility.	
Atemporal plausibility	Two dependent databases (e.g., database 1 abstracted from database 2) yield similar values for identical variables.	
Temporal plausibility (VA_TP_01)	Observed or derived values have similar temporal properties across one or more external comparators or gold standards.	Values share similar temporal distributions compared between sites
Temporal plausibility	Sequences of values that represent state transitions are similar to external comparators or gold standards.	
Temporal plausibility	Measures of data value density against a time-oriented denominator are expected based on external knowledge.	The sampling frequency distribution of all data items is uniform across sites

Table 4.3: Kahn data quality evaluation framework: validation process. Where domains have been implemented within inspectEHR the accompanying evaluation code has been provided.

Classification	Definition	Implementation
Value Conformance (VE_VC_01)	Data values conform to internal formatting constraints.	Data items are contributed with the correct data type
Value Conformance (VE_VC_02)	Data values conform to internal formatting constraints.	Categorical data items are contributed with the correct string representation
Value Conformance (VE_VC_03)	Data values conform to allowable values or ranges.	numeric data falls within limits of possibility
Value Conformance (VE_VC_04)	Data values conform to allowable values or ranges.	categorical data falls within allowable categories
Value Conformance (VE_VC_05)	Data values conform to allowable values or ranges.	date, date-time and time data falls within possibility
Value Conformance (VE_VC_06)	Data values conform to allowable values or ranges.	Episodes cannot be overlapping
Relational conformance (VE_RC_01)	Data values conform to relational constraints.	Referential integrity of the database (evaluated by application of database schema)
Relational conformance (VE_RC_02)	Unique (key) data values are not duplicated.	Primary key integrity of the database (evaluated by application of database schema)
Relational conformance	Changes to the data model or data model versioning.	
Relational conformance (VE_RC_04)	Changes to the data model or data model versioning.	Events originate from episodes that have passed QC
Computational conformance	Computed values conform to computational or programming specifications.	

Classification	Definition	Implementation
Completeness (VE_CP_01)	The absence of data values at a single moment in time agrees with local or common expectations.	Episodes must have a reconciled end datetime
Completeness (VE_CP_02)	The absence of data values at a single moment in time agrees with local or common expectations.	Missingness patterns in data items are a function of casemix, and not true missingness
Completeness (VE_CP_03)	The absence of data values measured over time agrees with local or common expectations.	Episodes do not originate within sectors of time that are sparse for other contributions
Completeness (VE_CP_04)	The absence of data values measured over time agrees with local or common expectations.	Missingness patterns in data contribution over time, controlled by each site
Completeness (VE_CP_05)		Metadata are present and complete
Uniqueness Plausibility (VE_UP_01)	Data values that identify a single object are not duplicated.	Patient level singular events are not duplicated across episodes. Examples include: episode start and end time, death, withdrawal time, body removal time
Uniqueness Plausibility (VE_UP_02)	Data values that identify a single object are not duplicated.	Events are not likely duplicates
Atemporal plausibility	Data values and distributions agree with an internal measurement or local knowledge.	
Atemporal Plausibility (VE_AP_01)	Data values and distributions for independent measurements of the same fact are in agreement.	Two or more coincident events obey a logical constraint.

Classification	Definition	Implementation
Atemporal plausibility	Logical constraints between values agree with local or common knowledge (includes "expected" missingness).	
Atemporal plausibility	Values of repeated measurement of the same fact show expected variability.	
Temporal plausibility (VE_TP_01)	Observed or derived values conform to expected temporal properties.	Admission occurs before discharge
Temporal plausibility (VE_TP_02)	Sequences of values that represent state transitions conform to expected properties.	The chronology of events is correct
Temporal plausibility (VE_TP_03)	Measures of data value density against a time-oriented denominator are expected based on internal knowledge.	Events occur within episodes
Temporal plausibility (VE_TP_04)	Measures of data value density against a time-oriented denominator are expected based on internal knowledge.	Events from a particular site demonstrate similar seasonal patterns
Temporal plausibility (VE_TP_05)	Measures of data value density against a time-oriented denominator are expected based on internal knowledge.	Events occur with anticipated patient level periodicity

Table 4.4: Kahn data quality evaluation framework: verification process. Where domains have been implemented within inspectEHR the accompanying evaluation code has been provided.

4.3.2 Best Practice Development in Restricted Environments

A major challenge to the development of inspectEHR [6] was the need to implement the data quality evaluation pipeline within the UCL DSH. There are key differences between an ideal for scientific reproducibility, which would include version control, testing of code and the use of reproducible environments [178], and what is possible in the DSH. These differences are outlined.

Version Control

Version control is a formal method to track changes to project files [181]. As software grows in complexity, version control becomes increasingly useful. In the context of scientific software development, version control has two core functions:

1. tracking of the state of the codebase so that research outputs can be associated with a specific fixed point in the development history of the software. In short, facilitating scientific reproducibility.
2. management of the interconnected dependencies that may exert downstream effects if not properly managed. Thus allowing the developer to isolate and revert any particular changes to the codebase should it be required.

In the DSH version control was not widely available until the latter parts of this project, making development challenging. As version control support was added (in the form of git and gitlab[4]) development was accelerated. Connections could subsequently be made between any particular stage of development for inspectEHR [6] and research outputs. It remains a challenge to connect the development of software between the DSH and external environments. It is not feasible to review the entire version control history of a project before exporting from the DSH, and so there is a practical limitation on exporting any version controlled software from the DSH. As a result, development either needs to proceed entirely inside the DSH, or fragments of code can be written external to the DSH and imported. Current work is ongoing to allow a more streamlined process whereby code could be developed external to the DSH and mirrored into the DSH in a more automated fashion. This

[4]git is the version control software, and gitlab is a suite of developer tools that utilise git

would permit a full version control tree both inside and outside the DSH. The main obstacles for this ideal include security concerns and a lack of comprehensive test data. There must be some reasonable guarantee that nefarious code is not going to be imported as part of this process. Additionally, since much development needs to take place in proximity to research data, this is only a realistic solution if development can take place with test data that is close enough to the original so as to have utility for development, but without any privacy concerns. This is an ongoing and as yet unsolved problem.

Testing

Testing code is a critical component of software development, ensuring that research code performs as expected. Failure to properly test research code can have disastrous consequences [182]. Each of the evaluation functions employed by inspectEHR [6] has a companion test suite written with the testthat package [183], which ensures that functions:

- handle anticipated inputs correctly.
- handle unanticipated inputs by failing safely.
- handle all possible cases where the function is sufficiently specific to allow these options to be enumerated.
- handle a range of appropriate test cases, including edge cases, where the input options are too numerous to be enumerated in entirety.

The main challenge to writing a testing suite, was the lack of good quality test data. The overcome this, I created a synthetic test cohort. As a generative cohort, this had perfect privacy preserving features (i.e. patients cannot possibly be real in this cohort). Specific tolerances could be set to produce cohorts with certain error prone characteristics. This was designed to mimic the kinds of errors that were commonly encountered in reality, and include for example:

- sites that contribute data in different units.
- missing data.
- duplicate data.

- patients that follow incoherent temporal patterns.
- contribution frequencies that were too high or too low.

Test data were abstracted into their own package "hic.data", along with common utilities and configuration files, so that it could be shared as a dependency for both inspectEHR [6] and wranglEHR [5].

Reproducible Environments

The UCL DSH has evolved in maturity over the course of this project. However, there remains one key element that is missing as a core element for producing high quality reproducible research; reproducible environments. Any research output will be a function of both data and the research environment. This includes the interaction between the operating system, software and hardware. Small perturbations in how systems are constructed can have surprisingly dramatic effects on research outputs [184]. Further, if we wish to repeat a previous study, even when presented with perfect copies of archived data and all software in the exact state when it was originally conducted, the interaction with a different research environment may cause the research pipeline to fail or produce a different result. This problem is largely solved by the implementation of project containers and virtual machines to create an encapsulation of the entire operating environment. Owing the security concerns, a solution to this problem is not currently implemented, though it remains of high priority to address.

4.3.3 Summary Data Quality Metrics

The quality evaluation process outlined is necessarily verbose, and so it has proven useful to have a set of summary metrics that help communicate an overall impression of data quality. Two scoring systems are presented, each providing an appraisal of data quality at the foundational elements of the CC-HIC data model; episodes and events (i.e. all individual data elements that are contained within each episode). The main goal of these metrics is to provide a simple summary measure that is largely independent of the size of the cohort, or the therapies provided.

The episode score is defined in equation 4.2. Simply, it is the proportion of

episodes that pass validation, with a penalty applied for large portions of the data contribution where episodes are highly likely to be missing. Although necessarily an estimate, this penalty factor is important to include as missing episodes are a common problem in the CC-HIC database. This penalty prevents an overly optimistic score that would be achieved by sites electing to not submit erroneous data.

$$\frac{\sum_{n=1}^{N_{present}} I(episode_n)}{N_{present} + N_{missing}} \tag{4.2}$$

Where:

- $I(episode_n)$ is the indicator function for whether or not the nth episode proceeds through the inspectEHR [6] evaluation pipeline without raising an error.
- $N_{present}$ is the total number of episodes submitted to CC-HIC.
- $N_{missing}$ is the total number of episodes thought to be missing from the submission. This is based on identifying calendar months within which the number of daily admissions falls two standard deviations below the seasonal daily average for that site for $\geq$ 10 days. This is a somewhat arbitrary cut off, but does serve to detect deficient episode contribution with little ambiguity, since such a pattern would be extremely unusual. This penalty therefore is likely to return a conservative estimate of the number of missing cases.

The event score is defined in equation 4.3. The event score is normalised against calendar months of data submission for each site. This is because at the patient or episode level, it should be expected that many data concepts will be missing; patients do not receive all treatments and investigations in each stay. At the scale of the month however, it would be reasonable to expect that at least one instance of each data concept in the CC-HIC data model should be observed.

$$\frac{\sum_{m=1}^{M} I(concept_m) \frac{\sum_{t=1}^{T} I(month_t)(event_p / event_f + event_p)}{T}}{M} \tag{4.3}$$

Where:

- $I(concept_m)$ is the indicator function for the mth concept being contributed to

CC-HIC.

- M is the total number of primary data concepts in the CC-HIC data model.
- $I(month_t)$ is the indicator function for the tth month of data submission containing any instances of the concept.
- T is the total number of contiguous months of submission, from the date of first submission, to the last observed submission in the cohort.
- $event_p$ is the number of events that do not yield any error codes for concept m.
- $event_f$ is the number of events that do yield error codes for concept m.

The primary advantage of these scores is that they encapsulate all aspects of the data QE process and present the results in an easy to understand format bound on $[0, 1]$ with 1 representing a perfect score.

4.4 wranglEHR

A common statistical work flow (and the one adopted in this thesis) is to represent data in the so-called "tidy" format [185], that is, a rectangular format with the following specification:

- one row per statistical unit.
- one column per unique variable.

In this case, the statistical unit is either an ICU episode (where time invariant data are concerned), or a period of time for each patient (often 30 minutes or an hour)[5]. This step of extracting data from the CC-HIC research database and transforming it into a rectangular format is so commonly required, that it was desirable to write a software package to standardise this process. The cleanEHR package [14] originally served this purpose, whereby a user could specify the extraction requirements via a configuration YAML Ain't Markup Language (YAML) file. wranglEHR [5] could be thought of as the spiritual successor to cleanEHR, though

[5]The underlying EHR rarely stores data more often than at 5 minute intervals, and so this is a reasonable expected upper boundary. At higher temporal resolutions, the methods described here are likely to fail

wranglEHR [5] was written from the ground up as a software contribution to this thesis. As with inspectEHR [6], wranglEHR [5] is written in the tidy style, follows best practices from the R Studio development team where possible, and implements the modern paradigm of tidy evaluation [179, 180]. wranglEHR exposes two main functions to the end user:

1. `extract_demographics()`
2. `extract_timevarying()`

Both allow flexible data extraction from the CC-HIC research database, according to a user specification which is supplied directly as arguments to the above exported functions. Extracted data are transformed into a rectuangular shape for analysis, with accompanying meta-data identified and arranged into appropriately labelled columns. The extraction process can be customised to suit a specific case use, including:

- setting the desired temporal cadence of the table (i.e. one row per hour, versus one row per day).
- defining a custom or user specified action if the data storage resolution is higher than the target row cadence.

The advantage of this approach, is that the data extraction process can be standardised for a large number of research questions. Accompanying unit tests have been written to ensure that the data extraction process takes place as expected, and is consistent across a broad range of requirements. wranglEHR [5] abstracts away a large amount of code that is required to extract data from the CC-HIC database. Further, the analyst does not need to take on the cognitive load in remembering exactly how data is represented within the database, or switch between SQL and the analysis language of their choice. An overview of how the end user is expected to interact with wranglEHR [5] is provided in the software vignette in appendix Section C.2 (page 275).

4.5 Results

Results are presented in accordance with the format of the Kahn data QE framework. Specific examples are provided to help illustrate important or recurring themes, or where the area of concern is likely to impact upon the clinical research studies that follow. Table 4.5 illustrates how the results are to be presented. Results are provided with the total number times that an error is observed, and the proportion of total errors that this represents. The number of concepts affected by the error are shown as a means to quantify the breadth of the error in the CC-HIC database.

eval code	description	count	proportion	concepts
PP-DD-##	Description of the inspectEHR evaluation code	1,000	1.0×10^{-3}	25

Table 4.5: Example of data quality error results. An inspectEHR evaluation code with its description are provided. **Count:** the number of occurrence of this error label in the CC-HIC research database. **Proportion:** the proportion of total errors that this code represents. **Concepts:** the number of concepts that this error code is observed against (to a maximum of 255.)

4.5.1 Value Conformance

Data values should conform to constraints defined internally in the CC-HIC data model, or by appropriate external bodies. Data should be present in the correct system of measure (i.e. provided in the correct units), fall within an appropriate numerical range or set, or follow a pre-specified pattern (e.g. post codes should conform to UK standards). A summary of value conformance errors for the CC-HIC research database are shown in table 4.6.

Commonly encountered value conformance errors include numerical data that were contributed in incorrect units, or data that did not conform to the correct pre-

eval code	description	count	proportion	concepts
VE-VC-01	Value does not conform to external standard	12,152	1.35×10^{-4}	6
VE-VC-03	Numeric data falls outside range of possibility	6,457,616	7.16×10^{-2}	87

Table 4.6: Summary of data quality errors found for the value conformance domain

specified pattern.

In many cases it is easy to find a solution to data that have been contributed in the wrong units. F_IO_2 is one such example where it is contributed variably as either a percentage (ranging from 21-100%) or a fraction (ranging from 0.21-1). The distributions of these values exist on a non-overlapping support, are site specific, and the underlying meaning behind the values is straightforward to imply. As such, one can be reassured about applying a *post hoc* transformation to align all values onto the correct scale. There are situations where such a conversion is not possible. Noradrenaline—a drug used to constrict the blood vessels and increase blood pressure—is best represented as micrograms per kilogram per minute (mcg/kg/min). This allows one to understand the mass (and hence dose) of drug administered. Two sites contributed this concept in what can be assumed to be millilitres per hour (mL/hr) of an unknown concentration. This representation of noradrenaline cannot be reconciled into interpretable units as the dose administered is unknown without knowing the concentration of noradrenaline used. The contributed units of mL/hr hold no clinical meaning—implying a range of possible doses—and so beyond knowing that the patient received noradrenaline, little further information can be gleaned.

4.5.2 Relational Conformance

Data values should conform to relational constraints, including those that are structurally imposed by the data model and database. A summary of relational conformance errors for the CC-HIC research database are shown in table 4.7. The single error type identified in this domain was the presence of events that originate from episodes that did not pass the minimum data quality standard to be considered safe for research use. Since the episodes themselves were not trustworthy, the data they contain are also likely to be questionable.

4.5.3 Computational Conformance

Data values that are derived or calculated should be replicable internally and compare positively to external standards. Computational conformance is not presently

eval code	description	count	proportion	concepts
VE-RC-04	Event originates in episode failing quality evaluation	$4,305,829$	4.77×10^{-2}	246

Table 4.7: Summary of data quality errors found for the relational conformance domain

implemented in inspectEHR [6]. Calculated fields that would benefit from evaluation would include the APACHE-II score and the daily total of basic and advanced life support. These are not currently implemented because the source data to recalculate and check these fields internally have a high degree of missingness. This would result in a poor performance in this domain, where the evaluation would, in effect, be re-evaluating for data missingness (which is already addressed by the "completeness plausibility" domain) rather than evaluating for computational conformance.

4.5.4 Completeness Plausibility

Data should be complete with respect to project specifications, local EHR data availability and case mix. This completeness should be, within acceptable tolerances, uniform over time. A summary of completeness plausibility errors for the CC-HIC research database are shown in table 4.8.

Missing data patterns across the CC-HIC research database are shown in figures 4.5-4.7. Not all data concepts are provided by all sites. This is not necessarily a point of concern as sites do not uniformly collect and store data on all concepts in the CC-HIC data model. Concepts with a high degree of non-contribution included systematic ICU components, cardiac output monitoring, and certain drugs that are less commonly administered.

eval code	description	count	proportion	concepts
VE-CP-02	Non temporal missing data pattern in data item	$8,956$	9.93×10^{-5}	104
VE-CP-04	Temporal missing data pattern in data item	$16,043$	1.78×10^{-4}	222
VE-CP-05	Metadata is absent	$16,658,827$	1.85×10^{-1}	21

Table 4.8: Summary of data quality errors found for the completeness plausibility domain

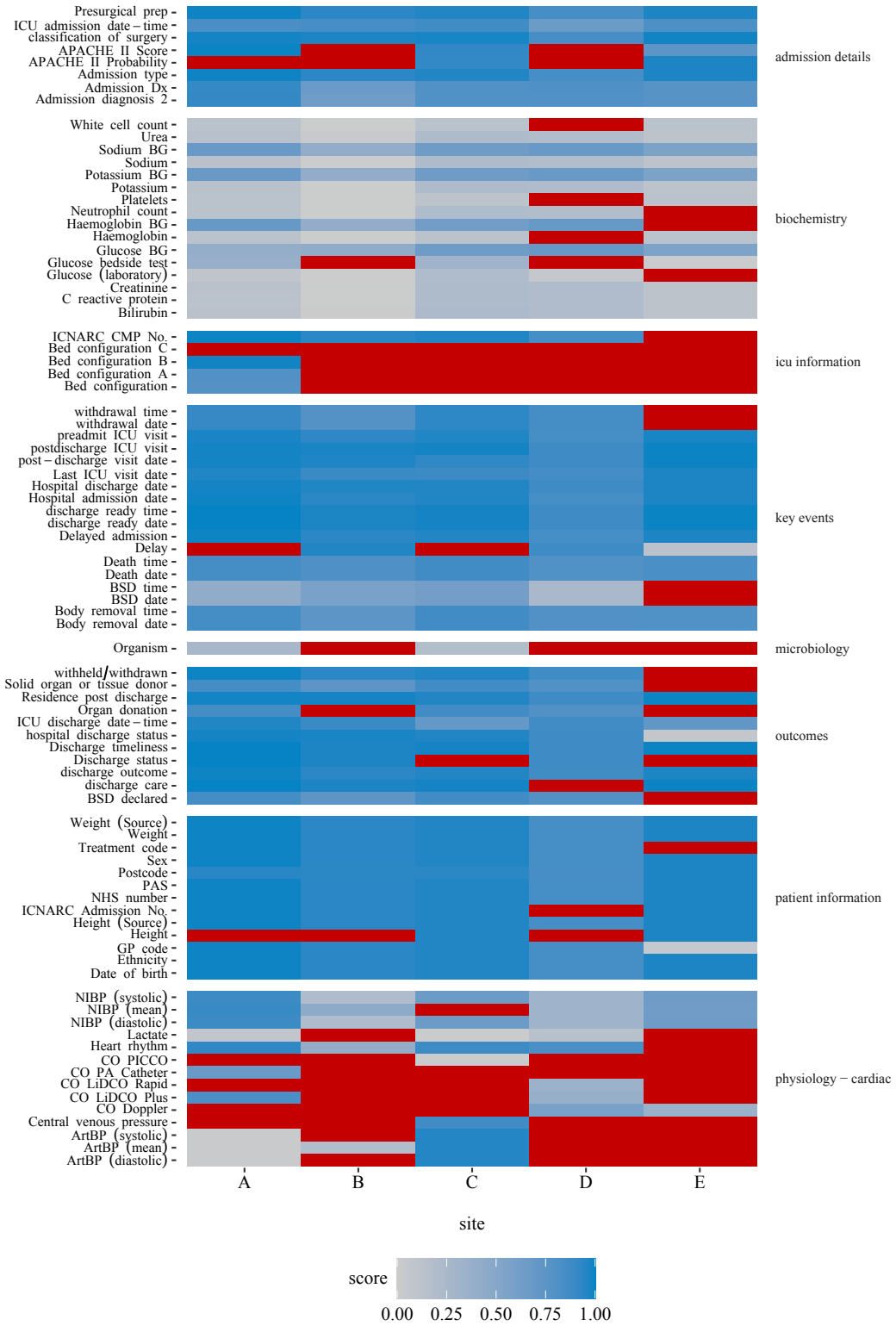


Figure 4.5: Data quality and data missingness patterns for the entire CC-HIC research database. Data concepts that are missing are highlighted in red. The data quality event score is shown in grey-blue (deeper blue hue indicating a higher performance).

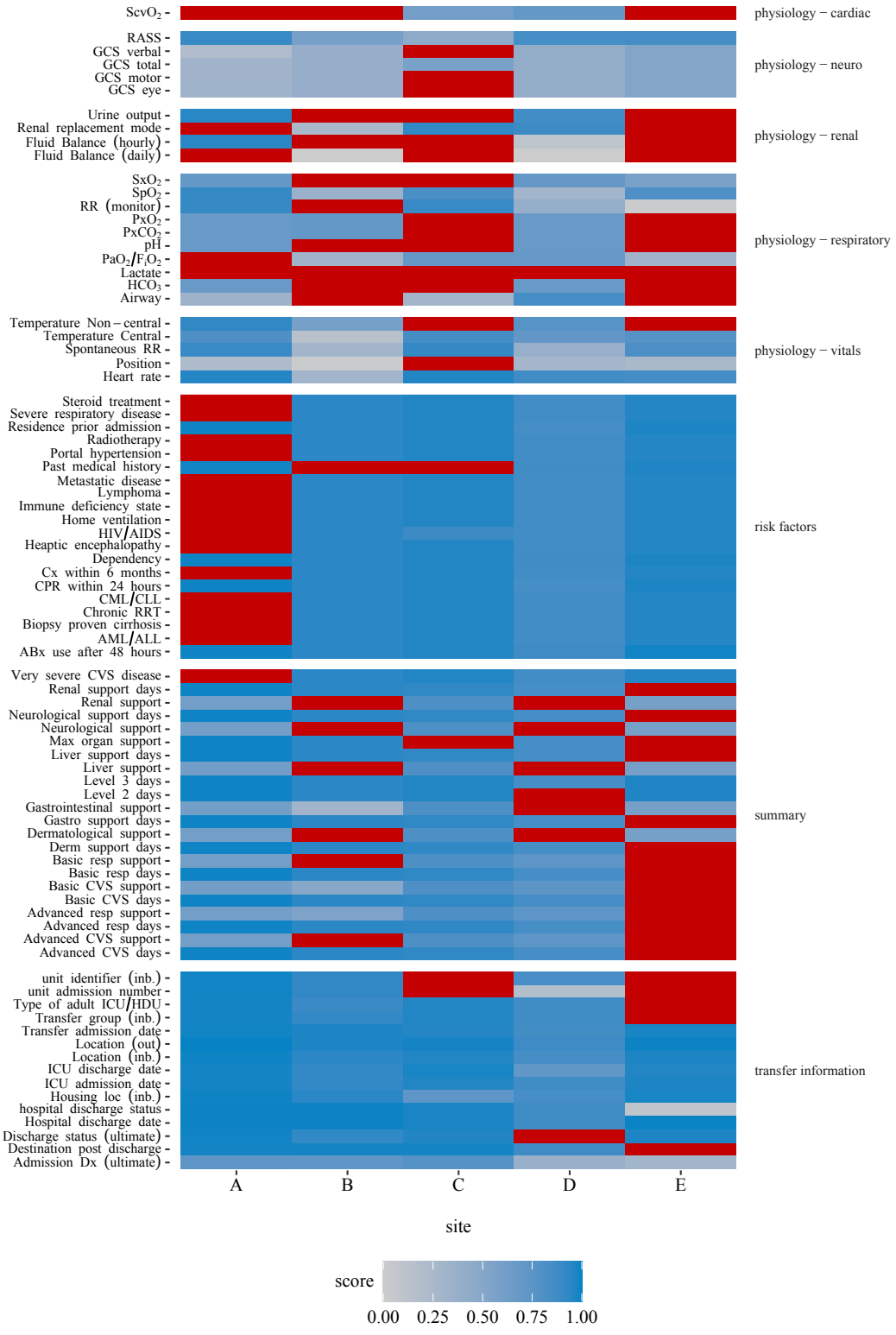


Figure 4.6: Data quality and data missingness patterns for the entire CC-HIC research database. Data concepts that are missing are highlighted in red. The data quality event score is shown in grey-blue (deeper blue hue indicating a higher performance).

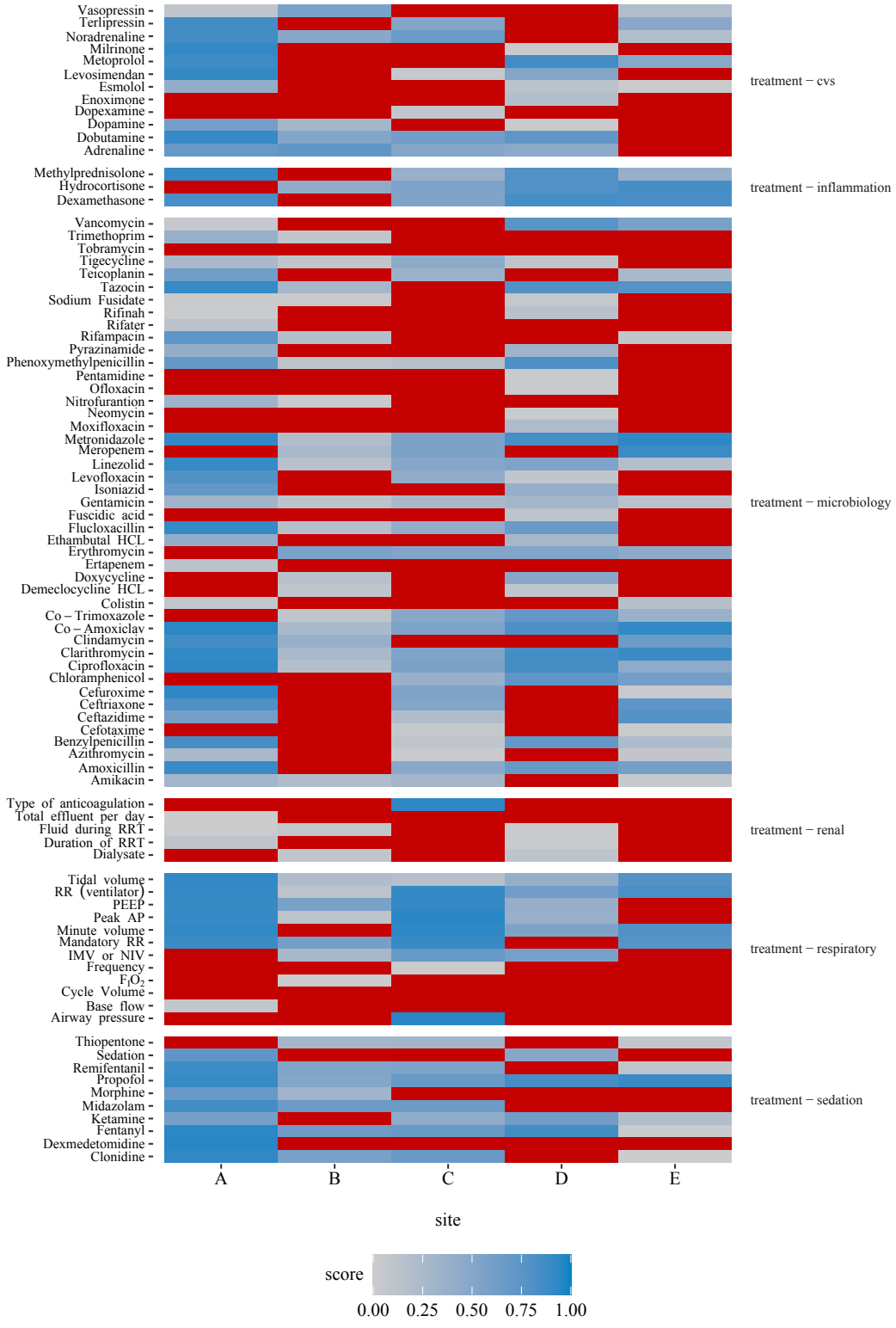


Figure 4.7: Data quality and data missingness patterns for the entire CC-HIC research database. Data concepts that are missing are highlighted in red. The data quality event score is shown in grey-blue (deeper blue hue indicating a higher performance).

eval code	description	count	proportion	concepts
VE-UP-02	Event is likely a duplicate	569,482	6.31×10^{-3}	97

Table 4.9: Summary of data quality errors found for the uniqueness plausibility domain

Systematic ICU components are likely missing as there is no logical means through which an episode centric data model can represent these concepts. Other missing concepts are site specific and may in places reflect data availability in the source EHR.

Meta-data showed a disproportionately high degree of missingness, accounting for 18% of all error codes generated. Meta-data in the CC-HIC data model often encodes information that is vital to the correct interpretation of primary data. For example—and with particular relevance to the clinical research studies that follow—this includes the anatomical source of a blood gas sample, without which the sample oxygen characteristics are uninterpretable.

Some important temporal patterns of missingness were observed. This included two sites who stopped contributing data for a range of drug infusions yet were still contributing the episodes. In both cases, the underlying cause of temporal missing data were back-end changes to the source EHR, causing certain concepts to no longer be captured properly by local research data pipelines.

4.5.5 Uniqueness Plausibility

Data that describes a singular concept should not be duplicated. It is only possible for data duplication to occur in CC-HIC among time varying events since, fortunately, there is referential integrity in the CC-HIC data model to prevent such duplications occurring in time invariant events. A summary of uniqueness plausibility errors for the CC-HIC research database are shown in table 4.9.

Data duplications in the CC-HIC research database are rare. In one investigation of data duplication, the duplicate findings were in fact genuine and existed in the source EHR. This was attributed to the existence of multiple systems to view clinical data, each of which had persisted data into the EHR. This is an interesting example of a “source” type error and it is open to debate as to whether it would be

eval code	description	count	proportion	concepts
VA-AP-01	Values do not share a common distributions across sites	10,652,827	1.18×10^{-1}	100
VE-AP-01	Two or more events not not obey a logical constraint	33,102	3.67×10^{-4}	3

Table 4.10: Summary of data quality errors found for the atemporal plausibility domain

beneficial to retain, but label, these particular data values.

4.5.6 Atemporal Plausibility

Data values and distributions should broadly conform to expected patterns. A summary of atemporal plausibility errors for the CC-HIC research database are shown in table 4.10. A large proportion of submitted data, across many concepts (11% of all errors) were found in the atemporal plausibility domain. This is despite setting a relatively high tolerance for the applied KS test of 0.5, which permits distributions to diverge from one another by a considerable amount. The large number of errors generated is in part related to how errors are assigned for *all* values for the data concept, since this is a distributional test and therefore applies to every value.

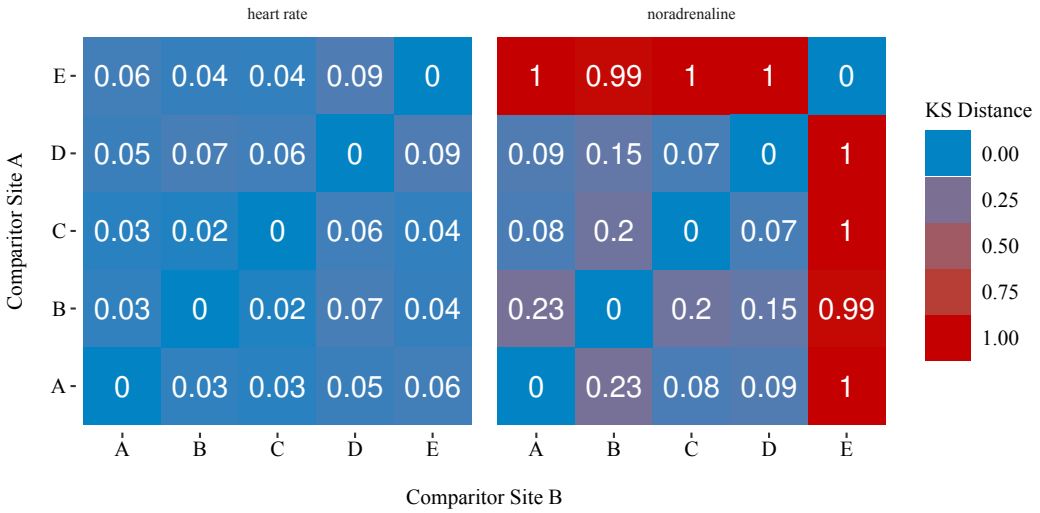


Figure 4.8: Atemporal plausibility KS distributional evaluation of numeric concepts. **Left panel:** Heart rate. **Right panel:** Noradrenaline. Note that the distributional error is clearly visible for noradrenaline, while the heart rate distributions compare favourably between sites.

Figure 4.8 illustrates this distributional check process for two data items, heart rate and noradrenaline. It is easy to discern from the test, that heart rate performs favourably, while noradrenaline does not.

4.5.7 Temporal Plausibility

Data value density should meet expectations when evaluated against a time-oriented denominator. Time dependent data that can be ordered, should appear in the correct order. A summary of temporal plausibility errors for the CC-HIC research database are shown in table 4.11.

Key findings of this domain include a high number of events that are reported from times outside of an ICU episode (20% of all errors) or contributed at a higher or lower frequency than would be expected (34% of all errors). Information contributed from outside an ICU episode is not a point of concern, other than it highlights that the data model has not been applied correctly. The frequency of data contribution is of concern, and is well illustrated by the urine output concept. The frequency with which a concept is contributed can be examined by calculating the time between samples for the same patient. The frequency of urine output (shown in figure 4.9) is extremely variable across the cohort. Some sites *only* contribute hourly data, which makes it extremely unlikely that non-catheterised[6] patients have been included in the cohort. On the contrary, one site *rarely* contributes hourly data, making it unlikely that catheterised patients form a large component of these data. In this example, zeros are variably contributed.

[6] a catheter is a small tube placed into the bladder. In an ICU, urine output is typically recorded hourly when such a device is present. Without a catheter, urine can only be measured according to the patient's own schedule.

eval code	description	count	proportion	concepts
VA-TP-01	Values do not share a common temporal distribution across sites	$1,852,155$	2.05×10^{-2}	31
VE-TP-02	Chronology of key events is correct	$11,778$	1.31×10^{-4}	3
VE-TP-03	Events occur outside the timespan of an episode	$18,652,675$	2.07×10^{-1}	145
VE-TP-05	Events occur outside anticipated patient level periodicity	$30,962,048$	3.43×10^{-1}	151

Table 4.11: Summary of data quality errors found for the temporal plausibility domain

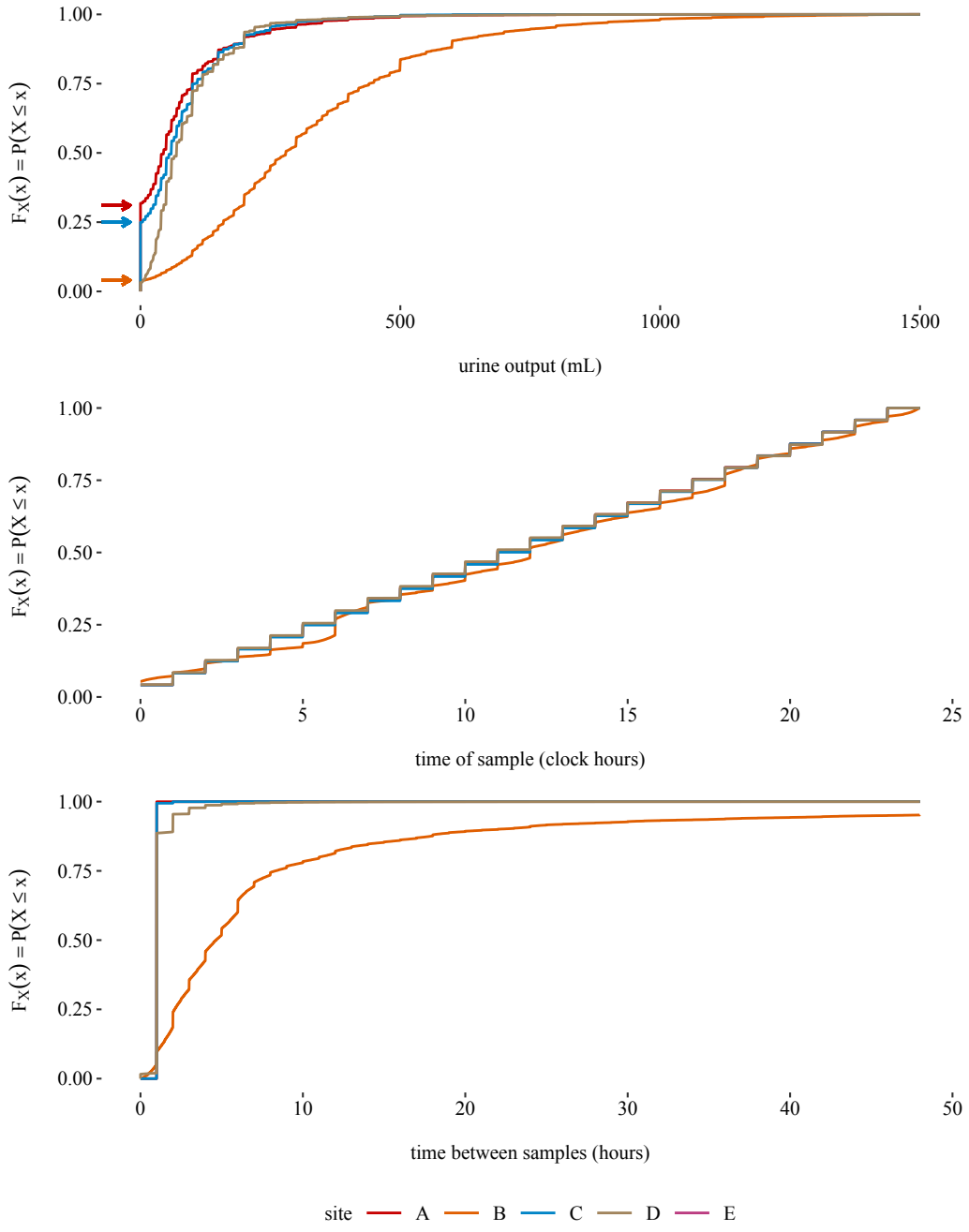


Figure 4.9: Empirical cumulative distribution functions are displayed. **Top panel:** value of urine output. Arrows are imposed to highlight the significant difference between sites recording zeros. **Middle panel:** time of day that urine output concept was contributed. **Bottom panel:** time (in hours) between samples.

This is particularly suspicious as an indicator for missing data, since it would be unusual for a UK ICU to never see any patients with anuria. Confusingly, the distribution of times for urine output events are strikingly uniform between the cohorts. Therefore, it is difficult to fully reconcile the data for this concept. One can speculate that the source EHRs store urine output for catheterised and non-catheterised patients separately and, in some instances, perhaps both have not been entered into the CC-HIC cohort. Regardless, the data for this concept are suspicious for not being representative of reality.

Figure 4.10 highlights errors where the temporal order of data concepts is logically incorrect. Timings relating to discharge from ICU, discharge from hospital and death were particularly problematic. In many instances, this could be attributed

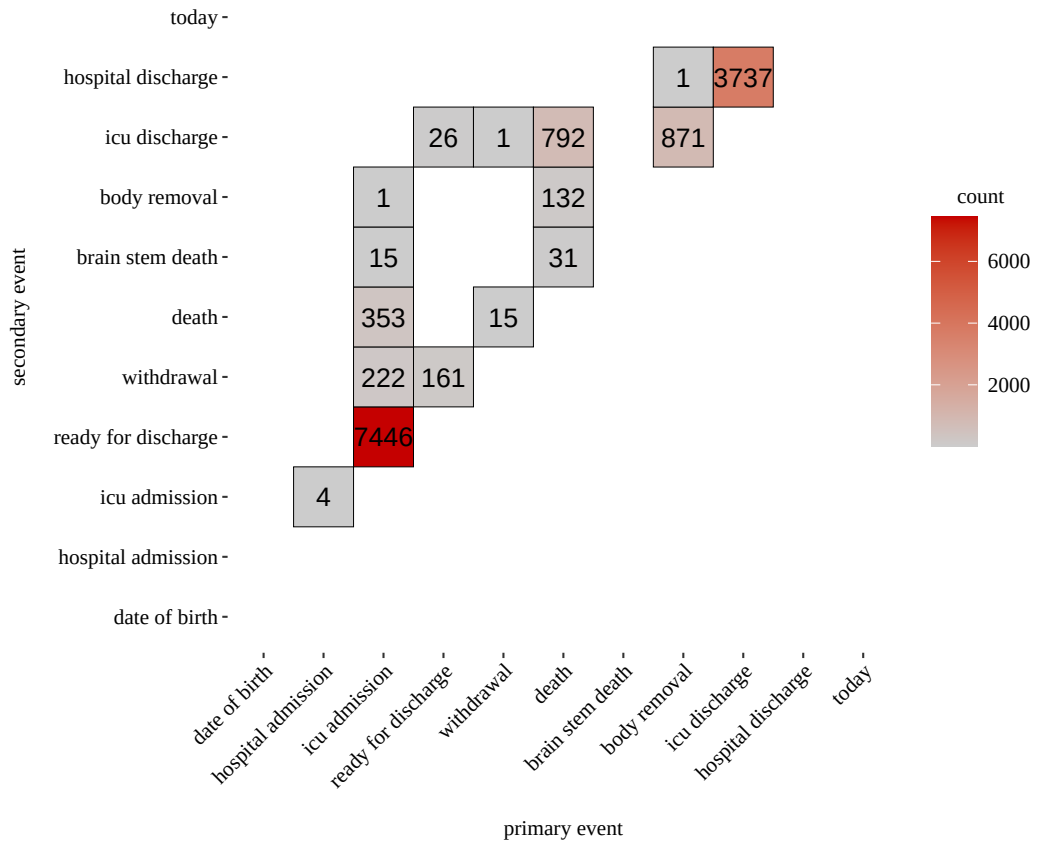


Figure 4.10: Chronology of key events. Cells highlight occurrences where key events that can be ordered in time, have appeared out of sequence. Commonly seen errors include patients: being ready for discharge prior to their arrival in the ICU, and discharged from hospital before discharged from the ICU.

to information that occurred from outside the ICU episode (for example hospital discharge or death) being erroneously connected to an ICU episode in the past or future.

4.5.8 Episode Characterisation

An area of data quality not evaluated directly by the Kahn framework is episode characterisation. This is a composite evaluation that requires many of the domains described previously to coordinate an answer. To be characterised, episodes must have an unambiguous:

1. start: when did the episode begin?
2. end: when did the episode finish?
3. outcome: did the patient survive or not?
4. patient: to whom does the episode belong?

Further, episodes must originate from areas of the database where data integrity are not of immediate concern. There are analytic scenarios where these missing properties could be approached through an appropriate set of methods (e.g. multiple imputation by chained equations). Here these properties are being employed as surrogates of overall data quality. It would be highly unusual for the four properties listed to not be identifiable from the EHR, since they are necessary for healthcare delivery. Where episodes are unable to provide unambiguous answers to these four properties, it suggests that the quality of these data has degraded to an unacceptable level.

The start of an episode is unambiguous in the CC-HIC data model, forming part of the primary key for the episode, and so mandatory for submission. The end of an episode is less clearly defined and must be reconciled by correlating the patient outcome with one of: the date-time of discharge, date-time of death, date-time of body removal or the last observed physiological data. The episode outcome is mostly described by a single concept (status at discharge) with support from a secondary concept for the special case where patients are declared brain stem dead[7].

[7] this is a relatively uncommon event in the CC-HIC database, but does create a situation whereby

Evaluation code	Description	n
VA-CP-01	No ICU outcome status	348
VA-VC-01	Invalid nhs number	38
VE-CP-01	Episode end cannot be reconciled	373
VE-CP-03	Episode originates in bad sector	146
VE-TP-01	Episode length ≤ 0	572
VE-UP-01	Duplicate and conflicting death times	199
VE-VC-04	Overlapping episodes	150

Table 4.12: Results of episode characterisation. Evaluation codes are as implemented by inspectEHR and defined in tables 4.3 and 4.4. A "bad sector" is a period of time where contribution from a particular site is in question.

Since a person is an attribute of the episode in the CC-HIC data model, then an incorrect identification of an individual would lead to the undesirable association of episodes between different patients. This is most commonly seen when the NHS number is absent, fails to conform to known standards or is contributed as a place holder (most commonly: "000000000")[8]. Exclusion on these grounds is particularly worrisome as invalid or missing NHS numbers have been shown to be more common in women, ethnic minorities, non-UK born individuals and those with social risk factors [186]. There is therefore a risk that the CC-HIC may introduce social bias into analysis by relying on the NHS number alone to identify patients as unique.

The results of episode characterisation are shown in table 4.12. Episode characterisation reduces the available size of the maximum cohort from 47,932 to 46,658 episodes (97.3%).

Table 4.13 shows the patient characteristics for this primary cohort, stratified by whether or not they were excluded by failing to meet the minimum specification for episode characterisation. Episodes that are excluded tended to:

- be for patients with lower weight (though likely not clinically relevant).
- be an emergency admission.
- be a medical admission.

a patient's body can be retained inside the ICU for a relatively long period of time, while still producing physiological data.

[8]While invalid, the "000000000" placeholder technically passes the NHS number checksum specification, which is likely the reason why it is used in trusts that require patients to have an NHS number, even when one has not been issued to a patient.

- have higher APACHE II scores (i.e. are sicker).
- have much higher rates of cardiopulmonary resuscitation (CPR) prior to admission (15% vs. 4%).
- have much higher rates of death (61% vs. 8%).

4.5.9 Spell Reconcilliation

A minor modification to the way in which episodes are represented is necessary before research can proceed. This is because one site in particular starts a new episode for each patient as they transition between physical ICU sites in the hospital. This is quite a common occurrence, for example, when a patient transitions from a level 3 ICU into a level 2 HDU during a phase of recovery from critical illness. In these instances, we are interested in a contiguous period of critical illness, rather than when a patient transitions between physical locations, which often occur on both clinical and administrative grounds (like bed availability). As a result, new episodes for the same patient that start within six hours of a previous episode, are linked as the same "spell".

Characteristic	Overall[1]	Excluded[1]	Included[1]	p-value[2]
N	47,932	1,438	46,494	
Height	1.69 (0.10)	1.69 (0.10)	1.69 (0.10)	0.2
Missing	50	1	49	
Weight	75 (65, 87)	73 (62, 85)	75 (65, 87)	0.014
Missing	0	0	0	
Sex				0.4
Female	19,663 (41%)	605 (3.1%)	19,058 (96.9%)	
Male	28,268 (59%)	833 (2.9%)	27,435 (97.1%)	
Missing	1	0	1	
Ethnicity				
White British	29,503 (62%)	893 (3%)	28,610 (97%)	
White Irish	821 (1.7%)	21 (2.6%)	800 (97.4%)	
White other	3,280 (6.9%)	87 (2.7%)	3,193 (97.3%)	
Mixed white/black	160 (0.3%)	3 (1.9%)	157 (98.1%)	
Mixed white/Asian	51 (0.1%)	1 (2%)	50 (98%)	
Mixed any other	145 (0.3%)	2 (1.4%)	143 (98.6%)	
Asian/Asian British	1935 (4.0%)	58 (3%)	1,877 (97%)	
Black/Brit. Carribean	1,197 (2.5%)	24 (2%)	1,173 (98%)	
Black/British African	1,505 (3.1%)	31 (2.1%)	1,474 (97.9%)	
Black/British other	1,006 (2.1%)	9 (0.9%)	997 (99.1%)	
Chinese	261 (0.5%)	7 (2.7%)	254 (97.3%)	
Other ethnic group	1,849 (3.9%)	43 (2.3%)	1,806 (97.7%)	
Not stated	6,124 (13%)	213 (3.5%)	5,911 (96.5%)	
Missing	95	46	49	
Admission priority				<0.001
Elective	17,206 (36%)	363 (2.1%)	16,843 (97.9%)	
Emergency	30,726 (64%)	1,075 (3.5%)	29,651 (96.5%)	
Missing	0	0	0	
Admission Type				<0.001
Medical	25,748 (54%)	871 (3.4%)	24,877 (96.6%)	
Surgical	21,793 (46%)	262 (1.2%)	21,531 (98.8%)	
Missing	391	305	86	
Surgical classification				<0.001
NA (medical)	25,748 (56%)	871 (3.4%)	24,877 (96.6%)	
Elective	12,396 (27%)	134 (1.1%)	12,262 (98.9%)	
Scheduled	2,520 (5.5%)	9 (0.4%)	2,511 (99.6%)	
Urgent	2,490 (5.4%)	41 (1.6%)	2,449 (98.4%)	
Emergency	2,796 (6.1%)	61 (2.2%)	2,735 (97.8%)	
Missing	1,982	322	1,660	
Organ system				
Cardiovascular	12,477 (26%)	313 (2.5%)	12,164 (97.5%)	
Dermatological	356 (0.7%)	7 (2%)	349 (98%)	
Endocrine	2,113 (4.4%)	37 (1.8%)	2,076 (98.2%)	
Gastrointestinal	8,112 (17%)	180 (2.2%)	7,932 (97.8%)	
Genito-urinary	5,308 (11%)	103 (1.9%)	5,205 (98.1%)	
Haematological	755 (1.6%)	28 (3.7%)	727 (96.3%)	
Musculoskeletal	1,539 (3.2%)	24 (1.6%)	1,515 (98.4%)	
Neurological	3,490 (7.3%)	115 (3.3%)	3,375 (96.7%)	
Poisoning	812 (1.7%)	6 (0.7%)	806 (99.3%)	
Psychiatric	29 (<0.1%)	0 (0%)	29 (100%)	

Characteristic	Overall[1]	Excluded[1]	Included[1]	p-value[2]
Respiratory	10,456 (22%)	263 (2.5%)	10,193 (97.5%)	
Trauma	2,094 (4.4%)	57 (2.7%)	2,037 (97.3%)	
Missing	391	305	86	
Level 2 days (HDU)	2 (1, 3)	1 (0, 2)	2 (1, 3)	<0.001
Missing	33	1	32	
Level 3 days (ICU)	0 (0, 2)	1 (0, 2)	0 (0, 2)	0.3
Missing	33	1	32	
Supported organs (max)	2 (1, 2)	2 (1, 3)	2 (1, 2)	0.017
Missing	5,117	182	4,935	
Apache II Score	14 (11, 18)	20 (15, 27)	14 (11, 18)	<0.001
Missing	10,685	951	9,734	
CPR prior to admission	2,051 (4.3%)	210 (10.2%)	1,841 (89.8%)	<0.001
Missing	7	6	1	
ICU outcome				<0.001
Survivor	43,179 (91%)	422 (1%)	42,757 (99%)	
Non-survivor	4,405 (9.3%)	668 (15.2%)	3,737 (84.8%)	
Episode open	9 (<0.1%)	9 (100%)	0 (0%)	
Missing	339	339	0	
Hospital outcome				<0.001
Survivor	33,227 (92%)	233 (0.7%)	32,994 (99.3%)	
Non-survivor	2,760 (7.6%)	143 (5.2%)	2,617 (94.8%)	
Episode open	228 (0.6%)	25 (11%)	203 (89%)	
Missing	11,717	1,037	10,680	
Site				<0.001
A	25,757 (54%)	296 (1.1%)	25,461 (98.9%)	
B	5,373 (11%)	345 (6.4%)	5,028 (93.6%)	
C	3,935 (8.2%)	120 (3%)	3,815 (97%)	
D	8,910 (19%)	542 (6.1%)	8,368 (93.9%)	
E	3,957 (8.3%)	135 (3.4%)	3,822 (96.6%)	
Missing	0	0	0	

Table 4.13: [1]Statistics presented: Mean (SD); Median (IQR); n (%). [2]Statistical tests performed: Wilcoxon rank-sum test; chi-square test of independence; Fisher's exact test. Percentages provided in the **Overall** column are column-wise. Percentages provided in the **Excluded** and **Included** columns are calculated row-wise.

4.5.10 Summary Data Quality Metrics

The result of the data quality metrics are shown in table 4.14. The episode scores are promising, with four of five sites scoring above 0.9. Many of the points of failure are caused by logical inconsistencies in these data. For example, episodes that finish before they start or those that have conflicting outcomes (i.e. episodes are tagged with outcomes for "survivor" and "non-survivor" simultaneously.) Since many of these conflicts are managed internally to EHRs as a matter of routine data

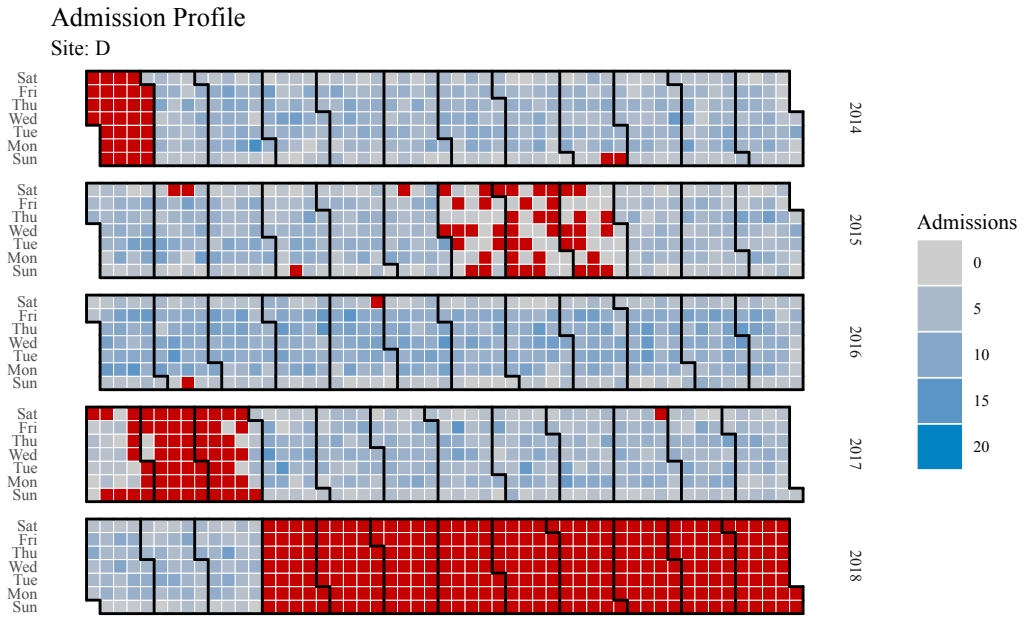


Figure 4.11: Overview of admissions from a single site. Squares in red indicate days in which no episodes have been submitted. There is a clear weekend effect visible, with lower number of admissions on Saturday and Sunday. Two blocks lasting several months exist where data submission dropped to a very low level. Data that was submitted during these time periods is also likely to be in question.

reconciliation for patient care, it seems probable that many of these inconsistencies originate as transcription errors when writing EHR data into the CC-HIC data model. Therefore, it is possible that enforcing these relational requirements in the CC-HIC data model, would result in an increase in the quality of submitted data. Site E performs well against the episode score, though there are likely many missing episodes that have not been detected by current procedures. The signal for potential missing episodes is for there to be a high proportion of missing episodes concentrated in a single month. This can be seen in the site admission profile for site D in figure 4.11. In comparison to the site admission profile for site E (figure 4.12), one will observe the generally low number of admissions. This is an extraordinarily low number of admissions to an NHS ICU, and certainly does not match expectations for this site.

The event scores are also shown graphically figures 4.5-4.7 (pages 131-133). The event score is likely a pessimistic view of the CC-HIC research database for two

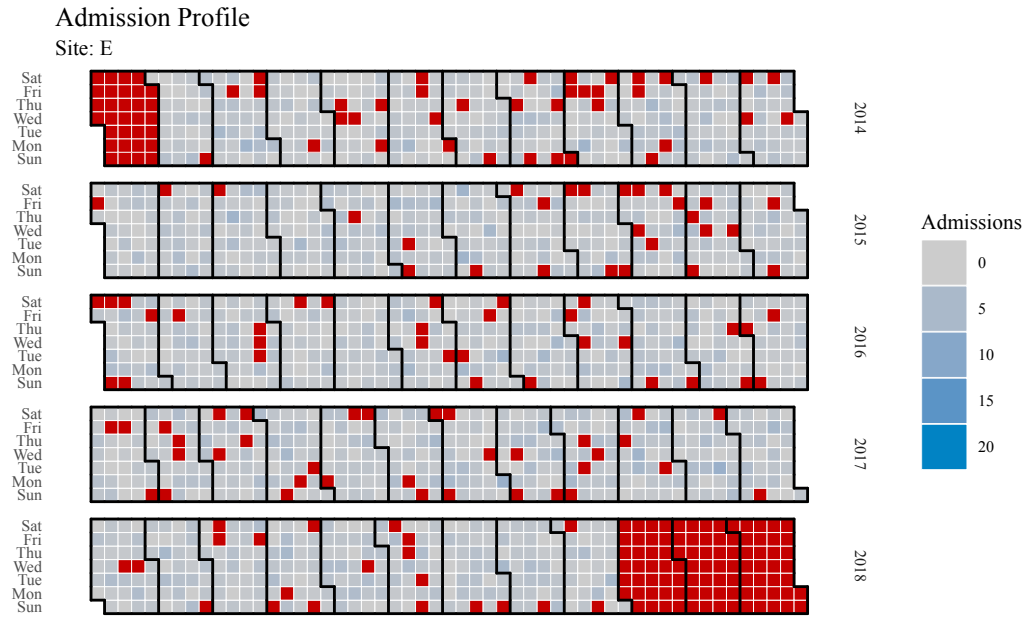


Figure 4.12: Overview of admissions from a single site. Case submission for this site is uniformly very low, and without a discernable weekend effect. There are numerous days where zero admissions occurred, but these are evenly distributed throughout the cohort. The low number of submissions is not commensurate to the size of the NHS trust from which these data originate, calling into question the likelihood of missing episodes.

reasons. First, the CC-HIC data model does not support the ability to indicate which data concepts cannot be contributed because they do not exist in the source EHR. This would negatively impact on sites unable to contribute a data concept that they do not have, which is not desirable. Second, the result is a consequence of being atypically comprehensive. inspectEHR [6] implements an extremely broad range

Site	Episode score	Event score
A	0.98	0.58
B	0.93	0.43
C	0.95	0.54
D	0.84	0.49
E	*0.96	0.39

Table 4.14: Data quality metric scores: episode and event scores are shown across all sites. The current scoring system penalises for events that are not contributed, and so the event scores may result in a lower score for sites that would be reasonable if those concepts are not persisted within the source EHR. *: site E scores well on the episode score, but their over all number of submitted cases are far lower than would be expected.

of data QEs. As each additional layer of evaluation is added, it can only possibly reduce the score for each site. This is off set by the hope that after demonstrating the issues with data quality, sites can be afforded to opportunity to correct errors and re-issue improved cohorts.

4.6 Discussion

This was the first, and most thorough, evaluation of the quality of data stored in the CC-HIC research database. This evaluation by necessity draws out deficiencies in these data, though there are many areas that evaluate favourably. Four of the five contributing sites were able to contribute a vast and comprehensive dataset. This contribution was maintained on a regular schedule throughout most of the project lifetime. Some major deficiencies that were reported back to contributing sites were either corrected at source, or external data were issued providing a correction within the CC-HIC research database itself. This demonstrates the successful movement of a comprehensive multicentre critical care dataset into a secure research environment and therefore indicates the successful attainment of the original aims of the HIC outlined in Section 2.3 (page 42).

Throughout the study period, modifications to data submissions were necessary to correct some of the data quality issues highlighted. Examples of some major improvements to data quality included:

- restoration of drug infusions that had ceased to flow with the primary data.
- contribution of a large portion of missing respiratory data via an external data patch.
- re-supply of drug infusions in the correct units via an external data patch.

These modifications were typically submitted as bespoke patches of data (in the form of Comma Separated Value (CSV) files) that could be added to the CC-HIC research database manually. While this may be a pragmatic solution, adding files in this way outside the normal pathway, introduces an unstable element that is prone to failure and not easily reproduced. Creating a provenance for such a process was challenging and so the source files for these patches, along with the code used to

import them into the database have been archived for reproducibility. On discussion with local data engineers, the following thematic reasons emerged as difficulties to changing the XML research pipeline:

1. resources were limited, and so there was not enough human resource to re-write the original data pipeline. However, a direct Structured Query Language (SQL) query on a case-by-case basis was quicker to action.
2. lack of institutional memory. The individuals who had written the original data pipeline had in some cases left the department, and these pipelines were written in a language that was not supported by remaining members of the team.
3. lack of familiarity with XML. Generally, individuals associated with the project at contributing sites had a much greater preference for working with SQL, rather than XML.

There were periods where data contribution for a given site either paused entirely, or for a certain subset of concepts. These instances were usually related to factors within the local hospital EHR itself, which were then unintentionally integrated into the research pipeline. This highlights the importance of have a continuous data quality monitoring process embedded within an EHR research pipeline.

Some deficiencies exist within these data that are yet to be fully accounted for. These include: an extremely high proportion of missing meta-data, non-contribution of certain concepts that are ubiquitous in critical care (for example, APACHE II scores or blood gas data), a general lack of alignment in concept units.

4.6.1 CC-HIC in Context

The CC-HIC research platform is compared and contrasted to others that exist in the critical care domain including ICNARC, MIMIC and Phillips eICU. A summary of this comparison is presented in table 4.15.

	CC-HIC	ICNARC	MIMIC-III	MIMIC-IV	DECOVID	PHILLIPS eICU
Timeframe	[2014, 2018]	[1993*-Present]	[2008-2014]	[2008-2019]	2020	[2014-2015]
Scope	multi-centre (regional)	multi-centre (national)	single centre	single centre	2 centres	multi-centre
ICU episodes	>40,000		53,423		Unpublished	200,859
Unique patients	>28,000		38,597	>40,000	Unpublished	139,367
Inclusion	all level 2/3 admissions	all level 2/3 admissions	all level 2/3 admissions	all level 2/3 admissions	all acute admissions ≥ level 1	Stratified sample
Scale	ICU only	ICU only (first 24 hours)	ICU only	ICU and some wards	all hospital	ICU only
Externally linkable	Partially (limited by data quality concerns)	No	No	No	No	No
Internally linkable	Partially (limited by data quality concerns)	Possible	Yes (at source)	Yes (at source)	Yes (via pseudo key)	Yes (pseudonym lookup not retained)
Longitudinal data	During ICU level care	No	Yes	Yes	Yes	-
Rich EHR representation	No	No	Yes	Yes	Partially	Yes
Findable	Partially	No	Yes	Yes	Partially	Yes
Accessible	Highly restricted	Restricted	Public	Public	Highly restricted	Public
Interoperable	No	No	Partially	Partially	Yes	No
Reusable	No	Yes	Yes	Yes	Yes	Yes
Sharing model	Centralised	Centralised	Centralised	Centralised	Centralised	Centralised
Local access to data	No	Curated	Yes	Yes	Partially	-
Data model	CC-HIC data model (bespoke)	ICNARC version 3/4	MIMIC version 3	MIMIC version 4	OHDSI version 5.3.1	bespoke
Free text	No	No	Yes	Yes	No	Structured
Imaging	No	No	No	Yes	No	-
Waveforms	No	No	No	Yes	No	-
Physiology	Some	Minimal	Comprehensive	Comprehensive	Comprehensive	
Treatment	Some	Minimal	Comprehensive	Comprehensive	Comprehensive	

Table 4.15: Comparison of major ICU data sharing collaborations. Note: *not all UK ICUs contributed data from the outset, however the capture is now extensive.

MIMIC is undoubtedly an exemplar and research leader in this field [12, 13]. There are a number of features that make MIMIC a natural choice for researchers who wish to work with routinely collected United States (US) critical care data. There are major differences in healthcare delivery between the US and the UK, and so one must be cautious about generalisations made when comparing research conducted against MIMIC to the UK. MIMIC presents a rich EHR representation of data, and have addressed many of the points that are drawn out in the discussion of the CC-HIC data model including sample versus result relationships and episode versus patient centric representations. Rather than proceed with a more generic CDM, MIMIC opted to developed their own bespoke data model. As early movers in this field, the landscape of widely used and validated CDMs at the time of MIMIC's creation was much less mature. Owing to the popularity of the OMOP CDM, the authors of MIMIC—Massachusetts Institute of Technology (MIT)—have written transformations of the MIMIC CDM into the OMOP CDM [187][9]. The MIMIC CDM is expressed as relational tables, and can be supplied in CSV files for import into a database of the researchers' choosing. This is a major distinction between the project infrastructure of the CC-HIC and MIMIC. MIMIC researchers are able to use the tools with which they are familiar, and even work on their own personal computer. This took considerable effort so as to address the concerns of confidentiality that would normally necessitate a more restricted model of access to data. This was supported by extensive anonymisation, including redaction of identifying features in free text notes and random offsets for all dates and times. The advantages for studies that focus on the analysis of non-identifying longitudinal physiology are potentially profound, since all the limitations of working within a restricted research environment have been removed. The cost of this accessibility and convenience is that any external linkage based on patient identifiers is no longer possible. Depending upon the broader goals of the research platform, this may be considered an acceptable loss.

Since data are so readily available within MIMIC in a standardised format, it

[9] though sadly the codebase for this transformation is no longer maintained.

has encouraged researchers to share their research code, producing a broad, high quality and validated codebase to perform common research tasks (for example the identification of ventilated patients) [13]. With a thriving and diverse research community, errors are identified through crowd-sourcing and reported back for correction via open issue trackers. This is analogous to the "inferential" stage of quality evaluation previously described and depicted in figure 4.1 on page 109. More recent developments have added imaging data to MIMIC in the form of chest radiographs [188, 189]. With each major revision, MIMIC moves closer to the ideal of presenting a full instantiation of the EHR in a research ready format.

An important driving factor behind MIMICs success is that it is conducted from a single private center, the Beth Israel Deaconess Medical Center in Boston, Massachusetts (US). It is undoubtedly easier to proceed through all the legal restrictions from a single centre. By comparison, attempts to release the CC-HIC research database as a public asset would require—in addition to a fundamental change in the underlying ethics provision for the project—oversight and approval from at least five different NHS trusts and a number of collaborator organisations including the NIHR.

In trying to learn from the successful open model of MIMIC, an anonymised representative sample cohort of 1000 patients from the CC-HIC research database was developed. This cohort was anonymised with the SDCmicro package for R [190], and applied data reduction, micro-aggregation and local suppression as methods to achieve a pre-determined degree of anonymisation. The goal of this process was that researchers could develop research code external to the UCL DSH and then import and continue to develop code internally at a later stage of development; hopefully shortening the development cycle and mitigating some of the challenges of working within a secure research environment. The anonymisation process encountered an unexpected conflict with the episode centric nature of the CC-HIC data model. Anonymisation removed the ability to link sequential episodes as spells, which would normally be performed by patient level identifiers, and so minor alterations to the data model were required for the cohort to keep its logical and time

ordered structure. This limited the success of the approach, since the data models inside and outside of the DSH were different enough to frustrate development when moving between them[10].

The Phillips eICU shares many similarities with MIMIC and can be thought of as a spiritual successor to MIMIC extended into the multi-centre domain. The eICU covers a large geographical area of the US where a tele-medicine service provided by Philips centrally aggregates certain physiological and treatment data feeds. This is quite distinct from the data sharing model of the CC-HIC where data harmonisation was required over a range of EHR vendors. Since all information comes from the same vendor, the process of data harmonisation was straightforward by comparison. There would be an expectation that a majority of core data feeds would be stored in a default pattern[11].

A core challenge for both the eICU and CC-HIC was that not all data feeds were available from all contributing centers. The information detailing which feeds are missing because they are not recorded in the first place is not accessible to the end user, and so it can be challenging to identify what information is missing because the center is unable contribute the data, and what is missing as an error.

ICNARC centrally aggregates summary patient data from a comprehensive number of ICUs across the UK bar Scotland. The ICNARC data model is mature, expressed in "strong specification" semantics, and tailored towards the specific audit and research requirements of the Case Mix Programme [24]. The current ICNARC data transfer process involves a lossy transformation of data from source hospital into the summary fields of the ICNARC data model. Because ICNARCs data collection is so vast, it covers many hospitals that do not implement an EHR. By comparison, the CC-HIC aggregate data from a much smaller number of digitally

[10]For interesting historical context, the OHDSI data model was developed in part so that the entire **person** table could be dropped from the data model, providing instant pseudonymisation without any loss of data integrity. This feature was developed to assist in the sharing of confidential healthcare data with pharmaceutical companies.

[11]There is some complexity here as sites are permitted to extend their local interface to capture unique data elements. The quality of how this is performed is therefore dependent on the individuals involved and can force the creation of semantic and ontological build errors without appropriate training.

mature sites, and tries to do so with much greater depth; attempting to translate much of the full longitudinal representation of the EHR into its data model. The goals and infrastructural set up of ICNARC and the CC-HIC are therefore distinct.

4.7 Conclusion & Recommendations

A comprehensive data QE has been implemented by inspectEHR [6]. In doing so, it has revealed many areas of concern within the CC-HIC research database that were previously unknown. This should impose some conservative expectations on the clinical research that is possible with this resource. A core goal of implementing a data model and sharing data across multiple healthcare organisations is that of data harmonisation. While some of the foundational elements for data harmonisation do exist within the CC-HIC, by large this process has not occurred, with alignment in many cases happening by chance rather than design.

Based on the review of the CC-HIC data model and the quality of the CC-HIC database, the following recommendations can now be made:

1. define the research data needs before developing a data model or sharing data.
2. make use of pre-existing open data modelling standards (for example OHDSI), and favour the extension of existing models, rather then developing new bespoke models, unless absolutely necessary to complete research goals.
3. consider the use of a person centric data model for healthcare data.
4. apply data normalisation principles to the degree necessary to fulfil analytic goals.
5. make use of interchange standards for data that are routinely used by data engineers working at contributing sites. For example, it may be prudent to exchange data in CSV format, even if XML provides a better technical standard, if the former is more familiar to key stakeholders.
6. transfer data in a format that can be inspected with a limited toolset by humans. Tabular structures are particularly useful in this regard.
7. avoid lossy transformations of data wherever possible.

8. prioritise the transfer of raw data concepts and document when data concepts are derived during the ETL process as meta-data within the data model.
9. employ a "weakly specified" data specification, that takes full advantage of local ontological mappings that already exist. Avoid hard coding data requests into the data model.
10. should a hospital not be able to contribute a requested data concept, document this information. Preferably as meta-data within the data model.
11. conform to international semantic and interoperability standards, such as SNOMED.
12. avoid arbitrary choices in data representations, and apply rules with consistency.
13. discourage data submissions outside the agreed pipeline, but ensure that the research pipeline is serviceable for the long term duration of the project.
14. develop a data evaluation pipeline in parallel to the data model, ensuring that an iterative feedback cycle is an intrinsic component of any data submission. Thereby creating an expectation that information exchange is bidirectional.
15. include both local and central data evaluations where possible.

These recommendations now form the basis of the next phase of the CC-HIC project. As of writing, the OHDSI data model is being implemented as a replacement for the CC-HIC data model in a modular fashion. Database tables are submitted one-by-one, and populated incrementally with data according to research need and data availability. This has placed data quality at the center of the next phase of the CC-HIC project. These recommendations also contributed to the development of the DECOVID data model, which similarly implements a variation of the OHDSI data model.

Part II

Clinical Research

Chapter 5

Cumulative Exposure to Excess Oxygen

The aim of this exemplar study is to determine whether cumulative exposure to oxygen levels in excess of clinical need are associated with increased ICU mortality. This study makes use of the unique features of the CC-HIC data resource that are not presently available elsewhere in the UK. First, longitudinal data capture, including arterial blood gas sampling. This allows for the exploration of oxygen exposure as a necessarily longitudinal drug exposure. Second, a large number of available cases enabling the detection of what are likely to be small statistical signals. The principal findings from this chapter have been published [4][1]. Details of the literature review and search strategy for this Chapter can be found in Appendix Section B (page 271).

5.1 Background

The possible harms associated with extremely high levels of oxygen administration are well documented in humans [51, 40]. What is less clearly understood is the dose response relationship between excess oxygen and mortality and whether or not the levels of oxygen that patients are administered in routine practice are detrimental.

There are several facets of this problem that make it methodologically challenging. First, oxygen exposure is longitudinal in nature. While there are biolog-

[1]Minor perturbations between these published results and the results presented in this thesis are anticipated. This reflects updates to the data quality procedures and progression in scientific thinking.

ical mechanisms that could contribute to harm on a relatively short time-scale (for example, vasoconstriction and absorption atelectasis), the primary means through which mortality is proposed to manifest—namely ARDS, lung fibrosis and increased inflammation—take time to accumulate and impart their harm. This has been a major limitation of most prior studies in this field of research, where data availability has been typically restricted to the first 24 hours of intensive care. In order to measure excess oxygen exposure, a window of observation is required to demonstrate an effect. This creates a tension with the attrition of patients from ICU from death and discharge; longer periods of observation are only possible for a cohort of diminishing size. As a longitudinal exposure in critical care, oxygen exposure is also potentially subject to informatively missing data patterns as previously outlined. This could lead to bias if not specifically accounted for. Second, the administration of oxygen is a therapeutic intervention for the correction of hypoxaemia, which is itself a manifestation of illness severity. The effect of exposure to oxygen on mortality is therefore confounded by acute illness severity and respiratory impairment; so-called confounding by treatment indication. There is no routinely observed feature of the patient that would allow one to isolate the causal pathway of potential harm between oxygen exposure and mortality. Therefore, the risk will always remain for residual confounding to hinder inferences, particularly if studying oxygen exposure directly as the F_IO_2. Last, there is no clear feature of the patient that is routinely monitored that provides an unambiguous definition of "oxygen excess", and so one must be created.

There are notable patient groups that may be more susceptible to the effects of excess oxygen, including those who are mechanically ventilated, or those with sepsis or COPD. Mechanically ventilated patients may be at increased risk of harm from oxygen toxicity mechanisms, since the mechanisms that are proposed to mediate harm could be potentially amplified by stresses to the alveolus caused by ventilation [68]. There are conflicting data as to whether or not high levels of oxygen could be harmful or helpful in sepsis [53]. COPD is established to have worse outcomes with high levels of oxygen exposure, mediated by a mechanism that is distinct from

those previously discussed [191]. In this regard, COPD could be used as a yardstick in models to determine if known effects have been appropriately captured.

5.1.1 Identifying Markers of Excess Oxygenation

Several features of oxygen administration are recorded in the EHR. This includes:

- fraction of inspired oxygen (F_IO_2).
- peripheral oxygen saturation (SpO_2).
- P_aO_2/F_IO_2 ratio or SpO_2/F_IO_2 ratio.
- partial pressure of oxygen in arterial blood (P_aO_2).

Each of these features represents an imperfect measure of oxygenation, or exposure to excess oxygen.

F_IO_2 is a logical primary candidate to investigate since this is the direct administration of oxygen to the patient [70, 192]. When investigated in RCTs, it is the F_IO_2 that is modified as the therapeutic intervention under investigation [52], typically to target a particular SpO_2 or P_aO_2. From an observational perspective, as a marker on its own, it is difficult to determine what constitutes "excess" F_IO_2, as opposed to what is in keeping with a patient's requirements. A patient who is maintained on pure oxygen (an F_IO_2 of 1.0) for long periods of time will already be at high risk of a poor outcome, regardless of the contribution made by oxygen exposure itself. F_IO_2 is therefore confounded by treatment indication, and it may not be possible to gain appropriate statistical control [193, 194]. To reiterate, the risk of naively including F_IO_2 in a statistical model to investigate the potential harms of oxygen exposure, is that high F_IO_2 will, in all likelihood, show a close relationship with increased mortality, regardless of how one applies statistical adjustment. But this will often be commensurate to the patient's increased clinical needs, rather than indicating direct harm itself. In lieu of a reliable instrumental variable within the CC-HIC database, or other means to apply a principled causal methodology, this is problematic in of itself.

Due to the sigmoid nature of the oxyhaemoglobin dissociation curve, SpO_2 demonstrates a ceiling at 100%, which is reached at even low levels of supplemen-

tary oxygen. This variable is thus limited in its capacity to reveal the effects of excess oxygenation. With the exception of chronic obstructive pulmonary disease (COPD)—a respiratory disease with well established oxygen sensitivity—most patients in intensive care are prescribed oxygen with a lower treatment threshold (e.g. to maintain an $SpO_2 > 92\%$). From a modelling standpoint, one could consider SpO_2 as a biomarker with a censored distribution, with an upper detection threshold at 100%. Re-considering SpO_2 as a threshold detection problem would provide a principled means through which SpO_2 could be studied, despite the restricted information it contains. Complicating matters, pulse oximitry (the method by which SpO_2 is measured) is subject to a relatively high degree of measurement error, and systematic biases from both acute physiology and patient ethnicity [195].

The P_aO_2/F_IO_2 and SpO_2/F_IO_2 ratios provide a reflection of lung function, though both can be augmented by a change in ventilation strategy, which is not necessarily reflective of an improvement in underlying lung function. The ARDSNet study showed that patients with better P_aO_2/F_IO_2 ratios (suggesting improved lung function) experienced worse outcomes [129]. The prevailing consensus attributes this to the more liberal ventilation settings used in this group. In this respect, it may be more useful to use the P_aO_2/F_IO_2 ratio as a direct means to control for impairment of the respiratory system, rather than to investigate exposure to oxygen itself. Given the sensitivity of the P_aO_2/F_IO_2 ratio to changes in ventilation strategy, it has been suggested that the P_aO_2/F_IO_2 ratio is also modelled alongside positive end expiratory pressure (PEEP) as a means of accounting for the effect of ventilation, though this approach has also been contested [196]. The P_aO_2/F_IO_2 ratio will have the inherent *disadvantage* of only being recorded intermittently on the ICU with a sampling frequency related to the illness severity of the patient. Conversely, the SpO_2/F_IO_2 ratio has similar properties to the P_aO_2/F_IO_2 ratio, but is measured more frequently and likely without the sampling bias observed in the P_aO_2/F_IO_2 ratio. The SpO_2/F_IO_2 ratio is subject to the same distributional artefacts caused by the ceiling effect observed in SpO_2. The relationship between SpO_2 and P_aO_2 may be altered by pathophysiology and ageing [197, 36]; the former may present

challenges to control for within an observational cohort.

From a biological standpoint it remains unclear which of these biomarkers (or combination thereof) provides the best measure to elucidate potential effects on outcomes.

The P_aO_2 represents the partial pressure of oxygen present in the artery. Under reasonable conditions encountered in clinical practice, the P_aO_2 cannot exceed 13.3 kPa without the administration of supplemental oxygen. This limit is governed by the alveolar[2] gas equation shown in equation 5.1. The alveolar gas equation describes the partial pressure of oxygen in the alveolus (P_AO_2), which must be higher than that of the artery (P_aO_2) due to the concentration gradient that drives the movement of oxygen into the body.

$$\begin{aligned} P_AO_2 &= (P_{atm} - P_{H_2O}) \times F_IO_2 - \frac{P_aCO_2}{RQ} \\ P_AO_2 &= (101 - 6.18) \times 0.21 - \frac{5.5}{0.8} \\ P_AO_2 &\approx 13.3 \end{aligned} \tag{5.1}$$

There is a limited clinical rationale that can be used to justify why a sustained P_aO_2 in excess of 13.3 kPa would be necessary in critical care[3]. This threshold allows the creation of a less ambiguous (albeit imperfect) definition of excess oxygen: a $P_aO_2 \geq 13.3$ kPa. This definition of oxygen excess has the following limitations:

1. F_IO_2 is the direct exposure of interest, and so a high P_aO_2 would likely act a surrogate for potential harm.
2. while a $P_aO_2 \geq 13.3$ kPa is unambiguously in excess for most clinical situations, there will be patients who are in relative clinical excess below this threshold. For example, patients with pre-existing lung disease who do not maintain a P_aO_2 of 13.3 kPa under normal circumstances and have acclimatised to lower levels.

[2] the alveolus is the functional anatomical unit of the lung; an air sac responsible for the exchange of gas between the atmosphere and the blood stream

[3] High flow oxygen is used in certain clinical conditions where the patient does not display hypoxaemia; this would include pneumothorax and carbon monoxide poisoning.

3. there is a complex relationship between F_IO_2 and P_aO_2, where different pathologies lead to different degrees and types of decrement in oxygen between the alveolus and bloodstream. An investigation targeted at P_aO_2 would, by default, ignore these differences.
4. there will invariably be patients who have too high a degree of respiratory impairment to increment their $P_aO_2 \geq 13.3$ kPa, and so these patients would not be able to contribute to the variability in this constructed variable.
5. imposing a threshold on P_aO_2, even on biological grounds, does not make use of full information and so could be considered statistically suboptimal.

Excess oxygen under the 13.3 kPa threshold definition can be converted into a longitudinal exposure by parametrising repeated P_aO_2 samples as the cumulative effects biomarker morphology (as discussed in subsection 2.4.1 on page 49). The cumulative effect of $P_aO_2 \geq 13.3$ kPa is calculated as the area under P_aO_2-time curve, bounded by 13.3 kPa, as shown in figure 5.1. For brevity, I refer to this parametrisation as "cumulative hyperoxaemia", which takes the units kPa.hours. To enhance comparisons between differing lengths for the observation window of *potential exposure* to oxygen, a time weighted average can be taken. This divides cumulative hyperoxaemia by the number of hours of potential exposure, returning the constructed variable of "hyperoxaemia dose" in the more natural units of kPa.

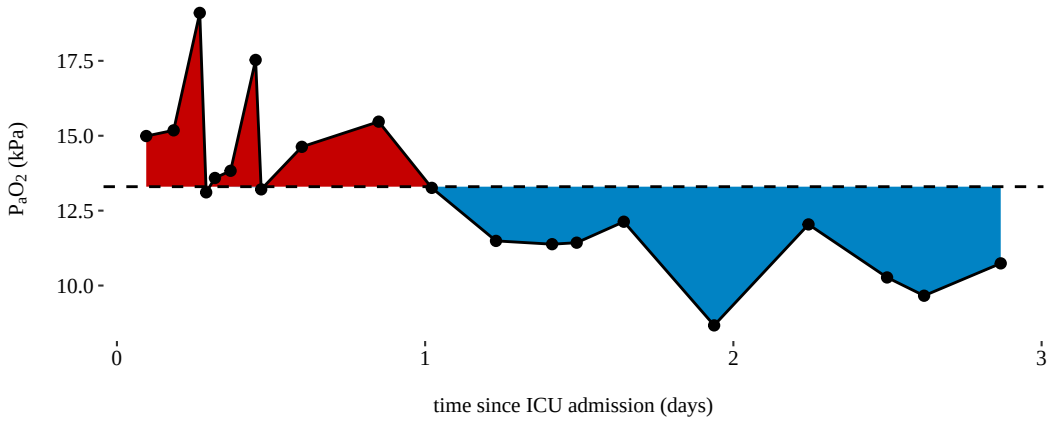


Figure 5.1: Illustration of cumulative hyperoxaemia for a random patient in the CC-HIC database. The red shaded area demonstrates the total exposure to a $P_aO_2 \geq 13.3$ kPa

For example, 1 kPa of hyperoxaemia dose for an exposure window of 24 hours describes that a patient's P_aO_2 was 1 kPa above 13.3 kPa on average for the duration of those 24 hours.

5.2 Hypothesis Statement

Models will be constructed to evaluate the following hypothesis:

- exposure to oxygen in excess to clinical requirements—as defined by the hyperoxaemia dose—is associated with increased mortality in the general critical care population.

I also explored particular patient subgroups who are at a perceived elevated risk from exposure to excess oxygen, including patients with: COPD, sepsis or those receiving mechanical ventilation.

5.3 Methods

5.3.1 Cohort Definition

The primary cohort for this study were all adult ($\geq$ 18 years) index spells submitted to the CC-HIC from 31st January 2014 to 31st December 2018. Spells were included in the study if they had sufficient quality represented by the following criteria:

1. spells (and their constituent episodes) could be unambiguously characterised.
2. spells contained at least one P_aO_2 sample.

Spells were excluded on the following grounds:

1. a spell length of stay less than 24 hours.
2. spells used for pre-surgical preparation only.
3. the presence of any treatment limitation orders.
4. cardiopulmonary resuscitation in the 24 hours preceding ICU admission.

Patients with a length of stay less than 24 hours were most commonly admissions for post-anaesthetic care following elective surgery. These patients have a very low risk of mortality, and limited observable exposure of oxygen, and so are unlikely

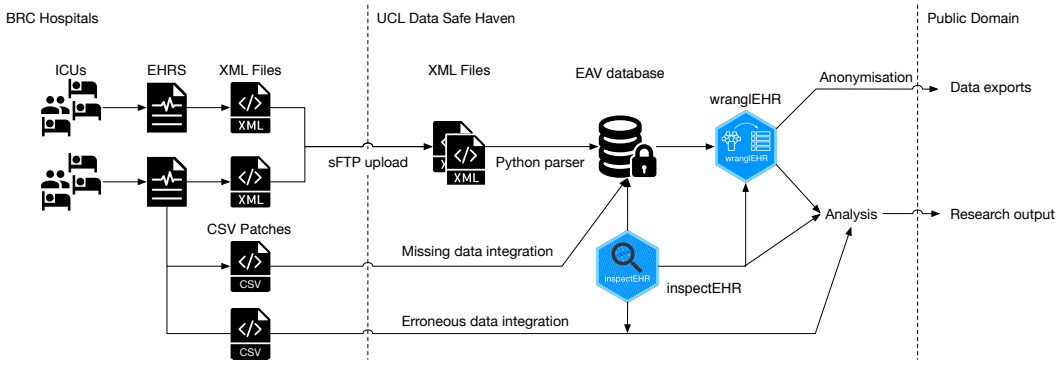


Figure 5.2: Manual data integration external to the CC-HIC pipeline. Data that are missing from the CC-HIC research database are processed into the correct format (EAV) and directly added to the database with their own provenance. Replacement data for concepts that are present in the CC-HIC research database, but erroneous, are processed downstream of the CC-HIC research database and integrated into the study directly.

to contribute a meaningful signal for detection. The same justification is true for patients admitted for pre-surgical preparation. Treatment limitations may preclude escalating oxygen to clinical requirement. Patients who undergo cardiopulmonary resuscitation have an exceedingly high mortality and commonly receive 100% oxygen with little regard for clinical requirement [65, 198, 199]. For the purposes of this study, they were removed, though they do present an interesting cohort to study in future work.

5.3.2 Maximising Available Data

Limited data availability via the main data pipeline (the XML pathway) for core concepts required for the study, jeopardised its viability. This stems from three main areas of concern:

1. one site was not able to contribute any blood gas data.
2. several sites contributed incorrect APACHE-II scores; the primary means of risk stratification built into the CC-HIC data model.
3. a large amount of meta-data used to distinguish P_aO_2 (arterial) from P_vO_2 (venous) were missing.

One site was not able to submit arterial blood gas data for their cohort via the XML pathway. The reasons for this relate to resource availability for ongoing main-

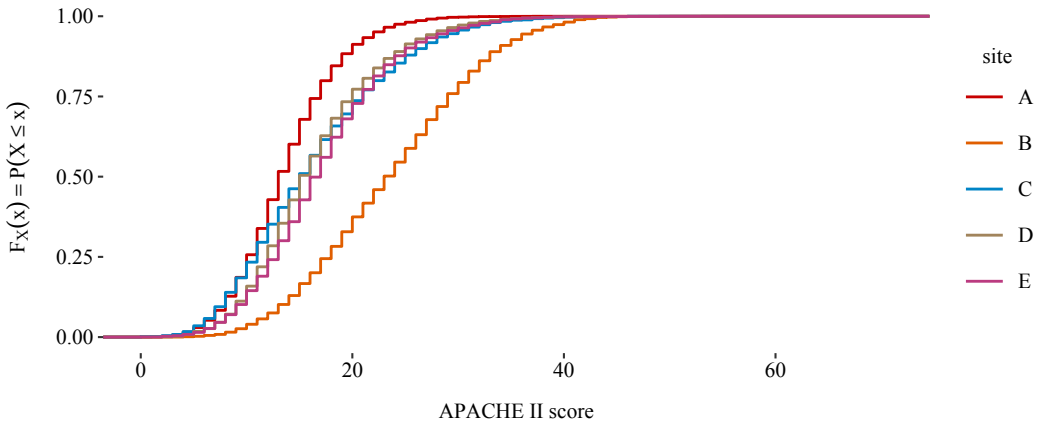


Figure 5.3: Distribution of the APACHE-II score after corrective submissions from each site. A known issue with the oxygen component positively biases the Cambridge (RGT) distribution.

tenance of the existing local XML pathway and has been previously discussed. In order to increase the number of cases that could be used in the study, a data patch was provided external to the XML pathway with P_aO_2, P_aCO_2 and pH values for all patients from this site. These data were manually integrated into the CC-HIC database as illustrated in figure 5.2. Patients in the manual extract were identified and linked into the main CC-HIC database by their NHS number and ICU admission date. Blood gas data were standardised to match the CC-HIC data model and assigned to an episode in the CC-HIC database should the datetime stamp of the blood gas fall between the start and end of a fully characterised episode that already existed in the CC-HIC database. A similar process was used to add new APACHE-II data. Since the APACHE-II data conflicted with existing data that had been contributed, these new data were *not* read into the CC-HIC research database as they would cause a conflict with existing data provenance. Instead these data were integrated downstream from the CC-HIC research database into the study directly.

Figure 5.3 shows the distribution for the APACHE-II score stratified by each site *after* the corrections had been made. Even after these corrections there was a large discrepancy between sites that is difficult to account for by case-mix alone. A known miscalculation for site B's data remains whereby the oxygen component of

the APACHE-II score had been miscalculated. Due to the nature of the error, this is likely to increase a patient's oxygen subcomponent score by a maximum of 2 points (from a total of 71) and so thought not to be particularly relevant.

5.3.3 Labelling Arterial Blood Gases

It is necessary to identify the anatomical source of any blood gas sample, to distinguish whether the sample originates from the arterial or venous circulation. The anatomical source of the blood gas is represented in the CC-HIC data model as meta-data, of which a significant proportion are missing. As of writing, 1.6×10^6 samples with a partial pressure of oxygen have been submitted to the CC-HIC database, of which 0.4×10^6 (25%) do not contain an anatomical label. This missingness was not isolated to any particular site or patient cohort that would explain a systematic fault for the missing data. This missingness is at odds with standard clinical practice since the clinical interpretation of a blood gas relies on knowing its anatomical source. Thus these labels should be complete to a high degree within any EHR. One possible explanation is that the true anatomical source of these unlabelled samples was something other than "arterial" or "venous" and therefore not specified in the CC-HIC data model. For example, extra-corporeal blood gas samples are taken for the routine monitoring of calcium levels during citrated RRT. Even if this is the case, it remains that a 25% missingness, without localisation to a particular patient group, is still much higher than would be clinically expected.

A logistic regression was used to model the anatomical source of unlabelled blood gases. The outcome variable was the anatomical source, with the positive indicator representing an arterial origin. The predictor variables chosen for inclusion were the P_xO_2[4], P_xCO_2, blood acidity (pH), all pair-wise and three-way interactions. Variables were transformed onto the unit-variance scale prior to inclusion in the model. There are more variables that form part of the standard blood gas panel that could potentially be used to improve this model. However, missing data patterns indicated that these other variables contained within the standard blood

[4]For clarity, I use the conventions of P_a, P_v and P_x to describe a partial pressure obtained from the artery, vein or unknown source respectively.

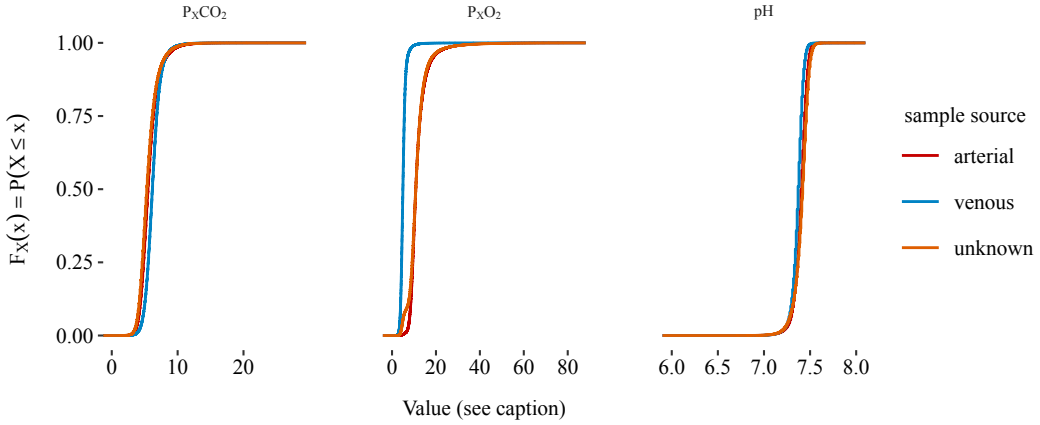


Figure 5.4: Univariate empirical cumulative distribution functions for PCO_2 (left), PO_2 (middle) and pH (right). There are overt differences between arterial and venous samples for PO_2. Differences in PCO_2 and pH are present, but subtle.

gas panel are far more frequently missing in the CC-HIC research database, and so inclusion of additional variables would likely fail to support the primary motivation of the model; to predict anatomical labels (rather than extract particular inferences). By restricting the problem to use only these three predictors, which are present for almost all samples, a larger proportion of blood gases that are missing anatomical labels could be relabelled. Individual blood gas samples were treated as independent, even if they came from the same patient. This approach renders many of the inferences from the model invalid, for example, by shrinking standard errors and making the model unreasonably confident in its estimates. A multilevel approach was initially taken (samples nested in patients), however, as convergence issues were encountered, this approach was abandoned in favour of a simpler model specification.

Figure 5.4 demonstrates the cumulative density functions for the three predictor variables stratified by their known anatomical labels (arterial or venous) or unknown status. A large difference exists between the arterial and venous distribution of the P_aO_2, with a notably smaller difference between those of the P_aCO_2 and pH distributions.

Figure 5.5 shows the resulting coefficients from the fitted model. Bootstrapped 95% confidence intervals are drawn using 100 resamples, though even these confi-

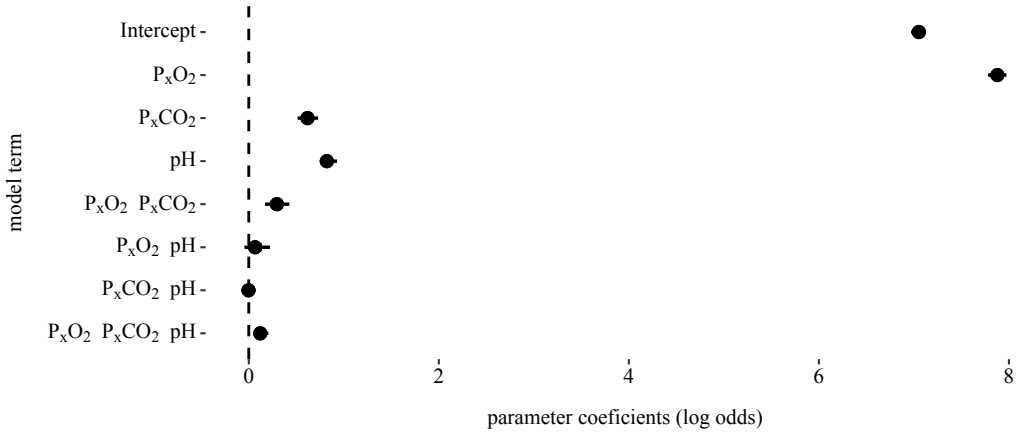


Figure 5.5: Coefficients for predictive variables in the blood gas re-labelling model.

dence intervals will be over confident given the model formulation described above. These coefficients should only be used to provide an indication as to the relative importance of variables within the model. The model intercept is positive, indicating the increased likelihood that a sample is arterial if PO_2, PCO_2 and pH are all measured at their mean values of 10.5 kPa, 5.8 kPa and 7.39 respectively. This reflects the higher number of arterial samples in the CC-HIC database.

The performance characteristics of the reclassification are shown in figure 5.6. The area under the receiver operator and precision recall curves was 0.98 [0.98, 0.98] (estimate [bootstrapped 95% confidence intervals]) and 0.97 [0.97, 0.97] respectively[5]. The performance characteristics were evaluated by applying the 100 bootstrapped models to a reserved test set (15% of all labelled samples). The high performance of this model can be attributed to how the logistic regression is modelling a physico-chemical environment; as such there are strict chemical laws that govern the relationship of the variables under investigation [200]. In other words, there are considerable differences between the arterial and venous circulation which can be taken advantage of by the model.

The resulting model coefficients were implemented as a function in inspectEHR [6] allowing other researchers to reproducibly re-label blood gases. The re-labelled blood gases were *not* embedded in the CC-HIC database as metadata labels as this would potentially cause confusion with regard to data provenance.

[5]These bootstrapped confidence interval were very tight around the mean.

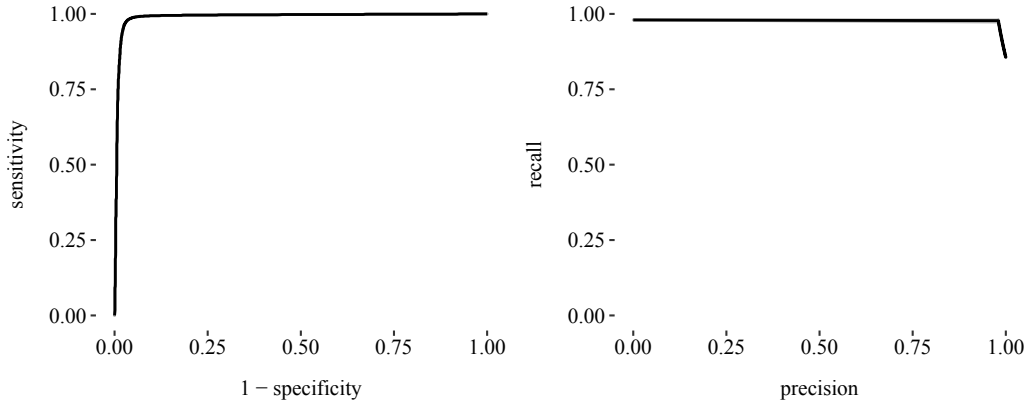


Figure 5.6: Discrimination curves for blood gas labelling model. Left: Receiver operator characteristic curve with area under the curve 0.98. Right: Precision recall curve with area under the curve 0.97. Confidence intervals are shown, though they are extremely tight around the central estimate.

5.3.4 Ventilation Phenotype

Following review of the CC-HIC data model and data quality evaluation, it was anticipated that defining ventilation within the cohort would be non-trivial. This stems from a lack of semantic expressiveness for ventilation coupled with the mixed use of measured and derived concepts relating to ventilation within the data model. Further, missing data patterns suggested that not all sites contribute all the concepts that relate to ventilation. Ventilation is a complex therapy, and represented through a large constellation of fields in the typical EHR. The absence of ventilation is unlikely to be explicitly asserted as it does not form part of routine documentation[6]. Rather, the absence of concepts relating to ventilation must be used to infer that a patient is not receiving ventilation.

The requirements for this study are modest since it is only necessary to distinguish patients who are ventilated from those that are not. A suitable scheme for this simple phenotype would therefore be the following:

1. no ventilatory support.
2. ventilatory support (optionally divided into invasive or non-invasive).

This would allow for a straightforward delineation of patients who are ventilated

[6]It would be quite strange in fact to document what therapies the patient is *not* receiving, since this list would be unfathomably long.

from those who are not, with an optional additional level of detail to identify those who receive invasive mechanical ventilation. There *are* concepts expressed in the CC-HIC data model that make triangulation on this simplified ventilation phenotype possible. However, systematic differences in the way in which sites contribute data render a general purpose ventilation phenotype more challenging.

Each site contributes the concepts that relate to ventilation in a unique way. This is demonstrated in figure 5.7 which shows the correlations between data submissions for these concepts in the CC-HIC database. The patterns of data contribution are unique to each site, and not wholly explainable through missing data alone. The monitoring of ventilated patients is standardised in the UK [201], and so it is likely that at least some of this pattern of data contribution represents transcription errors.

The potential concepts from the CC-HIC data model from which a ventilator phenotype could be derived are enumerated below (where "missing" denotes missing data, rather than a specific level called missing):

1. airway:
    - none (positive declaration).
    - missing.
    - endotracheal tube.
    - tracheostomy tube.
2. ventilation settings:
    - any of: tidal volume, ventilator respiratory rate or airway pressure present. These are all concepts related to a ventilator and should not appear in documentation unless ventilation is active.
    - missing.
3. ventilation status:
    - invasive ventilation.
    - non-invasive ventilation.
    - missing.

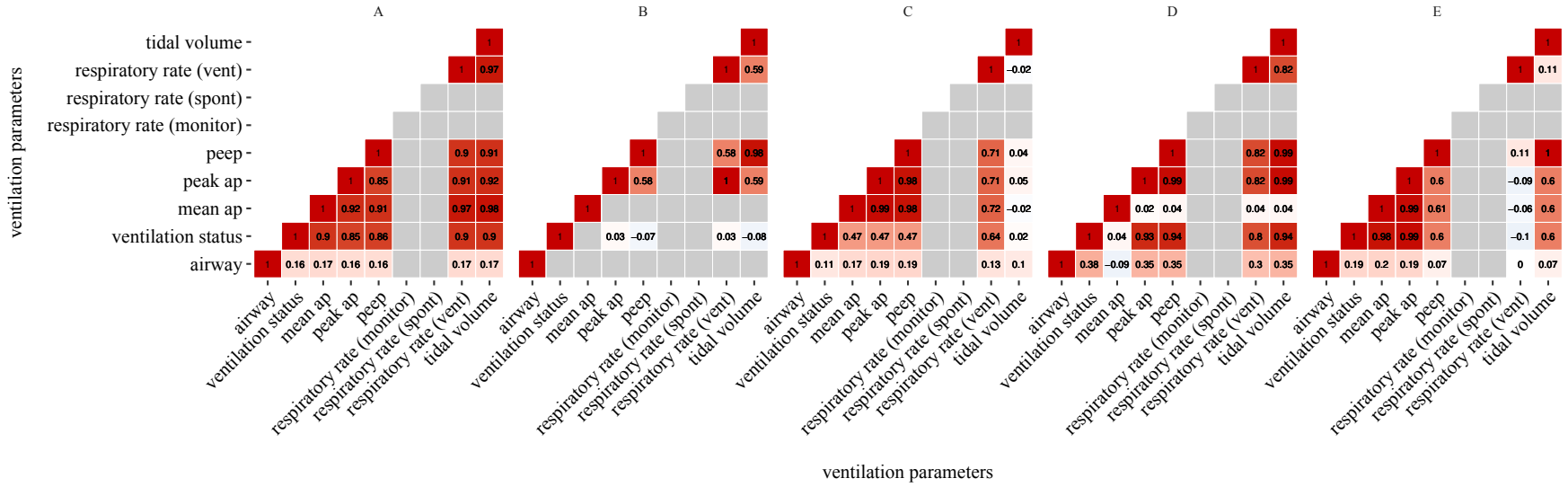


Figure 5.7: Correlations in missing data patterns for ventilation parameters. There are clear differences demonstrated in how concepts are documented between sites. Concepts that are not contributed are indicated as grey cells.

The combinations of these levels have been enumerated in table 5.1 alongside a proposed ventilator phenotype. The corresponding phenotype is provided with a level of confidence determined by how much missing information is being used to triangulate on a level. An important feature is that it is much easier under this phenotype to identify that a patient is receiving ventilator support, than to identify that they are not. Only one combination of concepts positively identifies a patient as being free of ventilation. This particular combination is dependent upon the use of "none" documented as the airway. It is important to consider this as a limitation of any study that makes use of ventilation inside the CC-HIC database. With a limited ability to determine who is *not* in receipt of ventilation, there may be a bias toward patients demonstrating more severe respiratory failure.

A number of illustrations are provided in figures 5.8 to showcase the varied underlying data available to construct the ventilation phenotype. These illustrations were chosen as they showcase typical patterns of data contribution seen, both in terms of the concepts and frequency with which they are contributed.

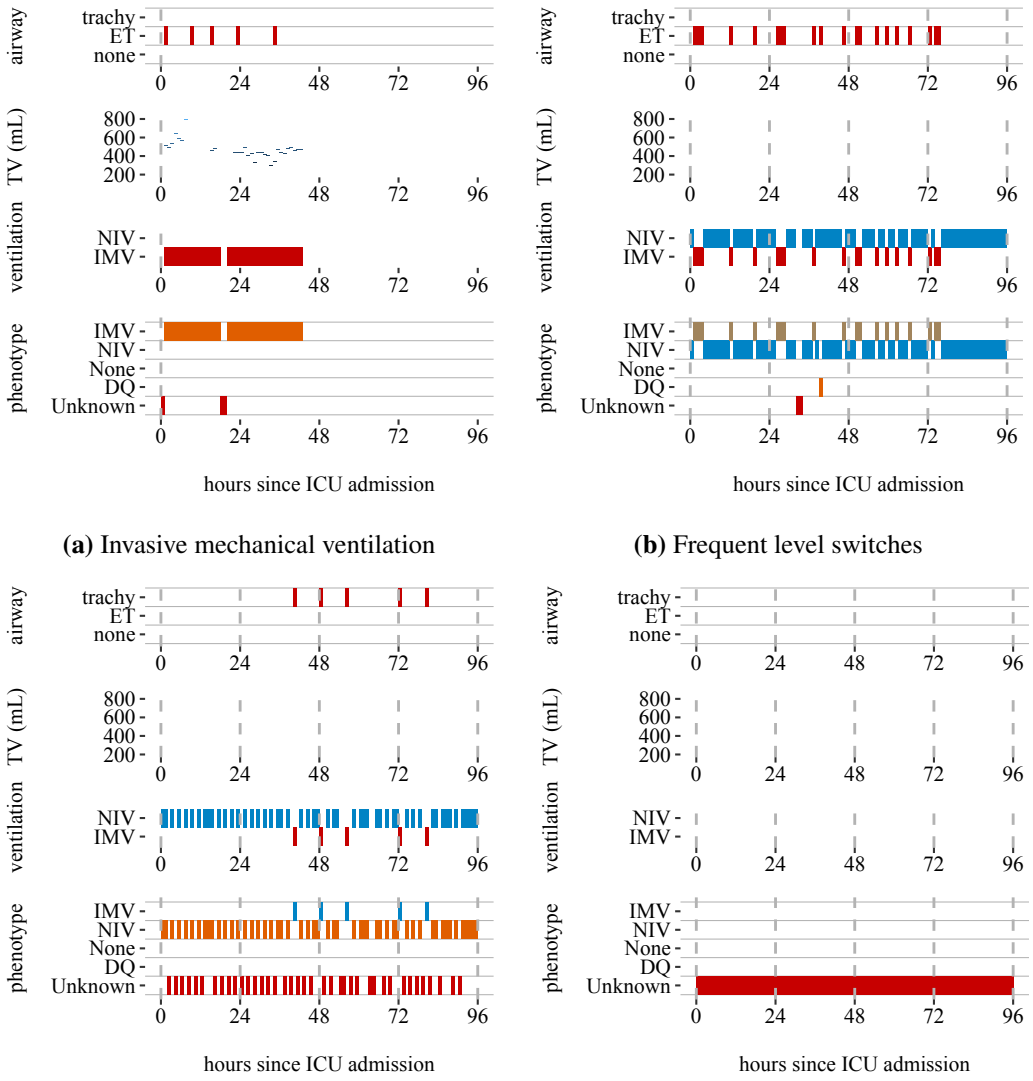
In order to evaluate the phenotype a random sample of spells was visually inspected. This provided a subjective—but domain knowledge-led—approach to determine the quality of the phenotype from the available data. The proportion of "error" phenotype labels (where logically inconsistent parameters are observed as highlighted in table 5.1) is enumerated in table 5.2. 7.5% of these data are shown to produce erroneous labels. The majority of the phenotype produces an "unknown" label, though this likely represents times when the patient is not receiving advanced respiratory support. From this review, I decided that the available data would likely neither support a phenotype at a higher than daily resolution, nor one that distinguishes invasive from non-invasive ventilation. This primarily related to the high number of oscillations between neighbouring levels of the ventilator phenotype which are unlikely to be observed in practice. By limiting the scope of the phenotype to distinguish between any form of ventilation and none, these oscillations are less likely to adversely impact on the analyses that follow.

Data concepts			Ventilator phenotype	
Airway	Ventilator settings	Ventilator status	Phenotype	Confidence
Tracheostomy	Present	Non-invasive	Error	-
Endotracheal	Present	Non-invasive	Error	-
None	Present	Non-invasive	Non-Invasive	High
Missing	Present	Non-invasive	Non-Invasive	Moderate
Tracheostomy	Missing	Non-invasive	Unknown	Low
Endotracheal	Missing	Non-invasive	Error	-
None	Missing	Non-invasive	Non-Invasive	Moderate
Missing	Missing	Non-invasive	Non-Invasive	Low
Tracheostomy	Present	Invasive	Invasive	High
Endotracheal	Present	Invasive	Invasive	High
None	Present	Invasive	Error	-
Missing	Present	Invasive	Invasive	Moderate
Tracheostomy	Missing	Invasive	Invasive	Moderate
Endotracheal	Missing	Invasive	Invasive	Moderate
None	Missing	Invasive	Error	-
Missing	Missing	Invasive	Invasive	Moderate
Tracheostomy	Present	Missing	Invasive	Moderate
Endotracheal	Present	Missing	Invasive	Moderate
None	Present	Missing	Non-Invasive	Low
Missing	Present	Missing	Unknown	Low
Tracheostomy	Missing	Missing	Invasive	Low
Endotracheal	Missing	Missing	Invasive	Moderate
None	Missing	Missing	None	High
Missing	Missing	Missing	Unknown	High

Table 5.1: Enumeration of all possible levels of the simplified ventilation phenotype in CC-HIC. Possible levels are: "Error"; the combination of values does not lead to a logical conclusion. "Unknown"; the combination cannot identify a particular level of respiratory support. "None"; no respiratory support (a positive statement). "Non-invasive"; non-invasive support. "Invasive"; invasive ventilation. A level of confidence in the phenotype label is provided determined by how much missing information is necessary for the label, with more missing information reducing the confidence in the label.

Ventilation phenotype	Count	Proportion
Invasive mechanical ventilation	276,258	30.5%
Non-invasive ventilation	191,496	21.2%
No ventilatory support	33,751	3.2%
Unknown status	336,600	37.1%
Conflict (data quality issue)	68,071	7.5%

Table 5.2: Frequency of different ventilator phenotypes in the CC-HIC database. "No ventilatory support" is under-represented from a lack of positive indicators in the CC-HIC data model for this level of the phenotype. There is a commensurate increase in "Unknown status" labels which, in many cases, will represent no advanced ventilatory support.



(a) Invasive mechanical ventilation

(b) Frequent level switches

(c) Non-invasive ventilation

(d) No data contributed

Figure 5.8: Illustrated ventilator phenotypes. Panels show different typical realisations of the ventilator phenotype for real patients drawn from the CC-HIC database. **a)** A patient is in receipt of invasive mechanical ventilation with all contributing data in agreement. The different frequency with which concepts are contributed (airway status four times per day) creates some gaps in the phenotype when creating a phenotype with hourly resolution and no accommodation of concept lag. **b)** There are frequent level switches between NIV and IMV, which is not clinically possible. It is likely that an error has been introduced on data export from the EHR (transcription error). **c)** Patient is mostly receiving NIV. Some occasional invasive episodes facilitated by a tracheostomy are plausible. **d)** No data are contributed. The likelihood is that the patient received only conventional oxygen via a facemask.

5.3.5 Power

Current information that could be used to inform a power calculation is limited by studies that have either targeted surrogate outcomes [52] or mortality but within a highly enriched population [202] that would not be representative of the more general cohort of the CC-HIC. Regardless, the current research landscape suggests that most studies to date have been powered optimistically. A simulation was conducted to provide an impression of the anticipated power for different possible effect and sample sizes. Mortality in the unexposed group was fixed at 5%, with varying proposed effect sizes that correspond to odds ratios of 1.1 (pessimistic), 1.3 and 1.6 (optimistic). Sample sizes were evaluated at various points from 5,000 to 40,000 patients. Each simulation was run 100 times to provide an empirical estimate of statistical power at the convention 0.05 alpha threshold. A summary of results are shown in table 5.3. For optimistic effect sizes (odds ratio 1.6) power ranged from 79% (n = 5000) to 100% (n = 40,000), while for pessimistic effect sizes (odds ratio 1.1) power ranged from 7% (n = 5000) to 39% (n = 40,000). It would be reasonable to conclude that even with the large convenience sample offered by the CC-HIC, this study question is unlikely to be overpowered and indeed may well be underpowered.

sample size	log odds ratio	odds ratio	power
5,000	0.1	1.1	0.07
	0.3	1.3	0.39
	0.5	1.6	0.79
10,000	0.1	1.1	0.07
	0.3	1.3	0.67
	0.5	1.6	0.98
20,000	0.1	1.1	0.18
	0.3	1.3	0.92
	0.5	1.6	1.00
40,000	0.1	1.1	0.39
	0.3	1.3	0.99
	0.5	1.6	1.00

Table 5.3: Results of simulated power analysis

5.3.6 Procedure

Characterised episodes were coalesced into spells as previously outlined. When combining data from episodes into spells, it was necessary to reconcile data that were duplicated across multiple episodes. Data concepts relating to the start of a spell (e.g. admission diagnosis or admission APACHE-II score) were taken from the primary episode of the spell. Data concepts that relate to the end of a spell (e.g. patient outcomes) were taken from the final episode of the spell. Should data concepts be missing from any particular episode, concepts were obtained opportunistically from any available episode, provided this did not cause a logical inconsistency (for example, a patient dying in their first episode, but progressing onto a second episode evidently alive).

Data were extracted from the CC-HIC database using the wranglEHR [5] package for R [137] using a 30 minute base cadence. Longitudinal data submitted at a higher frequency were summarised over this 30 minute period as either the mean, mode, maximum or minimum value as deemed clinically relevant to signify the *most* deranged physiology of that 30 minute window.

Since the CC-HIC database does not contain information on patients from outside their ICU episode, only the index spell was considered in the analysis. This was to limit confounding by an unknown exposure to oxygen either following discharge or between ICU episodes. Similarly, ICU mortality for the index spell was chosen as the primary outcome measure, in preference to hospital mortality or other distant outcome measure. The primary cohort was further narrowed to investigate the exposure to hyperoxaemia with each additional day of exposure, from 1 to 14 days (i.e. between 0-1, 0-2, 0-3 and so on to 0-14 days). Each cohort therefore had a window of *potential* exposure to oxygen that was the same between patients, and therefore unaffected by informative censoring from either ICU discharge or death. This approach allows for a fair comparison between patients in the presence of informative censoring from death and discharge.

A substantial proportion of spells had a hyperoxaemia dose of zero, i.e. no observed P_aO_2 above 13.3 kPa. To address this "spike at zero" an additional variable

indicating whether or not there was any exposure to a $P_aO_2 \geq 13.3$ kPa was added to the models [203]. For clarity, these variables are referred to as hyperoxaemia dose (containing continuous dose information) and hyperoxaemia indicator. The role of the hyperoxaemia indicator is to allow a discontinuity in the regression at the zero boundary for hyperoxaemia dose. Both variables should be considered in concert when interpreting the model, as they pertain to the same biological concept.

ICU mortality was modelled as a function of hyperoxaemia dose and hyperoxaemia indicator using multivariable logistic regression. A new model was fitted for each additional day of potential exposure, creating 14 models in total with potential exposure windows ranging from 0-1 day to 0-14 days.

Other predictor variables included: sex (male/female), age at admission (years), weight (kg), prior need for assisted daily living (independent or any level of dependence), primary admission reason (medical/surgical) and the APACHE II score. These variables were chosen on the basis of either salience to the underlying research question, scientific plausibility as treatment confounders, or a known strong association with mortality so as to improve model precision. Additional predictor variables were added as interaction effects to explore potential HTE within the cohort, including mechanical ventilation on each day of the exposure window (yes/no), a prior history of COPD and an acute diagnosis of sepsis for the spell under investigation. These variables were chosen as they are strong candidates for the presence of HTE within the cohort. Continuous variables were entered into the model without categorization. Age, weight, APACHE-II score and hyperoxaemia dose were modelled non-linearly using restricted cubic splines. Two internal knots were placed at the 0.25 and 0.75 quantiles of each distribution, with boundary knots at the value limits.

To regularise P_aO_2—which is measured irregularly when arterial blood gas samples are drawn—linear imputation was performed with a 12 hour window. Where P_aO_2 measures were still unavailable, the exposure was assumed to be zero.

The average treatment effect (ATE) of exposure to hyperoxaemia was calculated by fitting models with each individual's own recorded exposure to hyperox-

aemia, and contrasting this with the counterfactual scenario had this exposure been zero.

Model validation was performed using the non-parametric bootstrap with 100 resamples to provide optimism corrected calibration and discrimination indices. Calibration was evaluated through inspection of high resolution calibration plots and optimism corrected Brier scores. Model discrimination was evaluated using the area under the receiver operator characteristic and precision recall curves.

5.4 Results

Over the four year period of the study, 45,320 episodes were available. After exclusions and refactoring episodes into spells, a primary cohort with a minimum ICU length of stay of 24 hours of 24,348 spells remained. Conditioning on patients remaining alive and inside the ICU up to 14 days provided a cohort that tapers to 2,791 spells by day 14 as shown in the study flow diagram in figure 5.9. Baseline characteristics, stratified by exposure to hyperoxaemia, for the windows of exposure evaluated at days 1 and 14 are shown in table 5.4 (page 181). In total, 952,707 P_aO_2 samples were available for analysis, or 17 [10, 37] samples per spell (median [IQR]).

Exposure to hyperoxaemia was readily identified in the cohort, with a total of 18,968 (77.7%) of patients exposed to hyperoxaemia after 1 day. This increased in proportion to 97.3% of the cohort after 14 days. Figure 5.10 (page 183) shows the distribution of the hyperoxaemia dose variable over the period under investigation. An initial period over the first four days of relatively increased exposure give way to a stable pattern of exposure that follows.

The model coefficients for the hyperoxaemia indicator variable for each exposure window are shown in figure 5.11 (page 183). A consistent association was found between the hyperoxaemia indicator variable and increased ICU mortality. This association generally increased in effect size from day 1 to 11, with a simultaneous reduction in certainty in line with the reduction in cohort size over that time.

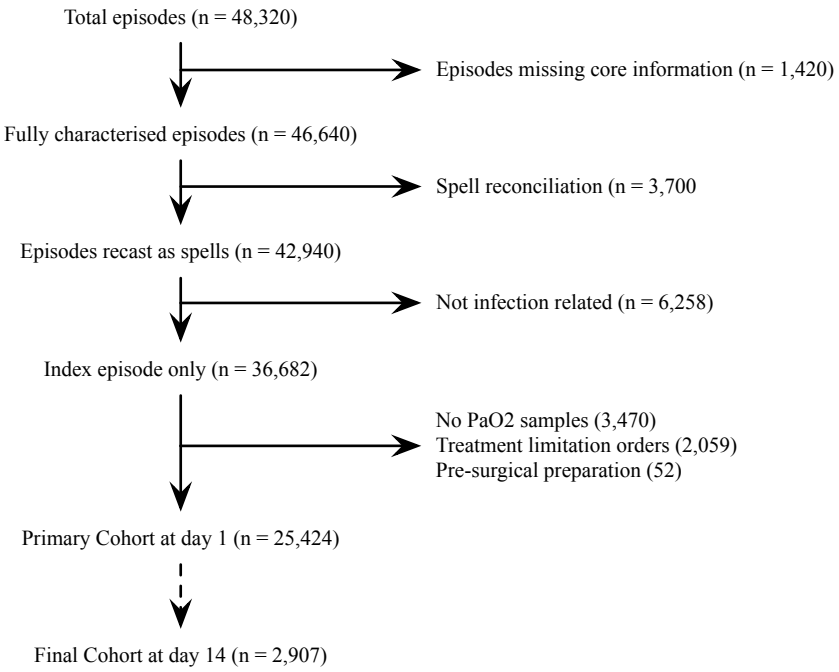


Figure 5.9: Hyperoxaemia study flow diagram.

Hyperoxaemia dose was modelled non-linearly with restricted cubic splines. The coefficients for the spline basis functions do not hold an intuitive meaning, and so partial dependence plots for this variable are presented in figure 5.12 (page 184). The curves displayed are mostly flat, suggesting no dose response effect. There is a small suggestion, particularly in the models beyond day 8, of some protective effect exerted by higher hyperoxaemia levels. However, there is a great deal of uncertainly in the confidence intervals shown, which are entirely consistent with no effect identified. In the models that do show some protective effect, this curvature is driven by a relatively small number of cases that saw very high levels of exposure to P_aO_2 and went on to survive. This would explain the slight downwards slope, but with very large confidence intervals.

There was evidence to support HTE between each of the *a priori* specified subgroups of interest. As would be expected, patients with COPD showed a positive association between exposure to hyperoxaemia and increased mortality. Patients in receipt of ventilation, and those with sepsis both showed a potentially protective association between exposure to hyperoxaemia and increased mortality.

	Exposure window 0-1 day		Exposure window 0-14 days	
Variable	Not Exposed	Exposed	Not Exposed	Exposed
N	5454	18968	76	2729
Age	62.81 (16.67)	60.05 (17.45)	62.13 (16.68)	57.52 (17.22)
Weight	78.51 (21.68)	77.27 (19.05)	73.58 (23.30)	78.82 (20.13)
Sex				
Female	2245 (41.2)	7415 (39.1)	34 (44.7)	956 (35.0)
Male	3208 (58.8)	11553 (60.9)	42 (55.3)	1772 (64.9)
Missing	1 (0.0)	0 (0.0)	0 (0.0)	1 (0.0)
Ethnicity				
White British	3384 (62.0)	11780 (62.1)	49 (64.5)	1703 (62.4)
White Irish	109 (2.0)	258 (1.4)	2 (2.6)	31 (1.1)
White other	392 (7.2)	1235 (6.5)	6 (7.9)	220 (8.1)
Mixed white/black	16 (0.3)	54 (0.3)	0 (0.0)	3 (0.1)
Mixed white/Asian	10 (0.2)	19 (0.1)	0 (0.0)	2 (0.1)
Mixed any other	20 (0.4)	63 (0.3)	0 (0.0)	3 (0.1)
Asian/Asian British	197 (3.3)	842 (6.1)	4 (5.2)	166 (6.2)
Black/Brit. Carribean	142 (2.6)	338 (1.8)	1 (1.3)	48 (1.8)
Black/British African	174 (3.2)	451 (2.4)	1 (1.3)	66 (2.4)
Black/British other	122 (2.2)	281 (1.5)	4 (5.3)	43 (1.6)
Chinese	25 (0.5)	95 (0.5)	0 (0.0)	14 (0.5)
Other ethnic group	189 (3.5)	833 (4.4)	2 (2.6)	165 (6.0)
Not stated	670 (12.3)	2693 (14.2)	7 (9.2)	258 (9.5)
Missing	4 (0.1)	26 (0.1)	0 (0.0)	5 (0.2)
Apache II score	16.60 (6.12)	16.51 (6.96)	18.22 (6.15)	19.82 (7.31)
System				
Respiratory	1824 (33.4)	2830 (14.9)	38 (50.0)	867 (31.8)
Cardiovascular	1131 (20.7)	6628 (34.9)	18 (23.7)	549 (20.1)
Gastrointestinal	804 (14.7)	3349 (17.7)	6 (7.9)	355 (13.0)
Neurological	298 (5.5)	1513 (8.0)	2 (2.6)	308 (11.3)
Genitourinary	522 (9.6)	1653 (8.7)	6 (7.9)	116 (4.3)
EMTP	269 (4.9)	605 (3.2)	3 (3.9)	34 (1.2)
Haem/Immunological	135 (2.5)	208 (1.1)	1 (1.3)	40 (1.5)
Trauma	172 (3.2)	1131 (6.0)	0 (0.0)	338 (12.4)
Other	268 (4.8)	850 (4.4)	1 (1.3)	59 (2.4)
Missing	31 (0.6)	201 (1.1)	1 (1.3)	56 (2.1)
Prior dependency	0.23 (0.42)	0.14 (0.35)	0.33 (0.47)	0.16 (0.37)
Medical	0.70 (0.46)	0.42 (0.49)	0.87 (0.34)	0.73 (0.44)
Surgical classification				
Elective	1004 (18.4)	6314 (33.3)	2 (2.6)	173 (6.3)
Scheduled	125 (2.3)	1684 (8.9)	3 (3.9)	71 (2.6)
Urgent	288 (5.3)	1259 (6.6)	2 (2.6)	110 (4.0)
Emergency	273 (5.0)	1842 (9.7)	4 (5.3)	386 (14.1)
Missing	3764 (69.0)	7869 (41.5)	65 (85.5)	1989 (72.9)
CPR	0.02 (0.15)	0.05 (0.21)	0.04 (0.20)	0.08 (0.27)
Oxygenation				
$\int_0^t P_aO_2 \geq 13.3$kPa	0.00 (0.00)	39.28 (49.76)	0.00 (0.00)	202.05 (258.56)
TW $\int_0^t P_aO_2 \geq 13.3$kPa	0.00 (0.00)	1.64 (2.07)	0.00 (0.00)	0.60 (0.77)
median SpO_2	95.36 (2.61)	97.72 (1.87)	93.80 (3.60)	96.37 (2.01)
Subgroups of interest				
Septic	0.18 (0.38)	0.09 (0.29)	0.30 (0.46)	0.24 (0.43)

	Exposure window 0-1 day		Exposure window 0-14 days	
Variable	Not Exposed	Exposed	Not Exposed	Exposed
COPD	0.13 (0.34)	0.04 (0.19)	0.12 (0.33)	0.05 (0.22)
Ventilated	0.29 (0.45)	0.34 (0.47)	0.38 (0.49)	0.55 (0.50)
Site				
A	3018 (55.3)	9093 (47.9)	45 (59.2)	1018 (37.3)
B	497 (9.1)	2496 (13.2)	5 (6.6)	570 (20.9)
C	385 (7.1)	2163 (11.4)	9 (11.8)	560 (20.5)
D	1112 (20.4)	3149 (16.6)	16 (21.1)	357 (13.1)
E	442 (8.1)	2067 (10.9)	1 (1.3)	224 (8.2)
ICU mortality	249 (4.6)	809 (4.3)	7 (9.2)	204 (7.5)
Spell LOS (days)	7.43 (10.94)	6.98 (10.90)	27.52 (16.21)	29.15 (20.33)

Table 5.4: Patient characteristics for the hyperoxaemia study. Patients have been stratified by exposure to PA $\geq$ 13.3 kPa as the primary treatment effect of interest. Characteristics over the first and last exposure windows are shown (days 0-1 and 0-14). EMTP = Endocrine, Metabolic, Thermoregulation and Poisoning. TW = time weighted. LOS = length of stay.

exposure window	ROC AUC	PR AUC	brier
1 day	0.82 (0.81, 0.83)	0.17 (0.16, 0.19)	0.038
2 days	0.79 (0.78, 0.80)	0.15 (0.13, 0.16)	0.043
3 days	0.77 (0.75, 0.78)	0.14 (0.13, 0.16)	0.047
4 days	0.75 (0.73, 0.76)	0.14 (0.12, 0.15)	0.051
5 days	0.73 (0.72, 0.75)	0.15 (0.13, 0.17)	0.055
6 days	0.73 (0.71, 0.75)	0.16 (0.14, 0.18)	0.059
7 days	0.72 (0.71, 0.74)	0.16 (0.14, 0.18)	0.063
8 days	0.72 (0.7, 0.74)	0.17 (0.14, 0.19)	0.065
9 days	0.73 (0.71, 0.75)	0.18 (0.15, 0.20)	0.067
10 days	0.73 (0.71, 0.75)	0.18 (0.15, 0.21)	0.068
11 days	0.73 (0.7, 0.75)	0.18 (0.15, 0.21)	0.067
12 days	0.74 (0.71, 0.76)	0.19 (0.16, 0.22)	0.067
13 days	0.74 (0.71, 0.76)	0.18 (0.16, 0.22)	0.066
14 days	0.73 (0.71, 0.76)	0.18 (0.15, 0.22)	0.065

Table 5.5: Model performance characteristics for hyperoxaemia models. Confidence intervals show the 95% bootstrapped confidence intervals.

The average treatment effect of hyperoxaemia was estimated by evaluating each patients risk of mortality with their observed oxygen exposure contrasted with an exposure fixed at zero. The resulting sampling distribution of average treatment effect is shown in figure 5.14. The mean value of the ATE distribution for each exposure window is consistently below 0, though there is too much uncertainty in the estimates to draw any conclusions from these findings. The confidence intervals are wide and reflect both great uncertainty in the overall estimates, as well as the potential signal for both benefit and harm from the different subgroups examined.

Overall model discrimination and calibration was good. Optimism-corrected area under the receiver operator and precision recall curves are shown in table 5.6. Illustrative calibration plots, with accompanying risk densities are shown in figure 5.16.

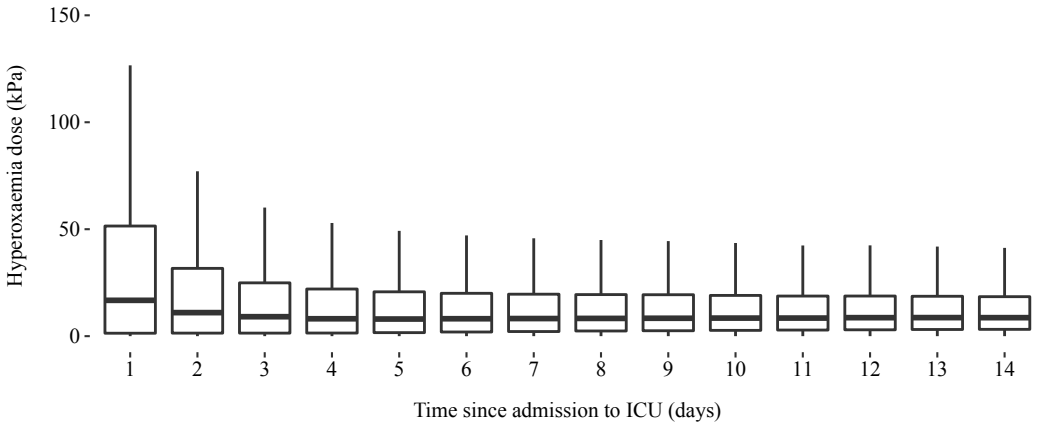


Figure 5.10: Distribution of exposure to hyperoxaemia dose. Exposure is readily seen inside the cohort.

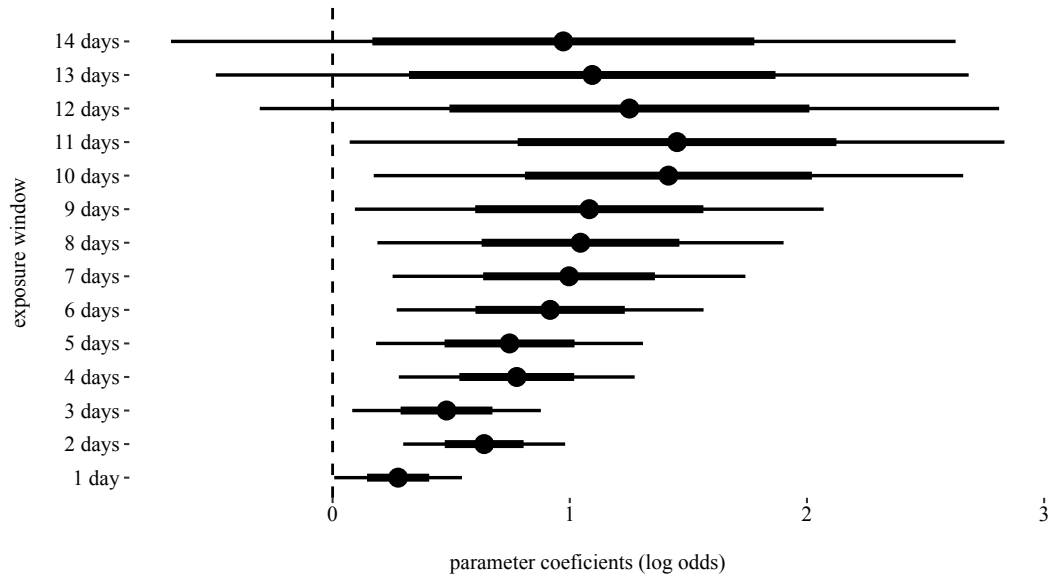


Figure 5.11: Model coefficients for hyperoxaemia indicator. Standard error derrived 68% and 95% confidence intervals are drawn. The vertical dashed line at log odds of zero indicates a null effect.

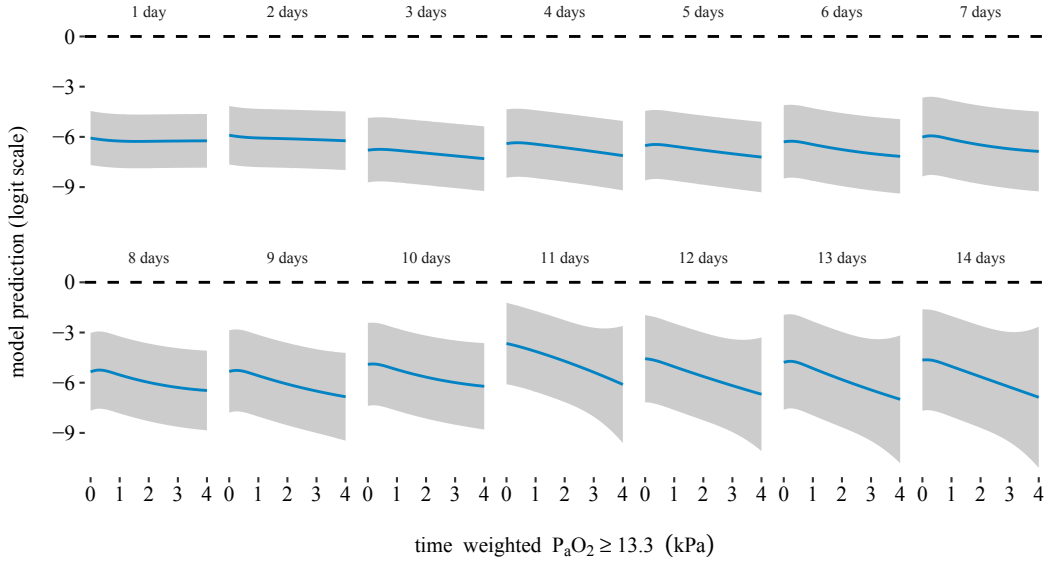


Figure 5.12: Partial dependence plot showing the non-linear effect on outcome as hyperoxaemia dose is modified.

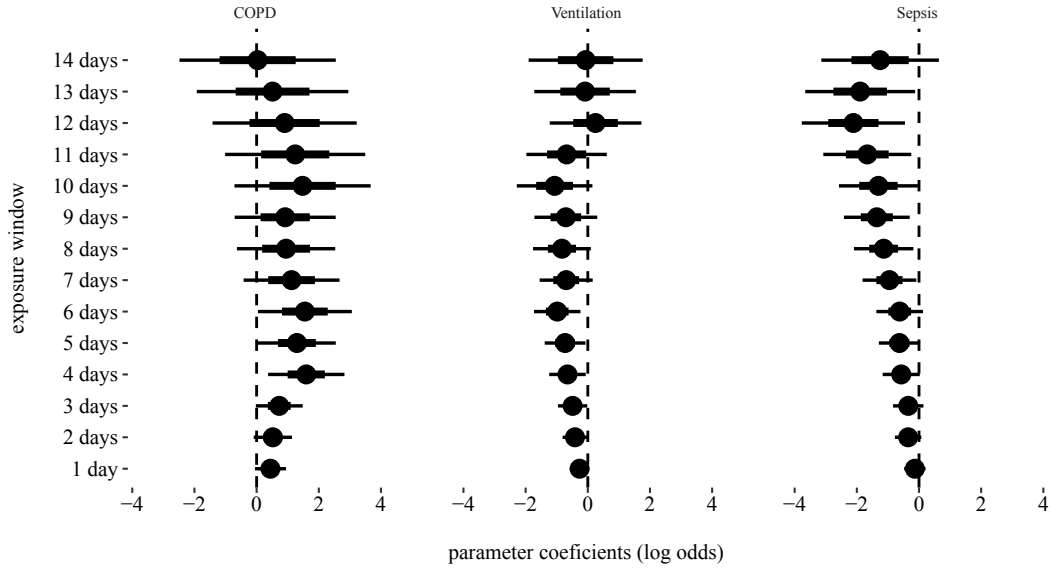


Figure 5.13: Interaction effects for COPD, ventilation status, and sepsis are presented. Standard error derrived 68% and 95% confidence intervals are displayed.

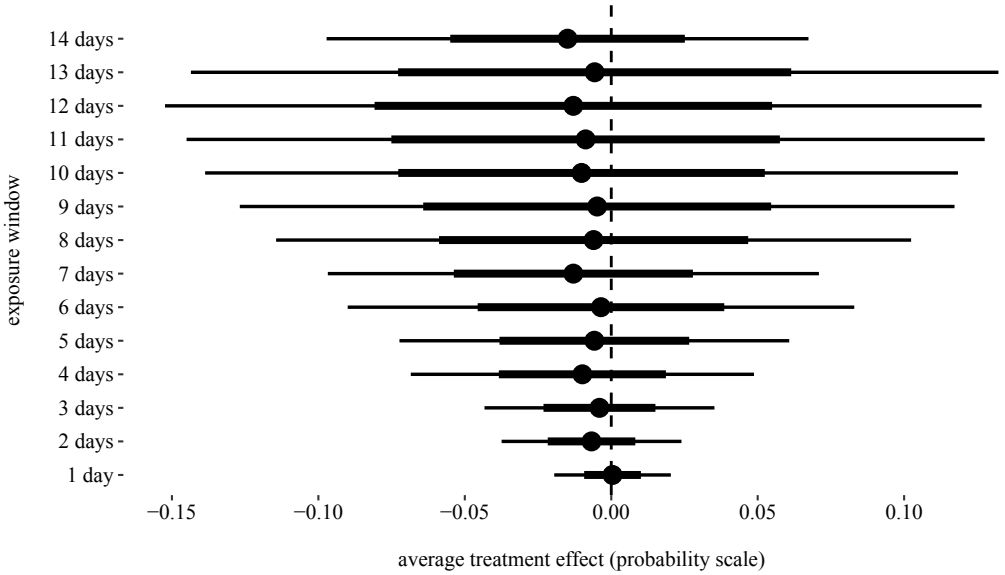


Figure 5.14: Average treatment effects

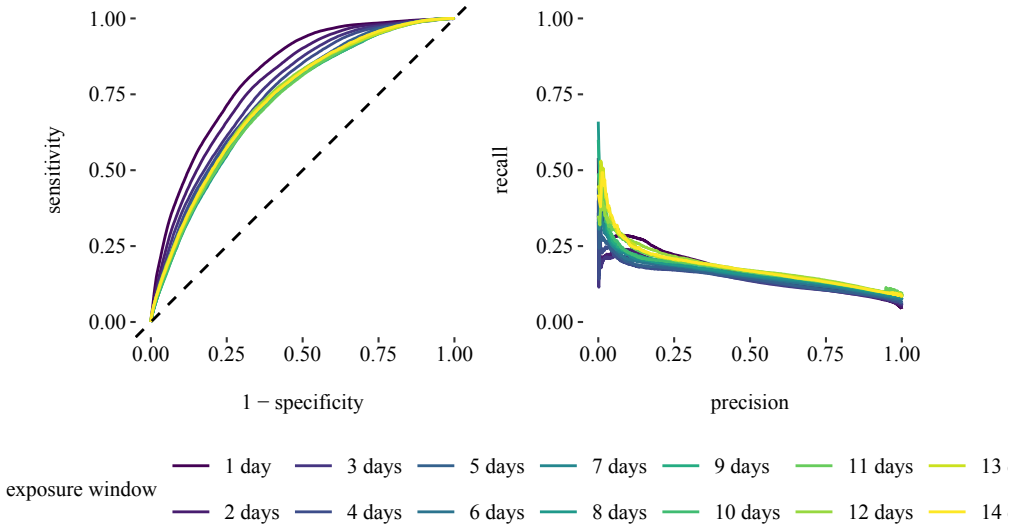


Figure 5.15: Hyperoxaemia model discrimination. **Left panel:** receiver operator characteristic (ROC) curves. **Right panel:** precision recall (PR) curves.

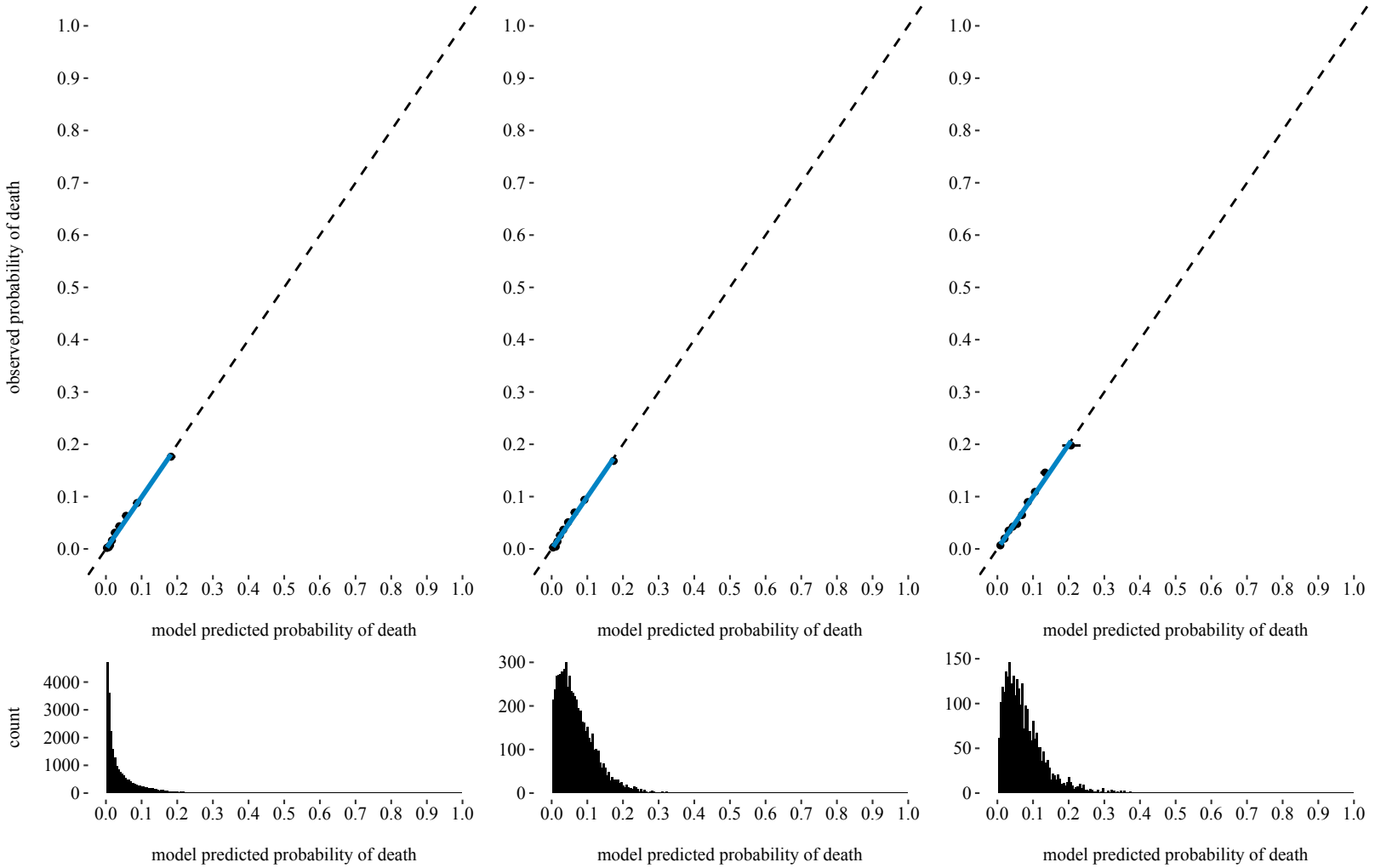


Figure 5.16: Hyperoxaemia models calibration curves for the exposure windows 0-1, 0-7 and 0-14 for the left, middle and right panels respectively. Point estimates and 95% confidence intervals are selected from each decile of risk, with a non-parametric locally estimated scatterplot smoothing (LOESS) fit highlighted in blue.

term	estimate	std.error	statistic	p
1 days				
Hyperoxaemia (indicator)	0.3	0.1	2.0	4.43×10^{-2}
Hyperoxaemia COPD interaction	0.4	0.3	1.7	8.18×10^{-2}
Hyperoxaemia ventilation interaction	−0.3	0.2	−1.7	9.46×10^{-2}
Hyperoxaemia sepsis interaction	−0.1	0.2	−0.8	4.52×10^{-1}
2 days				
Hyperoxaemia (indicator)	0.6	0.2	3.7	2.43×10^{-4}
Hyperoxaemia COPD interaction	0.5	0.3	1.7	9.68×10^{-2}
Hyperoxaemia ventilation interaction	−0.4	0.2	−2.0	4.31×10^{-2}
Hyperoxaemia sepsis interaction	−0.3	0.2	−1.6	1.07×10^{-1}
3 days				
Hyperoxaemia (indicator)	0.5	0.2	2.4	1.79×10^{-2}
Hyperoxaemia COPD interaction	0.7	0.4	1.9	5.81×10^{-2}
Hyperoxaemia ventilation interaction	−0.5	0.2	−2.0	4.20×10^{-2}
Hyperoxaemia sepsis interaction	−0.3	0.2	−1.4	1.66×10^{-1}
4 days				
Hyperoxaemia (indicator)	0.8	0.3	3.1	2.21×10^{-3}
Hyperoxaemia COPD interaction	1.6	0.6	2.5	1.08×10^{-2}
Hyperoxaemia ventilation interaction	−0.7	0.3	−2.2	2.92×10^{-2}
Hyperoxaemia sepsis interaction	−0.6	0.3	−1.9	6.04×10^{-2}
5 days				
Hyperoxaemia (indicator)	0.7	0.3	2.6	9.36×10^{-3}
Hyperoxaemia COPD interaction	1.3	0.6	2.0	4.23×10^{-2}
Hyperoxaemia ventilation interaction	−0.7	0.3	−2.2	2.89×10^{-2}
Hyperoxaemia sepsis interaction	−0.6	0.3	−1.9	6.29×10^{-2}
6 days				
Hyperoxaemia (indicator)	0.9	0.3	2.8	5.45×10^{-3}
Hyperoxaemia COPD interaction	1.6	0.8	2.0	4.41×10^{-2}
Hyperoxaemia ventilation interaction	−1.0	0.4	−2.6	9.95×10^{-3}
Hyperoxaemia sepsis interaction	−0.6	0.4	−1.6	1.05×10^{-1}
7 days				
Hyperoxaemia (indicator)	1.0	0.4	2.6	8.62×10^{-3}
Hyperoxaemia COPD interaction	1.1	0.8	1.4	1.53×10^{-1}
Hyperoxaemia ventilation interaction	−0.7	0.4	−1.6	1.10×10^{-1}
Hyperoxaemia sepsis interaction	−1.0	0.4	−2.2	3.12×10^{-2}
8 days				
Hyperoxaemia (indicator)	1.0	0.4	2.4	1.67×10^{-2}
Hyperoxaemia COPD interaction	0.9	0.8	1.2	2.40×10^{-1}
Hyperoxaemia ventilation interaction	−0.8	0.5	−1.8	7.93×10^{-2}
Hyperoxaemia sepsis interaction	−1.1	0.5	−2.3	1.97×10^{-2}
9 days				

Hyperoxaemia (indicator)	1.1	0.5	2.1	3.18×10^{-2}
Hyperoxaemia COPD interaction	0.9	0.8	1.1	2.67×10^{-1}
Hyperoxaemia ventilation interaction	−0.7	0.5	−1.4	1.71×10^{-1}
Hyperoxaemia sepsis interaction	−1.4	0.5	−2.5	1.20×10^{-2}
10 days				
Hyperoxaemia (indicator)	1.4	0.6	2.2	2.55×10^{-2}
Hyperoxaemia COPD interaction	1.5	1.1	1.3	1.86×10^{-1}
Hyperoxaemia ventilation interaction	−1.1	0.6	−1.7	8.54×10^{-2}
Hyperoxaemia sepsis interaction	−1.3	0.6	−2.0	4.40×10^{-2}
11 days				
Hyperoxaemia (indicator)	1.5	0.7	2.1	3.92×10^{-2}
Hyperoxaemia COPD interaction	1.2	1.1	1.1	2.80×10^{-1}
Hyperoxaemia ventilation interaction	−0.7	0.7	−1.0	3.01×10^{-1}
Hyperoxaemia sepsis interaction	−1.7	0.7	−2.3	2.12×10^{-2}
12 days				
Hyperoxaemia (indicator)	1.3	0.8	1.6	1.16×10^{-1}
Hyperoxaemia COPD interaction	0.9	1.2	0.8	4.46×10^{-1}
Hyperoxaemia ventilation interaction	0.2	0.8	0.3	7.40×10^{-1}
Hyperoxaemia sepsis interaction	−2.1	0.8	−2.5	1.27×10^{-2}
13 days				
Hyperoxaemia (indicator)	1.1	0.8	1.4	1.76×10^{-1}
Hyperoxaemia COPD interaction	0.5	1.2	0.4	6.80×10^{-1}
Hyperoxaemia ventilation interaction	−0.1	0.8	−0.1	9.16×10^{-1}
Hyperoxaemia sepsis interaction	−1.9	0.9	−2.1	3.59×10^{-2}
14 days				
Hyperoxaemia (indicator)	1.0	0.8	1.2	2.49×10^{-1}
Hyperoxaemia COPD interaction	0.0	1.3	0.0	9.80×10^{-1}
Hyperoxaemia ventilation interaction	−0.1	0.9	−0.1	9.41×10^{-1}
Hyperoxaemia sepsis interaction	−1.3	1.0	−1.3	1.95×10^{-1}

Table 5.6: Model coefficients for hyperoxaemia variables.

5.5 Discussion

A small yet consistent association was found in models between exposure to hyperoxaemia (by indicator) and mortality from day 2 to day 11 following admission to the ICU. While the 95% compatibility interval contained the null for models with exposure windows of $\geq$12 days. With regards to hyperoxaemia dose, there was no convincing association between dose response and mortality, with the curve being largely flat across observed hyperoxaemia dose and exposure windows. There was a small downward curvature of the hyperoxaemia dose curve, particularly at longer

exposure windows. This downwards deflection was informed by a small number of outlying patients with very high oxygen exposure levels who survived, which is reflected in the broad confidence intervals around the estimate.

It is important to exercise caution over "statistical significance" particularly in the observational context where sample size is largely a function of convenience, rather than an *a priori* power calculation. Large sample sizes can lead to parameter estimates with p values that fall below the arbitrary 0.05 threshold, regardless of effect size. It is therefore important to consider the Minimum Clinically Important Difference (MCID) for a given exposure. Given the ubiquity of oxygen as an exposure, it would be reasonable to accept a very low MCID, as even very small effect sizes could impart benefit when scaled to a large population.

An important consideration for the interpretation of these findings is the theoretical time-scale and over which exposure to excess oxygen could exert an effect, and therefore the plausibility of finding an association with mortality within such a window. Prior research has shown that in humans the onset of sequelae from exposure to pure oxygen are quite rapid. Evidence of worsening gas exchange, acute lung injury and mortality can be seen as swiftly as 24 hours following continuous exposure to pure oxygen [40, 51, 67]. It remains a more challenging question to answer how this effect may be mediated by exposure to the much lower concentrations of oxygen seen in typical clinical practice. Nevertheless, if exposure to excess oxygen does impart an effect on mortality, then it would be reasonable to expect to see this mediated over the time-scale of this study.

There was evidence to support the presence of HTE from exposure to hyperoxaemia acting in the presence of COPD, ventilation and sepsis. Figure 5.13 demonstrates the relationship between being exposed to a P_aO_2 $\geq$13.3 kPa and each of these patient groups. As would be expected, COPD has a positive association between oxygen exposure and increased mortality. This is an established mechanism and helps provide face validity and framing for other findings. Both ventilation and sepsis were found to have an association between oxygen exposure and improved survival. While this may be a true effect that warrants further investigation, it is

also important to consider whether or not the way in which these models have been conditioned may have produced the effect itself. To illustrate this potential issue, consider the cohort of septic patients. In order for these patients to be able to augment their P_aO_2 $\geq$13.3 kPa, these patients must have a certain degree of respiratory reserve. Severely unwell patients, with profound respiratory insufficiency, would not be able to increment their P_aO_2 to this level. By creating the variable "hyperoxaemia dose", a conditioning on patients who can achieve this state of blood oxygenation has been induced. This may result in those patients with highly performing lungs providing a strong signal for survival, far in excess of any small potential harm from oxygen administration. This type of confounding is well recognised, and in lieu of applying a specific causal methodology, can be difficult to rule out. One potential approach to address this would be to isolate patients who receive patient controlled analgesia (PCA) following surgery, and employ this as an instrumental variable. Oxygen is typically mandated—regardless of clinical need—for patients in receipt of PCA. Further, PCA is sufficiently common so as to provide a large cohort for study. These data are not present in the CC-HIC data model, though it remains a promising avenue to explore in future work.

Placing these finding in the context of existing literature, a small (albeit non-significant) signal of improved outcomes has previously been observed when liberal oxygenation strategies were applied in sepsis in a *post hoc* analysis of the ICU-ROX trial [53]. Challenging this, the HYPER2S study was stopped early due to increased adverse events in the high oxygen group [67]. The HYPER2S used a protocol of 100% F_IO_2, whereas the ICU-ROX trial used "usual care" for the high oxygen group, which would have been much lower. It is hypothesis generating therefore, that in septic patients, there may be a small signal of benefit from using modestly elevated oxygen levels, which may eventually become detrimental at higher levels. Having a mildly elevated oxygen level may also mitigate against ischaemic events, some of which have been observed when implementing conservative oxygen strategies in sepsis. For example, an increased number of mesenteric ischaemia events have been observed with conservative oxygenation strategies in ARDS [202]. These

events may in part contribute to the potentially deleterious effects of conservative oxygen strategies, and so finding the optimal oxygen level in this group remains elusive.

Many prior observational [70, 71, 204, 192] and interventional [59, 60, 68, 52] studies have shown an association between exposure to increased oxygen levels and harm. Eastwood *et al*, as the only exception in this field, could not find supporting evidence of this association [72].

A major limitation of these previous approaches has been the availability of longitudinal oxygenation data. A common approach has been to study the association of a single arterial blood gas result (for example, the "worst" sample in a 24 hour period) on outcome. As previously discussed, it takes many days for the deleterious effects of even pure oxygen to become apparent in humans [40]. And so there is a lack of biological plausibility to support that any single isolated measure of oxygenation could meaningfully alter outcomes.

Only one prior large study has investigated longitudinal oxygenation in the critically ill population [71]. The authors found "a dose-response relationship between supra-physiologic arterial oxygen levels and hospital mortality." This effect was observed in the upper category of oxygen exposure, and a gradient of worsening outcomes across oxygen exposure—which is a requirement to define dose-dependency—was not demonstrated. The most directly similar measure in their study to the approach presented in this thesis was a 96 hour area under the curve for P_aO_2. There was an association between this metric and increased hospital (but not intensive care) mortality. These findings are challenging to reconcile, and with these caveats, a convincing dose-response relationship was not identified.

A small study by Ruggiu *et al* [204] used a P_aO_2 >13.3 kPa to indicate hyperoxaemia, and so is more directly comparable to the methods applied here. They modelled mortality with survival analysis and arrived at the conclusion that a "dose-independent" exposure to hyperoxaemia was associated with harm.

It is possible that the unintended consequence of creating an unambiguous definition of oxygen excess ($P_aO_2 \geq$13.3 kPa) is that associations identified could rep-

resent artefacts of conditioning. This has already been discussed with regard to the potentially protective effect observed in sepsis. Counter to this, patient groups who have a higher mortality and a known increased opportunity to be exposed to high oxygen levels would be highlighted in the model. For example, patients who undergo transfers and procedures. These patients are inherently less stable, experience higher mortality [205] and morbidity [206] and may be placed on a high inspired oxygen concentration regardless of clinical need. Such scenarios would be conditioned by the model, with the likely result of a relationship being created between hyperoxaemia and mortality.

5.6 Limitations

The methodological approach taken in this analysis is undoubtedly *ad hoc* and suboptimal. An alternative approach would have been to use survival analysis, either by extending Cox's proportional hazards model to include hyperoxaemia dose as a time varying variable or to apply joint models with hyperoxaemia dose represented in the longitudinal submodel. With regards to the extended Cox model, there was concern that since the P_aO_2 samples were endogenous in nature (i.e. produced internally by the patient) then the model would not be valid [145]. The joint model would potentially address this concern, however there was a very high variability in serial P_aO_2 samples when viewed over time (i.e. low autocorrelation). There was a lack of confidence that this volatility of serial P_aO_2 samples could be appropriately captured by the longitudinal submodel in a joint model formulation. On reflection, it may have been prudent to apply these methods regardless of the concerns raised, since the underlying models described are more naturally suited to address the features of the research question in a parsimonious manner.

Missing data were problematic when devising the study cohort. Several strategies were implemented in attempt to reclaim missing or erroneous data, most notably linear imputation for P_aO_2 samples and the logistic regression model based recovery of anatomical labels. It is unclear what effect (if any) these approaches would have on final inferences. A multiple imputation approach could have been

advantageous in these situations, to better propagate uncertainty into the final models.

This analysis was conducted as a two-stage approach to the modelling of longitudinal data. In this approach, a longitudinal process, such as serial P_aO_2 samples, are collapsed into a single measure to be included within the model, so as to not violate the independence assumption mandated by the model. While a common approach, there is necessarily a loss of statistical information, and so this approach is only able to address questions related to the cumulative exposure to hyperoxaemia, rather than any other specific morphology (as previously outlined in subsection 2.4.1 on page 49). Under this approach, exposure to high levels of excess oxygen for a short period of time are thought of equally to low levels of excess oxygen for a long period of time.

There is likely to be a significant exposure to oxygen prior to ICU admission. Patients enter the ICU after a non-ignorable amount of time in either an operating theatre, emergency department or ward. It is reasonable to assume that most have had a prior exposure to oxygen, and so much of the potential exposure to oxygen is censored from the CC-HIC database. Indeed, even if normoxaemia is achieved after admission to ICU, a brief period of hyperoxaemia in the emergency department has been suggested to be detrimental [207].

A $P_aO_2 \geq 13.3$ kPa likely captures a surrogate of the mechanism that is causing harm (high F_IO_2). Much of the preclinical data favours high F_IO_2 as being causative for lung parenchymal damage [50]. However, there may be other unrecognised systemic and cellular effects acting other than in the lungs that result directly from a raised P_aO_2.

The cohort is notable for being low in overall risk, as seen in the calibration plots in figure 5.16. Despite being a relatively large cohort in the context of this research subject, the power of a logistic regression is related to the number of events observed in the smallest outcome group. A relatively small number of observed deaths, coupled with the anticipated small effect size of oxygen excess on mortality, increases the likelihood that this study was underpowered.

5.7 Conclusions

This study overcame a number of methodological and technical barriers including reduced data availability, and a lack of data harmonisation in key concepts. Novel solutions were implemented to recover sufficient data to conduct the primary research question. The study suggests that continued exposure to $P_aO_2 \geq 13.3$ kPa may be harmful, although there was a high degree of uncertainty in model estimates. Given the absent dose-response relationship, it would be prudent not to give these findings a causal interpretation. There was evidence to support the existence of HTE in all subgroups of interest, including COPD, sepsis and ventilation.

As a necessarily longitudinal treatment effect, studies that seek to use individual measures of oxygenation to extrapolate on outcomes should be avoided in future research. In the observational context, research efforts should be directed towards data resources that cover the whole hospitalisation period, so as to remove the limitation of left censored data that is likely a feature of this study. Should exposure to hyperoxaemia increase the risk of mortality, it is unclear over what time frame following exposure this risk returns to baseline. This is an area that remains largely unexplored, and will not likely be addressed by ongoing randomised studies. Further experimental investigation into this controversial field is thus warranted.

Chapter 6

Physiological Morphologies in Sepsis

In this chapter I will explore the relationship between the morphology of longitudinal physiology and survival in sepsis. Particular attention will be paid to informatively missing data patterns—which are endemic to critical care cohorts—through the application of joint models. Septic patients within the CC-HIC data resource will be used as the primary cohort for investigation. Details of the literature review and search strategy for this Chapter can be found in Appendix Section B (page 271).

6.1 Background

As described in Section 2.6, sepsis is a highly heterogeneous disease. Potential areas of heterogeneity that have yet to be explored with formal methodological approaches are the varied morphologies that can be demonstrated in longitudinal physiology. These morphologies are outlined in subsection 2.4.1 (page 49) and include:

- severity: the value of a biomarker at a point in time.
- velocity: the rate of change of a biomarker with respect to time.
- trajectory: the combination of severity and velocity sufficient to describe the path taken by the biomarker.
- cumulative exposure: the aggregate impact of the biomarker over time, typically defined by the area under the curve of the biomarker.

Patient physiology during sepsis is dynamic; while lacking a reference to disease time zero (which is generally unknown), it is nonetheless of interest to investigate

how these different morphologies relate to outcomes. This may reveal insights into the disease process, in lieu of a marker that reliably tracks the "stage" of sepsis. Due to the endemic presence of death in the critical care population, estimates that pertain to longitudinal biomarkers are subject to informative missingness. This has implications where longitudinal biomarkers are used as the target of inference, as in this instance. This study focuses on the longitudinal morphologies of the SOFA score, its individual constituent parts, and the serum level of C-reactive protein. C-reactive protein is included as it is routinely measured in critical care as a marker of inflammation. Inflammation is a driving mechanism of the pathophysiology of sepsis, and is not captured directly within the SOFA score. SOFA is itself well suited to the exploration of longitudinal morphologies in sepsis. Not in the least because SOFA was specifically developed to assist in the day-to-day tracking of organ dysfunction in sepsis [95]. The sepsis-III definition itself requires an *increase* in the SOFA score of at least two points [93]. We can consider that the diagnosis of sepsis *requires* a deterioration in organ function. The current definition of sepsis therefore already seeks to define sepsis in terms of a specific longitudinal morphology; disease velocity.

Prior research has investigated some specific morphologies of SOFA. Most commonly this involves modelling the change in SOFA (ΔSOFA) over a defined time frame. Common examples of ΔSOFA include Δ_{96}SOFA; the change in SOFA score between arrival to ICU and at 96 hours, and Δ_{max}SOFA; the change in SOFA score between arrival and the maximum observed SOFA score. This restricts the evaluation of the velocity of SOFA to an average taken between two points in time. Patients may improve, stabilise and deteriorate all within this time frame. If we wish to explore the morphologies of SOFA throughout the ICU episode, then it is necessary to consider the SOFA score as a continuous biomarker that we regularly sample. Rarely have steps been taken in this field to address informative censoring of patient data, and this has—in my view—led to some conflicting findings, particularly with respect to ΔSOFA.

Until recently, methodological constraints have been a limiting factor in de-

veloping these ideas more formally in the applied context. The development of a rich ecosystem of research software to support the application of joint models has permitted greater exploration of this topic.

6.2 Hypothesis Statement

Three joint models for each biomarker will be fit to explore the different physiological morphologies in sepsis: severity, trajectory, and cumulative effect. Building on prior knowledge in this area, evidence will be sought against the following testable hypotheses:

1. the trajectory parametrisation will yield a superior model fit over other parametrisations.
2. predictive performance of the models will reduce over time as patients enter a chronic phase of disease and acute physiological changes exert less of an influence on outcome.

It is also of interest to understand how the cumulative effects parametrisation relates to outcomes. With no prior research into this particular morphology, it is only possible to speculate on the associations that may be found.

6.3 Methods

6.3.1 Data Preparation

Data were pre-processed using inspectEHR [6] to apply standard data quality evaluation rules as outlined previously. Invalid episodes were removed, while those that remained were reconciled into spells. When aligning episodes into spells, baseline data pertaining to the start of the spell were captured from the primary episode (for example, co-morbidities), whereas data pertaining to the end of the spell were captured from the final episode (for example, patient outcome). In cases of data missingness, information was retrieved from whichever episode it was available from, provided the data concept was unlikely to change between episodes. For example, date of birth could be retrieved from any episode, whereas the APACHE II

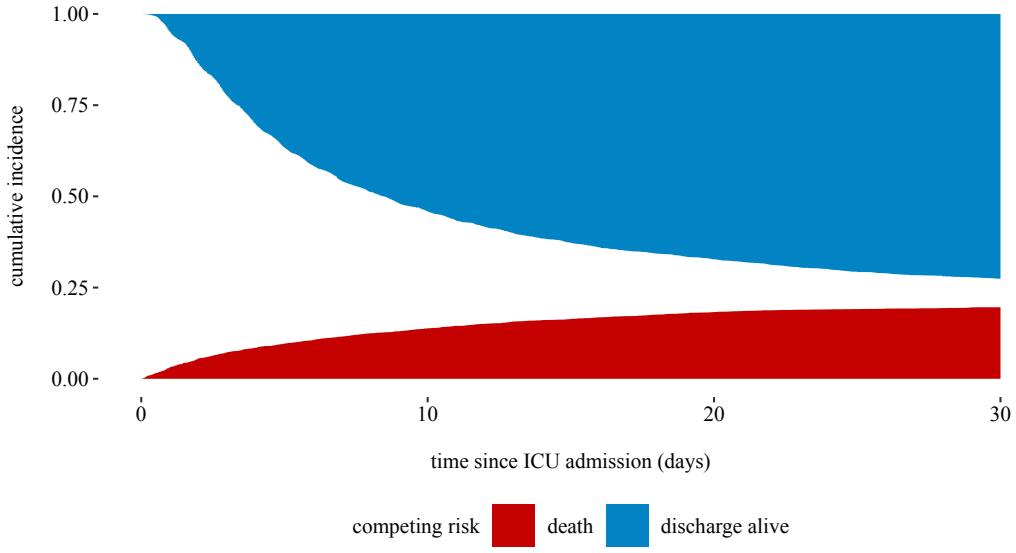


Figure 6.1: Cumulative incidence of the competing risks: discharge alive from the ICU (blue area) and death (red area). Cumulative incidence estimates were obtained using the cmprsk package [208].

score (which would be re-evaluated for each new episode) could not. Longitudinal data are often captured in duplicate across episodes that refer to the same spell. These duplicate data were removed, with data re-allocated to the correct episode (and hence spell) from which they originated.

The timing for spells was re-sequenced, so that the date-time information for all data concepts was referenced to the number of hours since ICU admission in the primary episode. Length of stay was capped at 30 days and all longitudinal results and outcome events beyond this point were censored. The outcome for patients who remained inside the ICU and alive beyond 30 days were marked as "survivors", regardless of their outcome at a later time. These patients are hence administratively censored. Of the 4,188 patients included in the study, 327 (7.8%) had lengths of stay greater than 30 days. Of these, 58 were non-survivors and 269 were survivors. The distribution of outcome times are shown for all patients in figure 6.1. This decision to limit the study to 30 days was informed by the need to balance a representative cohort, with the excessive computational burden of modelling the extremely long distributional tail of the relatively few patients who remained beyond 30 days. Regardless, models fitted against data beyond 30 days would be increasingly reliant

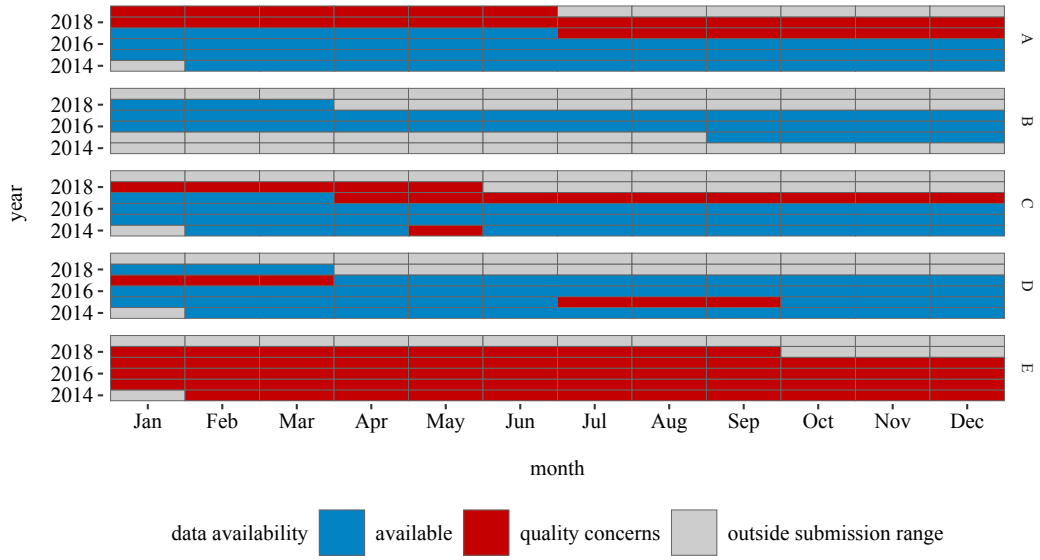


Figure 6.2: Overall data availability for the Sepsis study. Calendar months containing reliable data from each site are highlighted in blue. Regions where the overall data quality did not meet a minimum standard is highlighted in red. Regions for which no data have been received are shaded grey.

on a very small number of patients, with potentially limited generalisability.

The following data concepts were used to evaluate "safe" portions of the database that can be used for data extraction:

- invasive mean arterial pressure.
- lactate.
- noradrenaline.
- P_aO_2.
- F_IO_2.
- urine output.
- creatinine.
- bilirubin.
- platelet count.
- total Glasgow Coma Score (GCS).

I make reference to these data concepts as "bellwether" concepts, since their presence should be ubiquitous in a septic cohort. A high degree of data quality issues or data missingness arising in bellwether concepts raises concerns over the

broader reliability of these data. Months were flagged as "unsafe" and removed from analysis if they contained more than ten calendar days with an error free contribution of data that dropped below two standard deviations of the long running average for all bellwether concepts. This identified corresponding "safe" portions which are highlighted in figure 6.2 (page 199). Site "E" was unable to contribute to this study as there were no months of data contribution meeting these requirements. Sites "A" and "C" had their contribution to the study terminated early, mostly attributed to a drop off in reliable infusion data.

Standardised data extraction from the CC-HIC database was performed with wranglEHR [5] setting a temporal cadence of one hour. If a data concept is contributed more than once in an hour, multiple events were collapsed into a single event using the following summary functions:

- mean: for all continuous data except urine output.
- mode: for categorical data.
- sum: for urine output (to represent the total urine output in that hour).

All data concepts were visually inspected, and transformations made (typically at site level) where appropriate to align to common units. Examples include:

- F_IO_2 was transformed to be represented as a fraction.
- P_aO_2/F_IO_2 ratio was transformed to be represented in kiloPascals (kPa).
- urine output was transformed to be represented as millilitres (mL) and with the same sign (positive).

Multiple data concepts referring to the same semantic concept were coalesced into a single measure with a pre-specified order of precedence. For example, non-invasive and invasive blood pressure are both recorded simultaneously in the CC-HIC database. While there will invariably be some disagreement between these two readings for biological and technical reasons, the majority of the data describes a close relationship between these different measures of the same underlying biological phenomenon.

If outlying data fell outside a predefined range of plausible values it was entered as the limit value for that data concept. For example, the limit value for SpO_2

is 100%, and so a value of 102% would be capped at 100%. The limit values applied were obtained from either the ICNARC data specification, a case literature search or, in lieu of published evidence, clinical judgement. These limit values are listed alongside the primary data specification in appendix table A.1. In cases where the outlier was attributable to a misapplication of the CC-HIC data model, efforts were made to reconcile the outlier if reasonably achieved. Common examples have previously been discussed when reviewing the CC-HIC data model and include:

- categorical data incorrectly represented.
- continuous data contributed in the wrong units.
- data that deviate from an established pattern (e.g. ICNARC codes).

Resolving Outcomes

Following the spell reconciliation process, a small number of records continued to show some outcome variables that were logically inconsistent. In most cases this was caused by earlier ICU admissions for a patient being tagged with the outcome data from subsequent admissions. If patients die on a future re-admission, this results in labelling earlier admissions with their ultimate outcome. Table 6.1 shows occurrences of this inconsistency. Patient outcome is stratified by whether or not death timing data are present. The off axis diagonal elements of the table indicating records that have mismatching outcome data.

	Survivor	Non-survivor
Death date absent	3189	5
Death date present	128	874

Table 6.1: Outcome data stratified by death timing data. A perfect cohort would have no conflicting resulting in the off axis diagonal.

Of the 128 cases with a death date present and a conflicting outcome status of "survivor", only four cases have the death time represented as occurring within the time boundaries of the episode. In all four instances, this occurred because no time of death was present[1]. The time defaulted to midnight, artefactually bringing the

[1] dates and times are stored separately in the CC-HIC data model, and so it is entirely possible to have one without the other.

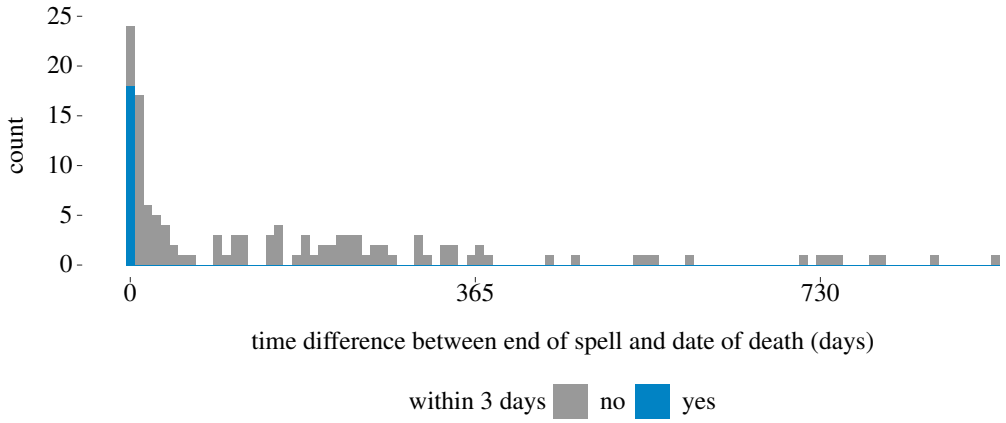


Figure 6.3: Histogram displaying the difference in hours between the declared end of a spell, and the associated time of death. A large number of timings occur within 24 hours of each other, which would be expected from the typical administrative processes governing end of life declarations. A long tail exists as spells have been unintentionally tagged with outcome data from future admissions.

time of death forward to occur within the episode. The outcomes for these episodes were reconciled by replacing the death time with the time that the episode was closed. It was assumed that the other 124 cases in this category were tagged with outcome data from a future admission, and so this information was disregarded. The five cases that are documented as non-survivors, but with missing date information for their outcome, have this value replaced by the documented time of the episode ending. Figure 6.3 shows the difference in time between the episode end time and the documented death time. In a considerable proportion of cases there is a discrepancy between the episode end and the time of death, with the latter occurring in the distant future (relative to the episode). This suggests that these recorded deaths are related to events that occur outside the ICU episode and, for the purposes of this study, were ignored.

6.3.2 Identification of Sepsis

Sepsis was identified using a modification of the approach used by the ICNARC case mix program [97]. The presence of any underlying infection in the CC-HIC primary or secondary diagnostic codes, was identified using the ICNARC coding

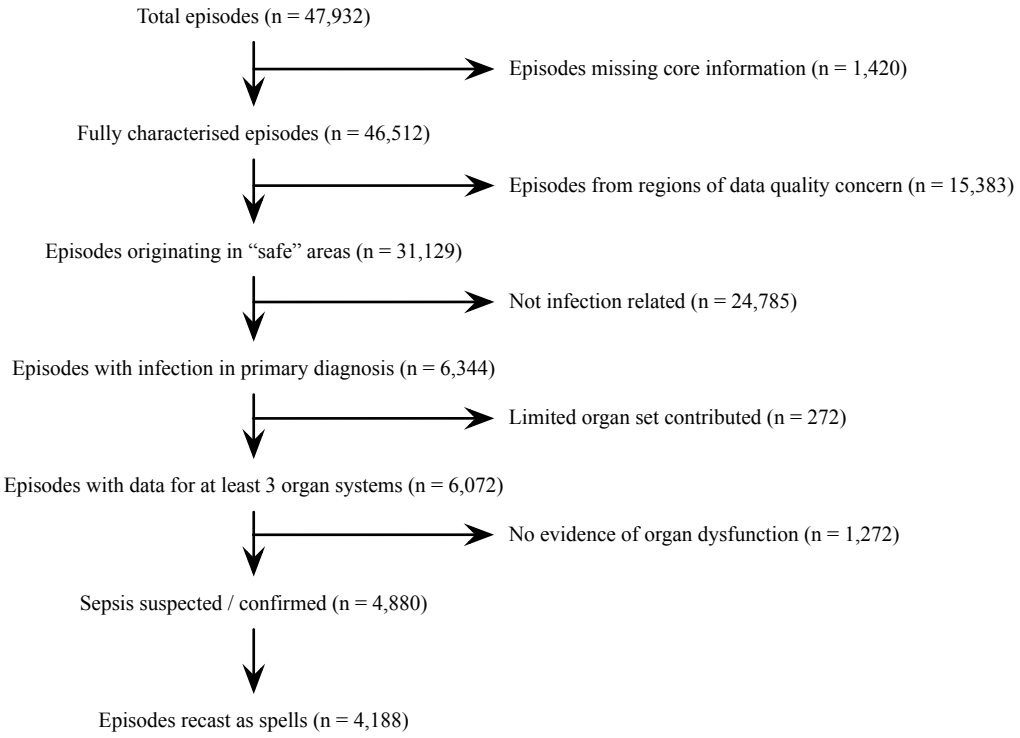


Figure 6.4: Sepsis study flow diagram

method[2]. Preliminary data were extracted from the CC-HIC research database for the first 24 hours of each episode. It was assumed that prior to ICU admission, the patient SOFA score is zero. Evidence was sought for a maximum SOFA score of two or more within the first 24 hours following ICU admission. Patients who had previous evidence of fulminant chronic liver or dialysis dependent renal dysfunction had these respective components of SOFA excluded. Most prior data quality checks are aimed at the site level. As an additional patient level data quality check, only episodes that were able to contribute at least three organs to the SOFA score were considered. Urine output was not used in the evaluation of SOFA as it had been contributed in distinct ways by each site in CC-HIC and it was not possible to properly reconcile these values. The renal component therefore relied upon creatinine only. A septic shock cohort was further identified from those who had a persistent lactate >2 mmol/L, or the presence of ongoing vasopressor use during the initial 24 hour window. A study flow diagram is shown in figure 6.4.

[2]INCARC code ##.##.##.27.## (27 at the 4th level of hierarchy) [98]

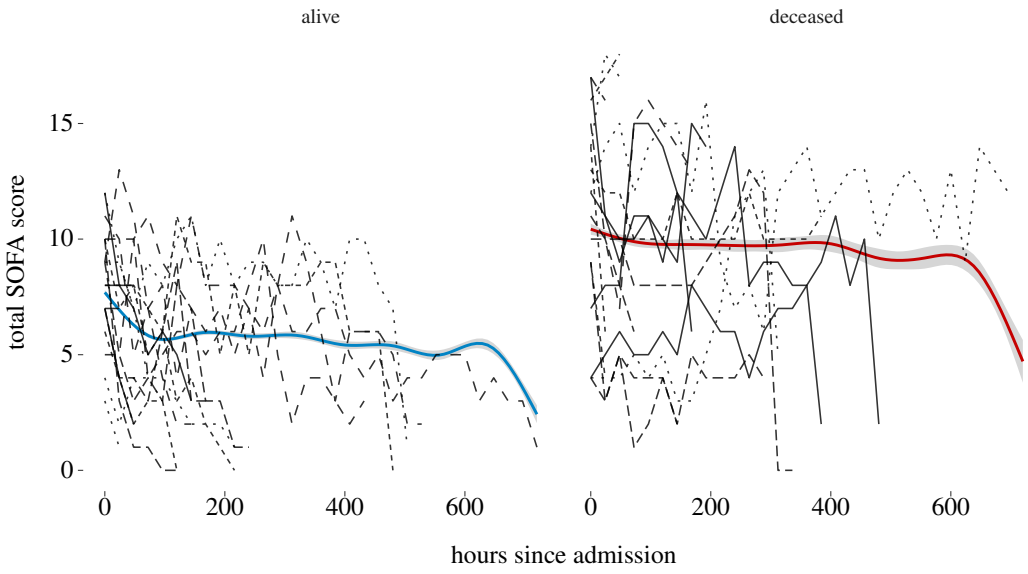


Figure 6.5: The naive marginal trajectory for maximum daily SOFA score is shown for survivors (left, blue) and non-survivors (right, red). A generalised additive model is applied against the whole sepsis cohort, without consideration of censoring from death. A sample of patients are illustrated to highlight some individual SOFA evolutions that are observed. A sudden artefactual decrease is seen in the final 24 hours of observation for the cohort due to the partial contribution of SOFA score componenents in this period.

6.3.3 Calculation of SOFA Score

The SOFA score is shown in table 2.4 (page 59). Minor modifications were necessary to implement SOFA against routinely collected data. The default behaviour of the SOFA score is to assume a zero score for an organ should that organ system not be observed. Since only partial measurements of the total SOFA score were possible in the final 24 hours of a spell, this approach would result in an artefactual *decrease* (i.e. an improvement) in SOFA score shortly before death or discharge. To correct for this, following the last completely observed 24 hour period, any remaining data were entirely disregarded and no SOFA score was provided. This artefactual decrease can be seen in the naive marginal trajectories for the whole population SOFA score shown in figure 6.5.

When there was a higher-than-expected quantity of missing data within the cohort, the score was not returned for that day if a particular data concept was missing. The typical protocol for SOFA is to return a zero if no information is

available. This is based on the assumption that if an organ system is not measured, it is likely that the clinician views the organ system as functioning normally and so does not require investigation or intervention. By making explicit assumptions about the whole evolution of the longitudinal biomarker, the mixed effects model can account for some intermittent missingness. This would likely return less biased estimates for the biomarker, than if the input data to the SOFA score were incorrect.

6.3.4 Data Missingness

Figure 6.6 (page 206) shows the missing data proportions at spell level of all longitudinal data used in the present study.

Drug infusions in CC-HIC are typically described on the basis of the times at which a continuous infusion rate is changed.[3] When data are extracted on a regular time cadence "missing data" for a particular hour represents instances where infusions were either not running, or rate changes had not occurred. In this situation, a last one carried forward (LOCF) procedure is appropriate, since this procedure mirrors the clinical actions that are taking place[4]. An univariate LOCF procedure was applied to vasoactive drug data, filling gaps up to a maximum of six hours. With longer gaps, it was assumed that the infusions were switched off, and so no imputation was performed. "Zeros" were not seen from any of the sites for drug infusions, which likely reflects documentation practice whereby infusions are not documented as zero when they are switched off. As such, infusions were manually "zeroed" either after the last recorded entry, or when the imputation window was greater than six hours. This procedure was inspired by the concept of "persistence windows" from the field of comparative effectiveness research, whereby a sustained period of missing data is used as a signal to decide that a particular drug regimen has been stopped [209].

Similar steps were undertaken as for the hyperoxaemia study to reclaim additional data, including a model based approach to re-labelling blood gas data.

[3] unlike conventional drug investigations, vasoactive drugs in ICU are typically administered via a continuous infusion into a central vein. The half life of these drugs is measured in minutes, and so the rate of infusion (rather than the absolute dose) is the metric of clinical importance.

[4] LOCF are seldom appropriate methods for the imputation of biological data. Drug infusions on ICU are not however biological in nature, and do not display a continuous change in dosing patterns.

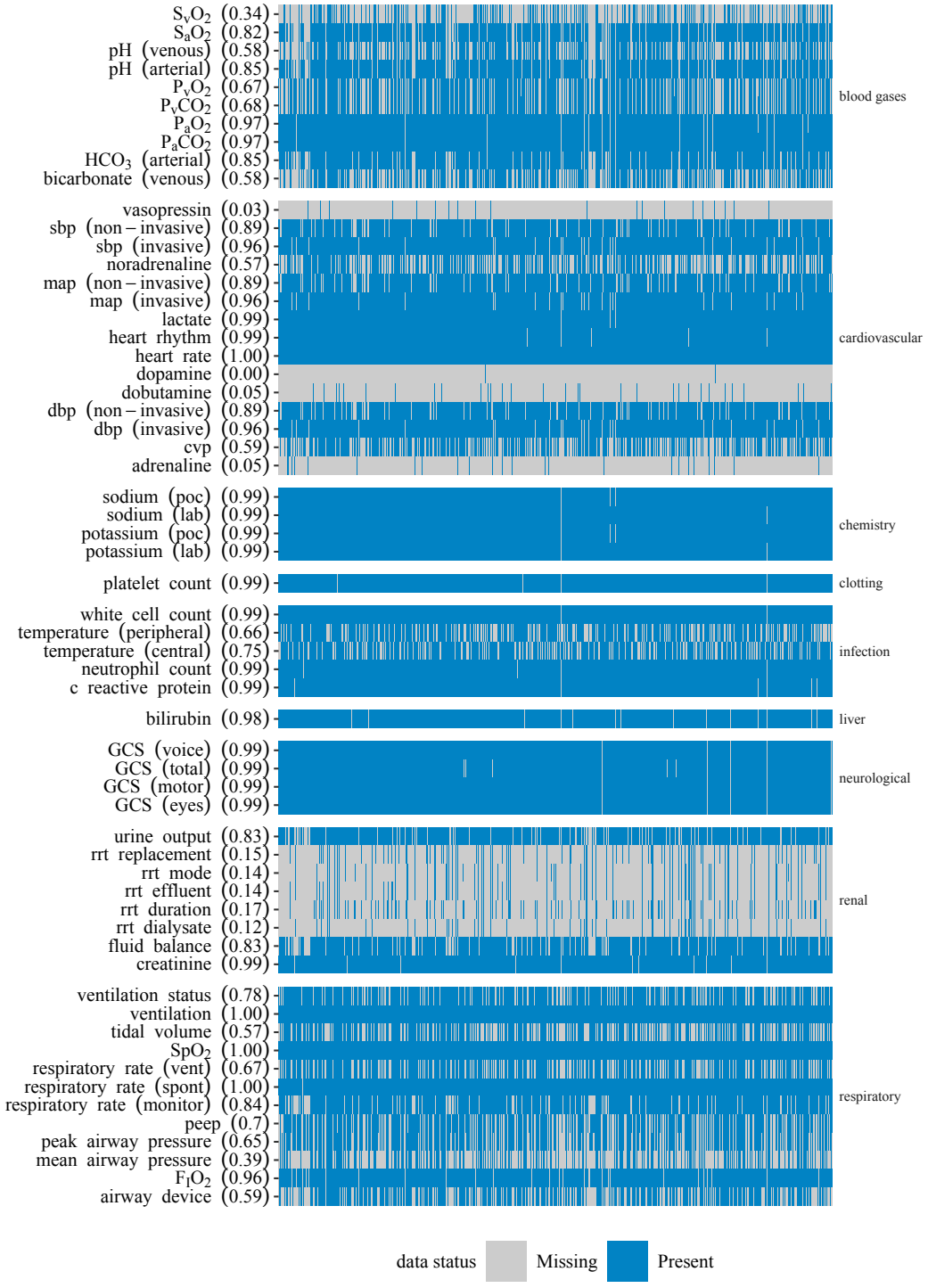


Figure 6.6: Spell level missing data patterns. Variables are shown as grey when no such data exists for the entire spell. Variables are shown as blue when there is at least one result for the spell. Aligned vertical lines across biomarkers indicates data that are missing (or present) for the same patient. There are numerous such alignments, suggesting that missing data are non-random as they are dependent upon the patient. The percentage of complete data are provided in parenthesis on the x-axis labels. Some missing data will invariable exist because the patient did not receive a therapy, for example renal replacement.

6.3.5 Model Fitting

In order to elucidate the relationships between the morphologies of longitudinal biomarkers and outcome, a number of univariate joint models were fitted under maximum likelihood. The following biomarkers were used:

- biomarkers of inflammation:
  - C-reactive protein (CRP).
- biomarkers of individual level organ dysfunction:
  - bilirubin (liver).
  - creatinine (renal).
  - P_aO_2/F_IO_2 ratio (respiratory).
  - maximum daily noradrenaline equivalents (cardiovascular).
  - total GCS (neurological).
  - platelets (clotting).
- summary measures of global organ dysfunction:
  - maximum daily SOFA score.

For all models, the baseline variables of Cox's sub-model included:

- age (unit variance transformed).
- weight (unit variance transformed).
- sex.
- major comorbidities (any vs none)[5].
- prior dependency (any vs none).
- cardiopulmonary resuscitation (CPR) prior to admission to the ICU (yes vs no).
- first 24 hours maximum SOFA score.

These baseline variables were chosen because they either have a known relationship with mortality, or can be used as surrogates for acute severity of illness, co-morbid

[5]Major comorbidities were defined as any of: cirrhosis, recent chemotherapy, chronic lymphocytic leukemia, dialysis dependency, congenital immunosuppression, hepatic encephalopathy, acquired immune deficiency syndrome, home ventilation, lymphoma, metastatic cancer, portal hypertension, radiotherapy, severe respiratory disease, steroid use, or severe cardiovascular disease.

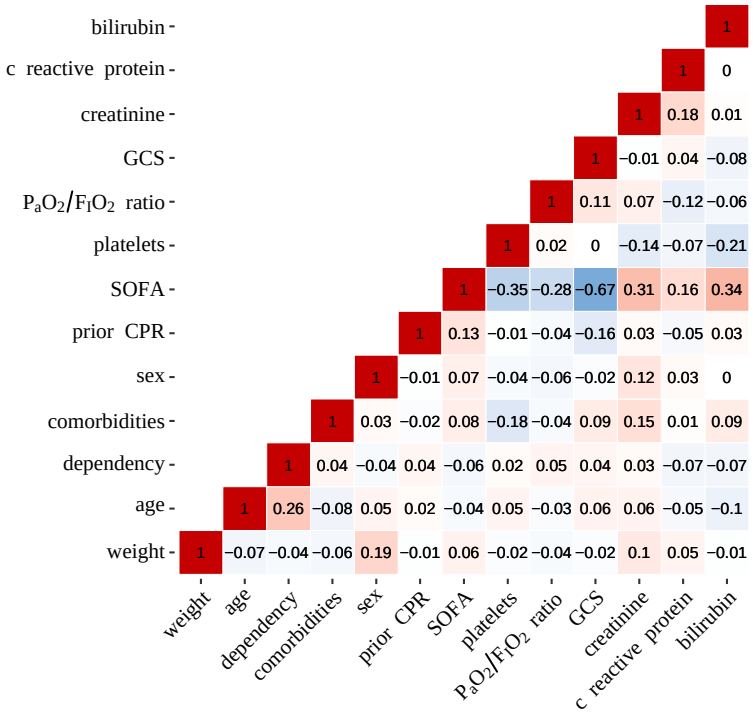


Figure 6.7: Correlogram of core components used in joint models. There is some degree of correlation between SOFA score and its individual components as would be expected.

disease burden, or clinical frailty. Significant collinearity in the baseline variables was examined prior to model inclusion as shown in figure 6.7. SOFA does show a degree of collinearity with its constituent components, as would be expected. It was worthwhile including the SOFA score as a baseline variable in the non SOFA models, since there is no other variable contributing to these models that represents an overall notion of acute disease severity. Baseline SOFA score itself is omitted from the SOFA models because the longitudinal submodel makes an implicit assumption about a patients SOFA score at time zero, without having to add SOFA score to the Cox sub model.

The linear mixed effects sub-model were fitted to each of the listed biomarkers flexibly using natural cubic splines. Fixed effects were fitted using a third degree natural spline[6] basis with boundary knots at days 0 and 30. Internal knots were placed asymmetrically at days 7.5 and 15 to accommodate more flexibility toward

[6]otherwise known as restricted cubic splines

Biomarker	System	Transformation
C-Reactive Protein	Inflammatory	Square root
Bilirubin	Hepatobiliary	Square root
P_aO_2/F_IO_2	Respiratory	Unit variance
Platelets	Clotting	Square root
Noradrenaline equivalents	Cardiovascular	-
GCS	Neurological	-
SOFA Score	Global	-

Table 6.2: transformations for biomarkers in univariate joint models.

the first half of the model where more events are occurring, and so we might reasonably expect more volatility in the biomarker. The random effects were fitted with splines of the same specification, allowing the accurate capture of individual non-linear trajectories without specifying an *a priori* functional form. Since the association structure is of primary inferential importance to the study, it is important that these potentially non-linear and individual patterns are captured correctly prior to any simplifying assumptions that may necessarily follow.

Some transformations of biomarkers (detailed in table 6.2) were necessary to facilitate model convergence. In most instances a square root transformation was applied in order to stabilise the variance of the biomarker, which are often strictly positive values with a positive skew.

Although creatinine was used to calculate the total SOFA score, it was not used to model the renal organ system in an univariate analysis. Without the accompanying renal replacement therapy information, the creatinine value can give a false impression of renal organ failure in the ICU.

Blood pressure is maintained between fairly tight limits in intensive care, and so is unlikely to provide useful variation to understand cardiovascular dysfunction in an ICU context. This is the primary motivation for the presence of vasopressor use in the SOFA score. Noradrenaline is the predominant vasopressor in use in the cohort. There is less frequent, but non-ignorable use of adrenaline, and so a composite marker for "noradrenaline equivalents" was created. Noradrenaline equivalents are the sum of both noradrenaline and adrenaline in the units of mcg/Kg/min.

6.3.6 Model Morphologies

For each biomarker three models were specified, each with a different association structure representing a different morphology of interest:

- severity; an instantaneous measure of the magnitude of disease.
- trajectory (severity + velocity); the path of severity at a given moment. This is what most clinicians rely on for intuition when evaluating the prognosis of a patient, or the relevance of a biomarker result. All new results are contextualised (where possible) to what came before.
- cumulative effect; the history of exposure up to and including the point of interest. If any biomarker is thought to exhibit an effect in an aggregate manor, as was seen in the previous chapter, then the cumulative effect parametrisation is best placed to elicit these effects.

In precis of subsection 2.4.1 (page 49), these conform to the morphological patterns shown in figure 6.8. This figure provides an illustration to orientate on the results that follow.

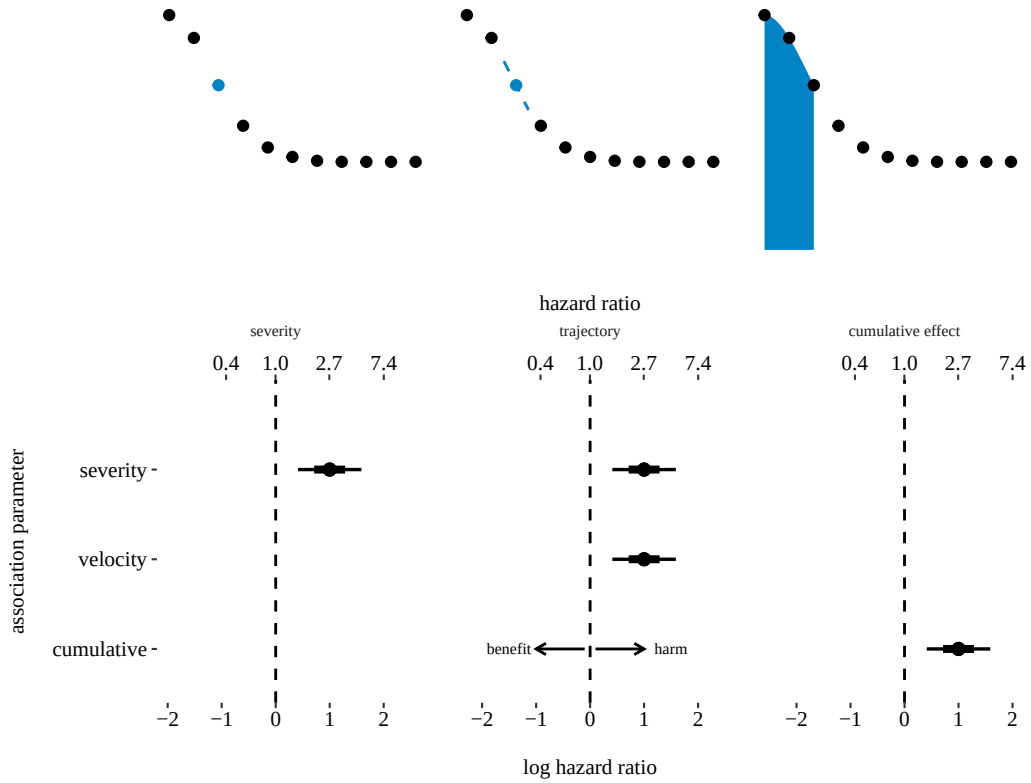


Figure 6.8: Graphical illustration to aid in the interpretation of joint model results. Results from the three morphologies (severity, trajectory and cumulative effect) will be presented graphically with the same approach. Each column of plots corresponds to a particular morphology. To assist in orientation, a graphical depiction of the morphology is presented here in the first row. Left column: the severity morphology, represented by the blue observed point. Middle column: the trajectory morphology, represented as a dashed blue line and point representing the simultaneous measure of disease severity (value) and velocity (slope). Right column: the cumulative effect morphology, represented by the shaded blue area under the biomarker. Each row of results corresponds to joint models for the same biomarker. The three possible association parameters are displayed on the y axis; velocity, severity, and cumulative effect. The severity model only displays the severity association parameter. The trajectory model displays both severity and velocity association parameters. The cumulative effect model only displays the cumulative effect association parameter. The null effect is shown as a vertical dashed line at a hazard ratio of 1 (log hazard ratio of 0). Coefficients displayed to the right (larger hazard ratio) show a stronger association for potential harm. The opposite is true for coefficients displayed to the left. Point estimates, 68% and 95% confidence intervals are displayed as points and lines of diminishing thickness.

6.4 Results

The characteristics of patients included in the sepsis cohort are included in table 6.3. Overall, this cohort is typical for those with sepsis in ICU, with a baseline mortality of around 20%. As would be expected, most patients are medical (86%), male (60%) and without any comorbidies (73%). There was good representation of different organ systems affected, although the majority of cases were respiratory insults (57%).

Characteristic	Overall	Sepsis	Septic shock	p-value
N	4,188	3,216	972	
Admission type				0.2
Medical	3,588 (86%)	2,743 (85%)	845 (87%)	
Surgical	600 (14%)	473 (15%)	127 (13%)	
Organ system				
Cardiovascular	403 (9.6%)	272 (8.5%)	131 (13%)	
Dermatological	82 (2.0%)	56 (1.7%)	26 (2.7%)	
Endocrine	53 (1.3%)	46 (1.4%)	7 (0.7%)	
Gastrointestinal	422 (10%)	297 (9.2%)	125 (13%)	
Genito-urinary	374 (8.9%)	263 (8.2%)	111 (11%)	
Haematological	107 (2.6%)	75 (2.3%)	32 (3.3%)	
Musculoskeletal	89 (2.1%)	73 (2.3%)	16 (1.6%)	
Neurological	255 (6.1%)	210 (6.5%)	45 (4.6%)	
Poisoning	9 (0.2%)	8 (0.2%)	1 (0.1%)	
Respiratory	2,381 (57%)	1,905 (59%)	476 (49%)	
Trauma	13 (0.3%)	11 (0.3%)	2 (0.2%)	
Age	64 (51, 75)	63 (50, 75)	65 (52, 75)	0.11
Sex				0.7
Female	1,691 (40%)	1,293 (40%)	398 (41%)	
Male	2,497 (60%)	1,923 (60%)	574 (59%)	
Comorbidities				0.046
0	3,038 (73%)	2,336 (73%)	702 (72%)	
1	752 (18%)	593 (18%)	159 (16%)	
2	308 (7.4%)	227 (7.1%)	81 (8.3%)	
3	80 (1.9%)	52 (1.6%)	28 (2.9%)	
4	10 (0.2%)	8 (0.2%)	2 (0.2%)	
CPR	128 (3.1%)	69 (2.1%)	59 (6.1%)	<0.001
Any dependency	1,183 (28%)	903 (28%)	280 (29%)	0.7
Spell LOS (days)	6 (3, 13)	5 (3, 12)	7 (3, 16)	<0.001
ICU mortality	822 (20%)	462 (14%)	360 (37%)	<0.001

Table 6.3: Patient characteristics for the sepsis cohort. Statistics presented: n (%); Median (IQR). Statistical tests performed: chi-square test of independence; Wilcoxon rank-sum test; Fisher's exact test. LOS = length of stay.

In order to provide an indication as to the representativeness of the models, example model fits are shown for a sample of patients for the trajectory model

across all biomarkers in figure 6.9. The varied and often non-linear paths taken by patients are demonstrated.

The model coefficients for the baseline patient features are shown graphically in figure 6.10 The majority of baseline coefficients across all joint models exhibit the same overall pattern; that is, the baseline coefficients are stable and relatively insensitive to the specification of the longitudinal biomarker. In general, baseline SOFA, the presence of dependencies and comorbidities, and age were positively associated with mortality. Weight and male sex were negatively associated with mortality. CPR prior to arrival had a variable relationship with mortality, with the confidence interval for this variable often containing the null. Noradrenaline is not shown, as models for this longitudinal outcome would not converge under the primary joint model specification.

The coefficients for the association parameters are shown in figure 6.11. In all cases, the cumulative effects parametrisation has a clinically negligible association with mortality. The trajectory models consistently demonstrate a strong relationship between both the velocity and severity parameters, and mortality. This relationship always occurred in the direction that would be expected, depending upon which directional change in the biomarker indicates a deterioration (for example, P_aO_2/F_IO_2 decreases with deterioration, while SOFA score increases with deterioration.) In all instances, the velocity parameter has a much larger effect size on the outcome than the value parameter. The coefficients obtained for the severity parameter were insensitive to the inclusion of the velocity parameter. In comparison to the baseline coefficients shown in figure 6.10, we can observe that changes in the longitudinal biomarker demonstrate a much large effect size.

The tabular representation of the association parameters is shown in table 6.4. Coefficients for all model parameters are shown in appendix table A.2 on page 269.

There was a concern that the precision of the estimates shown in figure 6.11 were too high (i.e. that the confidence intervals are too small). Even in the presence of relatively large volume of data, it seemed unlikely that the confidence intervals demonstrated would cover the correct proportion of cases. In the semiparametric

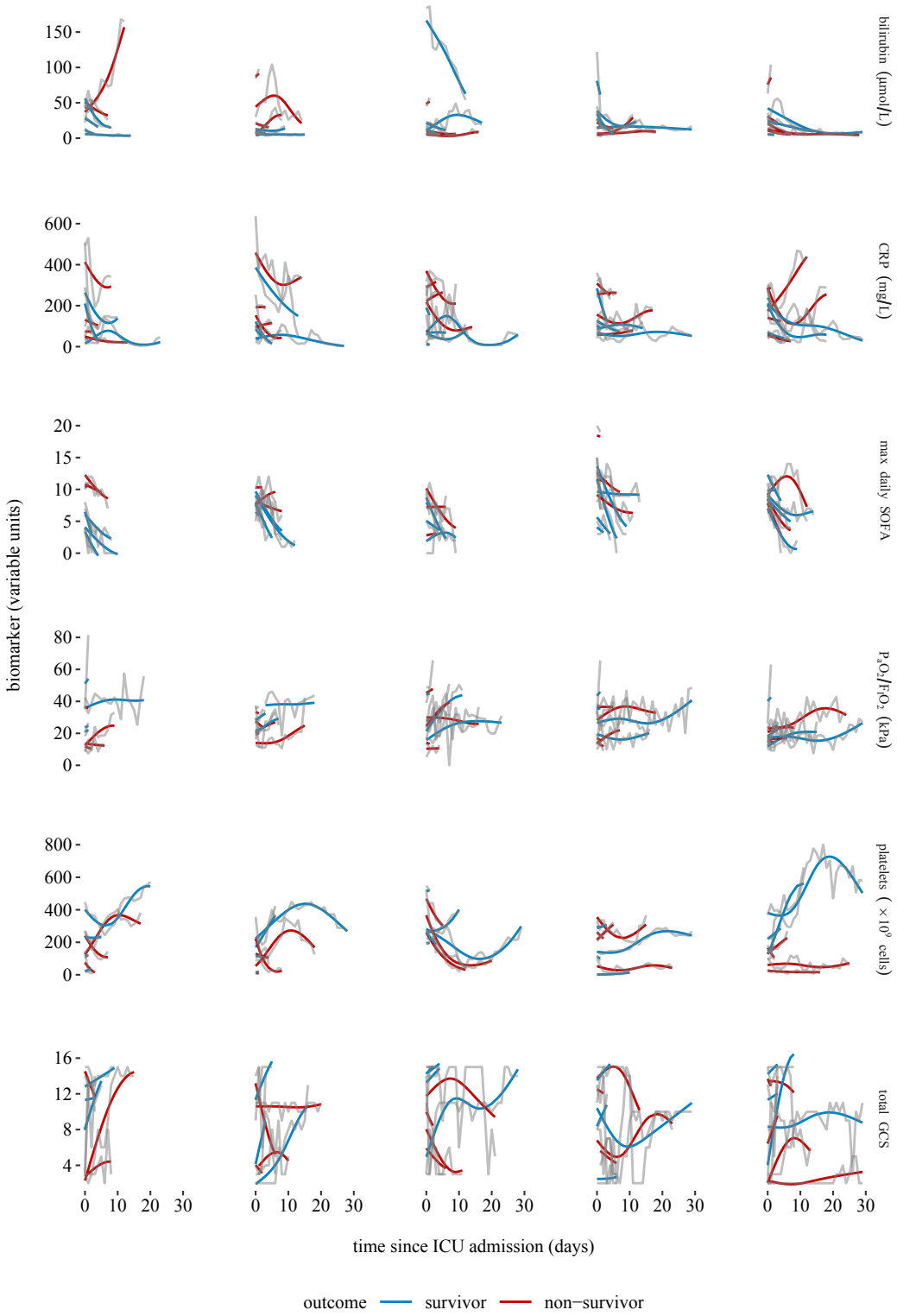


Figure 6.9: Example model fits drawn from the trajectory joint models

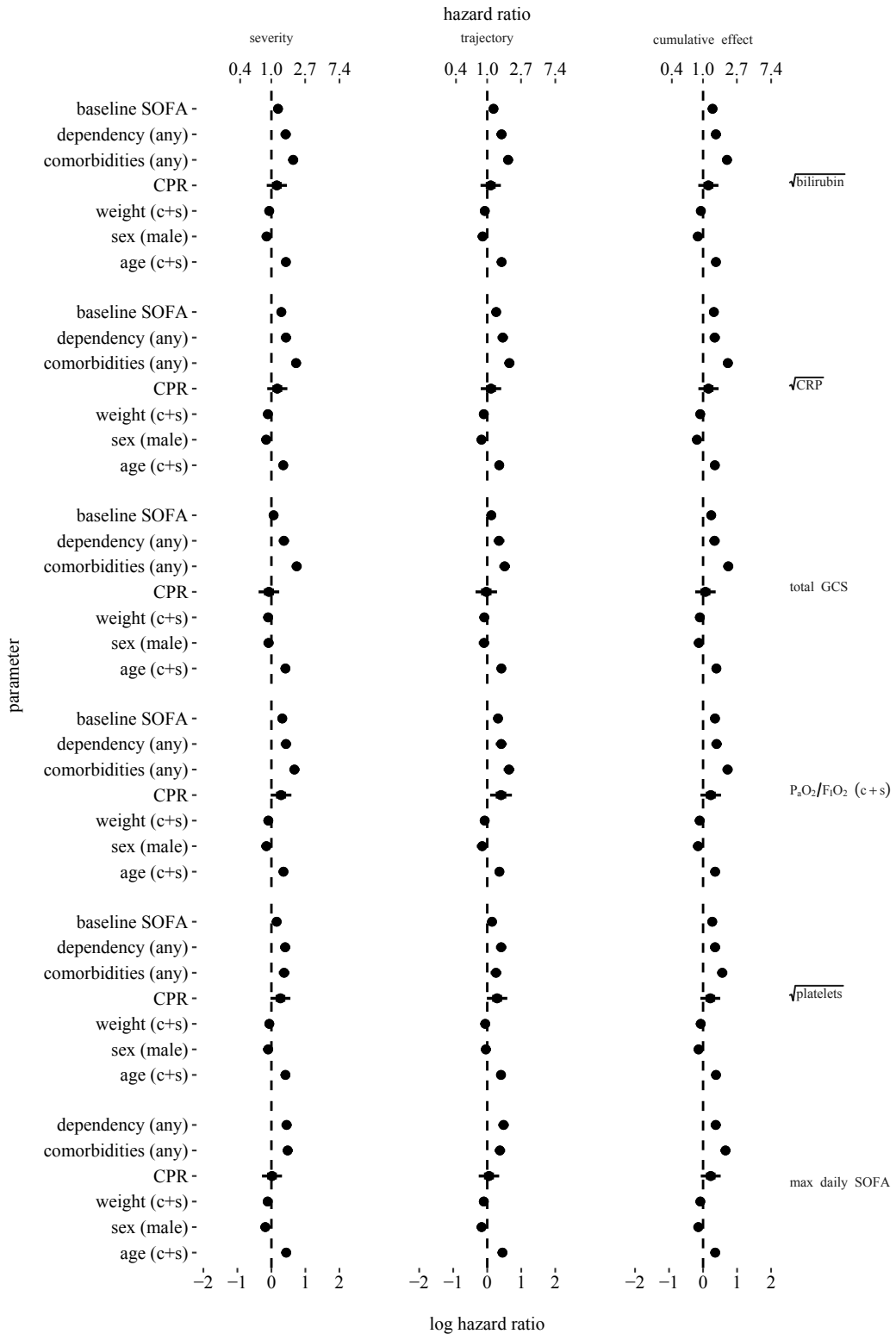


Figure 6.10: Baseline coefficients for all univariate joint models. A very similar pattern is seen across nearly all biomarkers and association structures, highlighting a lack of sensitivity of the baseline variables to these model elements. Other features are as described in figure 6.11. c+s = centered and scaled variable.

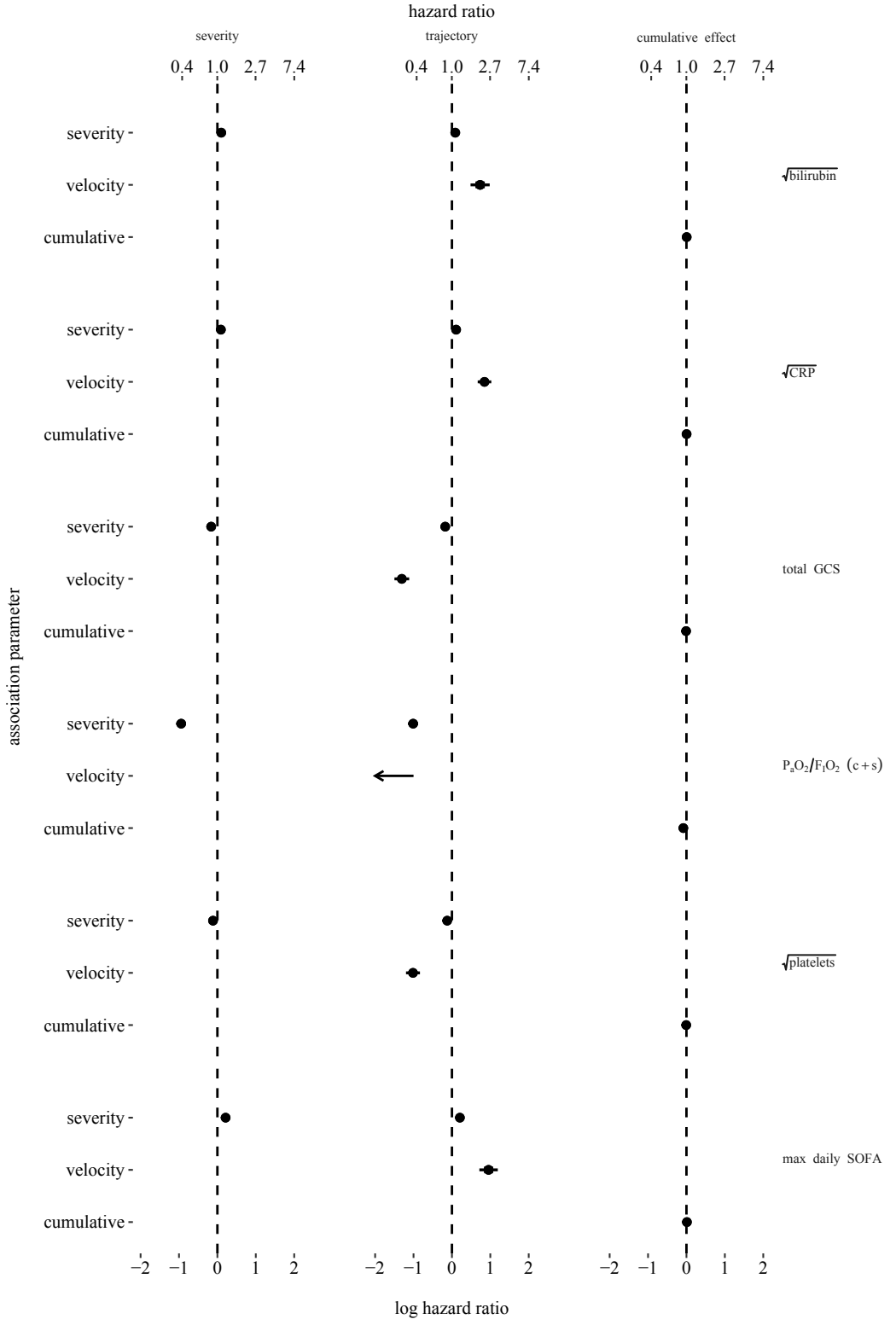


Figure 6.11: Association coefficients for all univariate joint models. Models for the same biomarkers are shown in rows. Models for the same morphology are shown in columns, from left to right: severity, trajectory, and cumulative effect. Results for the velocity parameter from the P_aO_2/F_IO_2 ratio trajectory model have been displayed as an arrow since the effect size is outside the limits of the plot. The log hazard ratio scale is displayed on the bottom x axis, the hazard ratio scale is displayed on the top x axis. Point estimates, 66% and 95% confidence intervals are shown. The dashed veritcal line is the point of null effect.

		log hazard ratio			hazard ratio				
Model	**Parameter**	**Coef**	**L95**	**U95**	**Coef**	**L95**	**U95**	**z**	**p-value**
Bilirubin									
Severity	Severity	0.10	0.08	0.12	1.11	1.09	1.12	12.37	3.72×10^{-35}
Trajectory	Severity	0.09	0.07	0.10	1.09	1.07	1.11	9.97	1.97×10^{-23}
Trajectory	Velocity	0.73	0.48	0.98	2.08	1.62	2.67	5.77	8.01×10^{-9}
Cum E.	Cum E.	0.01	0.00	0.01	1.01	1.00	1.01	7.23	4.68×10^{-13}
CRP									
Severity	Severity	0.09	0.07	0.11	1.09	1.07	1.12	9.03	1.71×10^{-19}
Trajectory	Severity	0.11	0.09	0.13	1.12	1.09	1.14	11.10	1.20×10^{-28}
Trajectory	Velocity	0.85	0.67	1.03	2.34	1.96	2.79	9.40	5.58×10^{-21}
Cum E.	Cum E.	0.00	0.00	0.01	1.00	1.00	1.01	3.25	1.14×10^{-3}
GCS									
Severity	Severity	−0.16	−0.18	−0.14	0.85	0.83	0.87	−14.37	8.19×10^{-47}
Trajectory	Severity	−0.17	−0.20	−0.15	0.84	0.82	0.86	−15.04	4.27×10^{-51}
Trajectory	Velocity	−1.30	−1.50	−1.11	0.27	0.22	0.33	−13.32	1.73×10^{-40}
Cum E.	Cum E.	−0.01	−0.01	−0.01	0.99	0.99	0.99	−8.01	1.11×10^{-15}
P_aO_2/F_IO_2									
Severity	Severity	−0.94	−1.07	−0.82	0.39	0.34	0.44	−14.72	4.48×10^{-49}
Trajectory	Severity	−1.01	−1.14	−0.88	0.36	0.32	0.41	−15.30	7.75×10^{-53}
Trajectory	Velocity	−6.44	−7.91	−4.96	0.00	0.00	0.01	−8.56	1.12×10^{-17}
Cum E.	Cum E.	−0.08	−0.09	−0.07	0.92	0.91	0.94	−11.53	9.65×10^{-31}
Platelets									
Severity	Severity	−0.11	−0.13	−0.10	0.89	0.88	0.91	−13.04	7.66×10^{-39}
Trajectory	Severity	−0.12	−0.14	−0.11	0.88	0.87	0.90	−13.72	7.46×10^{-43}
Trajectory	Velocity	−1.01	−1.20	−0.83	0.36	0.30	0.43	−10.97	5.51×10^{-28}
Cum E.	Cum E.	−0.01	−0.01	−0.00	0.99	0.99	1.00	−8.53	1.47×10^{-17}
SOFA									
Severity	Severity	0.21	0.19	0.23	1.24	1.21	1.26	21.69	2.66×10^{-104}
Trajectory	Severity	0.21	0.19	0.23	1.23	1.21	1.25	20.75	1.34×10^{-95}
Trajectory	Velocity	0.95	0.71	1.19	2.60	2.04	3.30	7.80	5.96×10^{-15}
Cum E.	Cum E.	0.01	0.01	0.01	1.01	1.01	1.01	13.39	6.86×10^{-41}

Table 6.4: Association coefficients for univariate joint models. Findings are grouped by biomarker and morphological parameterisation. Cum E. = Cumulative effects.

model	association parameter	log hazard ratio (95% CI)	hazard ratio (95% CI)
severity	severity	0.21 (0.20, 0.24)	1.24 (1.22, 1.27)
trajectory	velocity	0.95 (0.73, 1.19)	2.62 (2.07, 3.28)
trajectory	severity	0.21 (0.19, 0.23)	1.23 (1.21, 1.26)
cumulative effect	cumulative	0.01 (0.00, 0.01)	1.01 (1.00, 1.01)

Table 6.5: Bootstrapped association parameters for SOFA joint models. empirical means and 95% confidence intervals are drawn from the sampling distribution of 250 resamples for each model.

approach taken by Cox for the proportional hazard model, the baseline hazard function is left unspecified[7] and the partial maximum likelihood is maximised. With the joint model specification, the baseline hazard function is unavoidably part of the maximum likelihood specification due to the shared random effects. This can result in a systematic under-estimation of the model standard errors when the baseline hazard is left unspecified [145]. Two general purpose solutions have been proposed to solve this problem. First, a parametric baseline hazard function can be used for full maximum likelihood inference to proceed without bias. Second, the non-parametric bootstrap is a useful general purpose tool that can be applied to provide confidence intervals. The general flexibility of the bootstrap comes at a high computational burden, particularly when applied to joint models which are themselves computationally demanding. The models demonstrated in figure 6.11 are all fitted using a piecewise-constant baseline hazard function. Although the standard errors should be valid under this approach, the models have been refitted under a weibull baseline hazard function to provide an alternate fully parametric baseline hazard. Bootstrapped confidence intervals have also been drawn from the weibull models for the trajectory SOFA model with 250 resamples[8] to provide corroborating evidence for this model. The model coefficients and standard errors (either those computed analytically in the weibull models, or those derived empirically from the bootstrap samples) are largely comparable between both piecewise-constant and weibull hazard functions. The bootstrapped parameters are shown in figure 6.12 and table 6.5.

[7] hence the use of the term *semi*-parametric

[8] the weibull models fit in a shorter time than the piecewise models, and so a bootstrap approach was more feasible with these models.

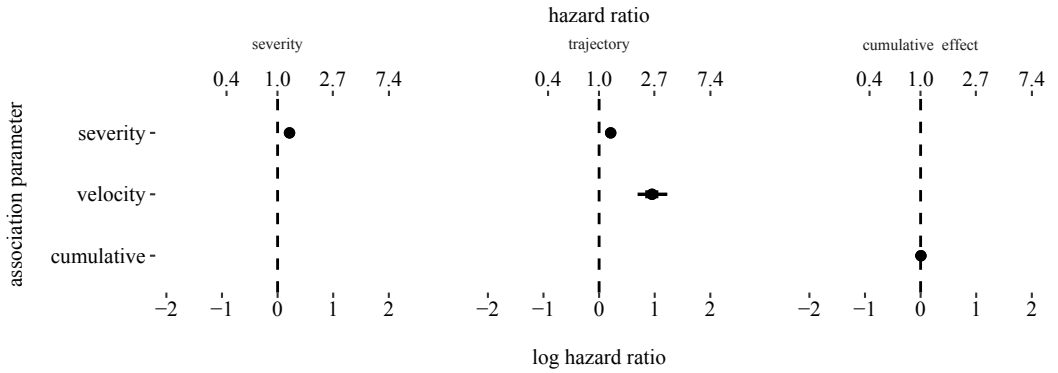


Figure 6.12: Association parameters for SOFA joint models with empirical bootstrapped 95% confidence intervals.

6.4.1 Model Evaluation

Model comparison was performed by evaluating each model's likelihood, Akaike information criterion (AIC) and Bayesian information criterion (BIC) (table 6.6). The trajectory model consistently proves to be the best representation of these data across all three criteria for each biomarker. Because the severity and trajectory models are nested (the trajectory model contains one additional parameter, in the form of the velocity association parameter), the likelihood ratio test can be applied to formally compare these models. In all cases, the likelihood ratio test when comparing the trajectory model to the severity model demonstrates a significant improvement in model fit by the inclusion of the velocity parameter.

A comparison can now be made between the marginal (population average) biomarker evolutions derived from the trajectory joint models and those from a naive analysis that does not consider informative censoring from death. These findings are shown in figure 6.13, which allows for a direct comparison for each longitudinal biomarker. On the whole, the generalised additive model (GAM) evolutions tend to be flatter than their joint model counterparts. In some cases, the GAM models fail entirely to capture a good representation of the evolution of the biomarker. This is particularly evident in the GCS, P_aO_2/F_IO_2, and SOFA models where the evolution in the GAM models trend toward more extreme values as they are likely dominated by those who remain alive and critically unwell inside the cohort. There is often an inflection in the biomarker evolutions at around seven days. Since this

model	LRT	p	logLik	AIC	BIC
bilirubin					
cumulative effect	-	-	$-49,189$	98,425	98,577
severity	-	-	$-49,151$	98,349	98,501
trajectory	30.5	3.30×10^{-8}	$-49,135$	98,321	98,479
CRP					
cumulative effect	-	-	$-88,943$	177,934	178,086
severity	-	-	$-88,913$	177,873	178,025
trajectory	79.6	4.64×10^{-19}	$-88,873$	177,796	177,954
GCS					
cumulative effect	-	-	$-92,384$	184,816	184,967
severity	-	-	$-92,302$	184,652	184,804
trajectory	178.4	1.11×10^{-40}	$-92,213$	184,476	184,634
noradrenaline					
cumulative effect	-	-	12,487	$-24,927$	$-24,775$
severity	-	-	12,564	$-25,080$	$-24,928$
trajectory	49.7	1.76×10^{-12}	12,589	$-25,128$	$-24,970$
PF					
cumulative effect	-	-	$-38,008$	76,063	76,214
severity	-	-	$-37,956$	75,959	76,109
trajectory	54.8	1.34×10^{-13}	$-37,928$	75,906	76,063
platelets					
cumulative effect	-	-	$-81,066$	162,180	162,332
severity	-	-	$-81,008$	162,064	162,216
trajectory	76.7	2.02×10^{-18}	$-80,970$	161,989	162,148
SOFA					
cumulative effect	-	-	$-84,731$	169,508	169,653
severity	-	-	$-84,585$	169,217	169,363
trajectory	55.7	8.26×10^{-14}	$-84,557$	169,163	169,315

Table 6.6: Piecewise joint model comparison characteristics. The Log-Likelihood (logLik), Akaike information criterion (AIC) and Bayesian information criterion (BIC) are detailed for all univariate joint models. Where models are nested (as is the case with the severity and trajectory models) a likelihood ratio test (LRT) has been performed. All tests have 1 degree of freedom. In all biomarkers, the trajectory model provides a better fit across all criteria than severity or cumulative effect models. The noradrenaline model failed to achieve proper convergence and is shown for completeness only.

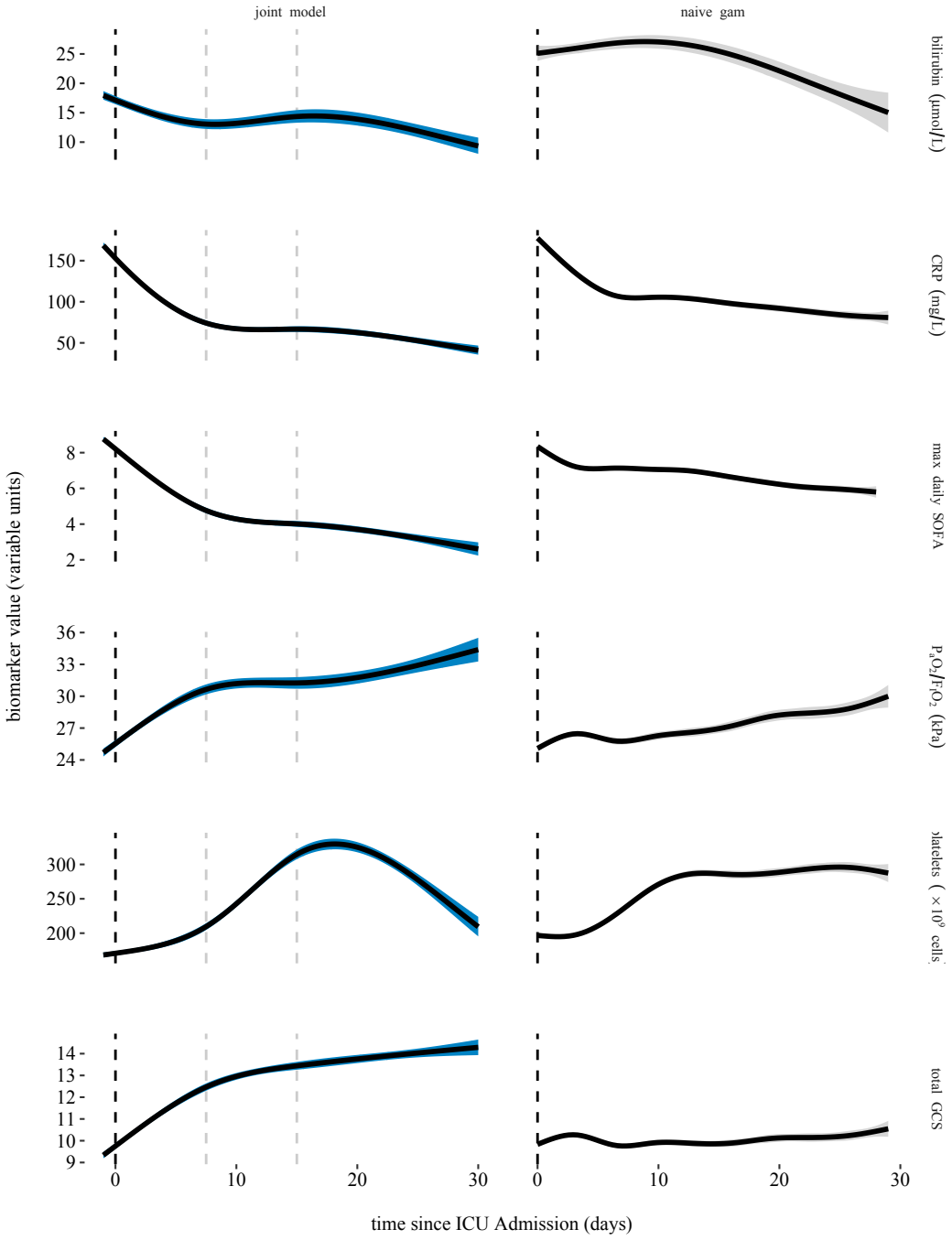


Figure 6.13: Comparison between joint model and naive marginal trajectories. The left column shows the marginal trajectories for each biomarker drawn from the trajectory joint models. Grey vertical dashed lines are provided at the point of spline knots (7.5 and 15 days). The right column shows the corresponding trajectories drawn from a GAM that does not take informative censoring into consideration.

coincides with the position of the first spline knot, it would be prudent to re-model at different knot positions to evaluate if the inflection point is sensitive to the spline specification. Though similar (albeit less pronounced) inflections are seen in some of the GAM models, where the number of spline knots (and hence their placement) is informed by the data itself, rather than as an *a priori* specification.

In each of the joint models, the evolutions have been cautiously extrapolated backward in time to one day prior to arrival in the ICU[9]. In all cases, the biomarkers are on an "improving" trajectory, suggesting that patients are improving at the point of ICU arrival.

The model performance was evaluated with dynamic area under the receiver operating characteristic (AUROC) (discrimination) and dynamic Brier scores (calibration). In each case, a time horizon of interest must be defined, over which the metric can be evaluated. The dynamic AUROC and dynamic Brier scores are shown graphically in figures 6.14-6.15 and 6.16-6.17 respectively. In each case, performance is evaluated from days 2-14, with a time horizon of 7 days into the future.

In all cases, discrimination and calibration of each biomarker decreased over time as the length of the time horizon increased. This highlights the increasing difficulty in predicting outcomes further into the future. The composite SOFA score yielded better performance than its individual components. While the trajectory models were superior in formal model comparisons, the picture is less clear when examining model performance stratified by time of interest. As is more clearly seen in the difference heatmaps in figures 6.15 and 6.17, there are regions of both improved and worse performance in the trajectory model when compared to the severity model. In general, the trajectory model was favourable within the first week of the cohort, and then performance deteriorated thereafter. The cumulative effect model consistently performed worse than the severity model, with the possible exception for CRP, which showed some possible improved discrimination in the second week.

[9]The use of restricted splines means that they are linear in their tails, and so short extrapolation beyond the limits of the data can be appropraite.

page intentionally left blank.

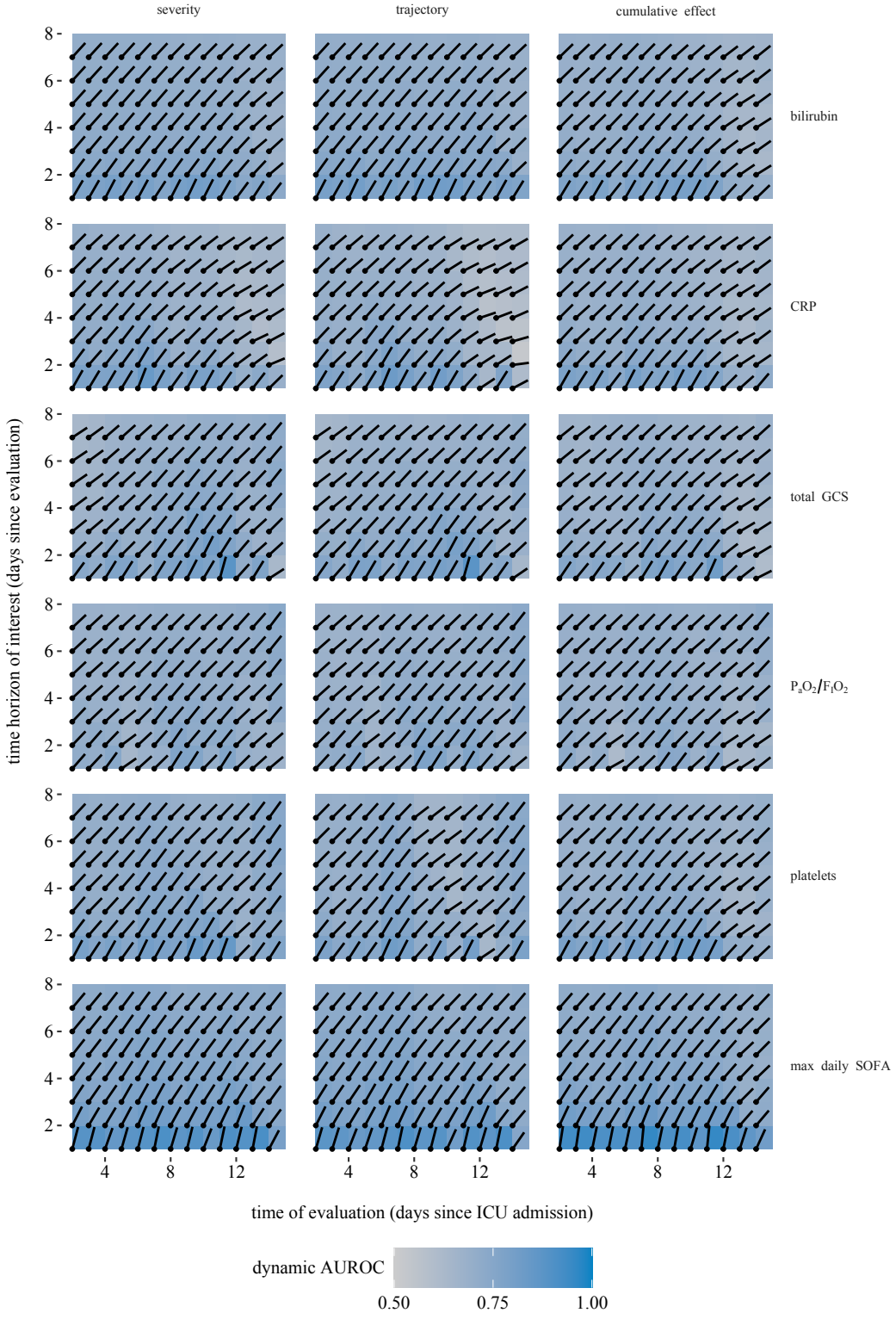


Figure 6.14: Dynamic area under the receiver operating characteristic (AUROC). The dynamic AUROC is shown for all biomarkers (rows) and morphologies (columns). The point of evaluation is shown on the x-axis from day 2-14. A time horizon of interest of up to 7 days is shown on the y-axis. Colour mapping is provided with deeper blue hue indicating better performance. Radial lines are drawn to emphasise this same effect with lines vertical at 90° indicating better performance.

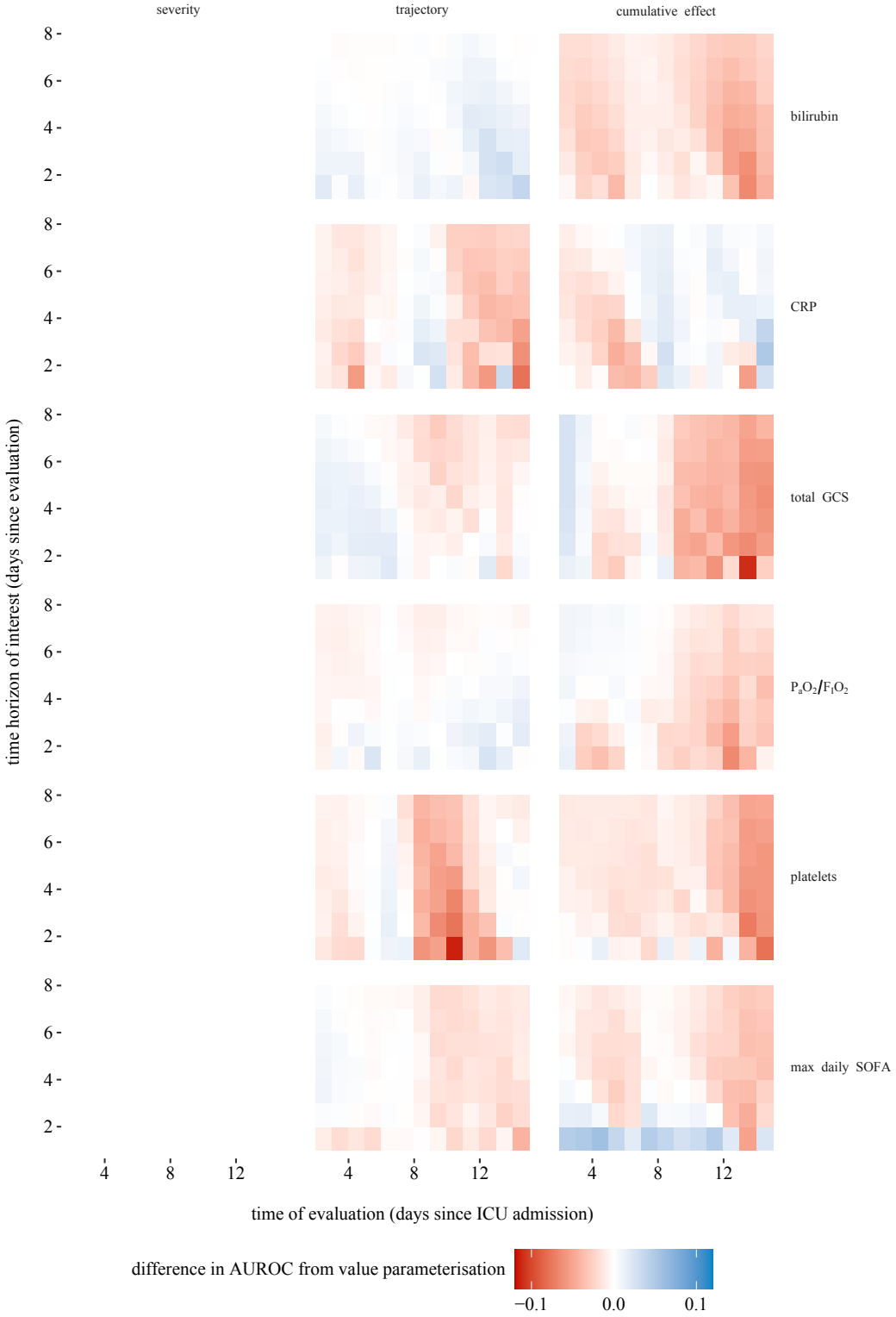


Figure 6.15: The setup of the plot is as for figure 6.14. The value in cell for the trajectory and cumulative effect models has been subtracted from the severity model. The result is displayed as a difference heatmaps showing the difference for each time point compared to the severity model, which is shown in white as the reference. Areas of red and blue indicate worse or better performance respectively.

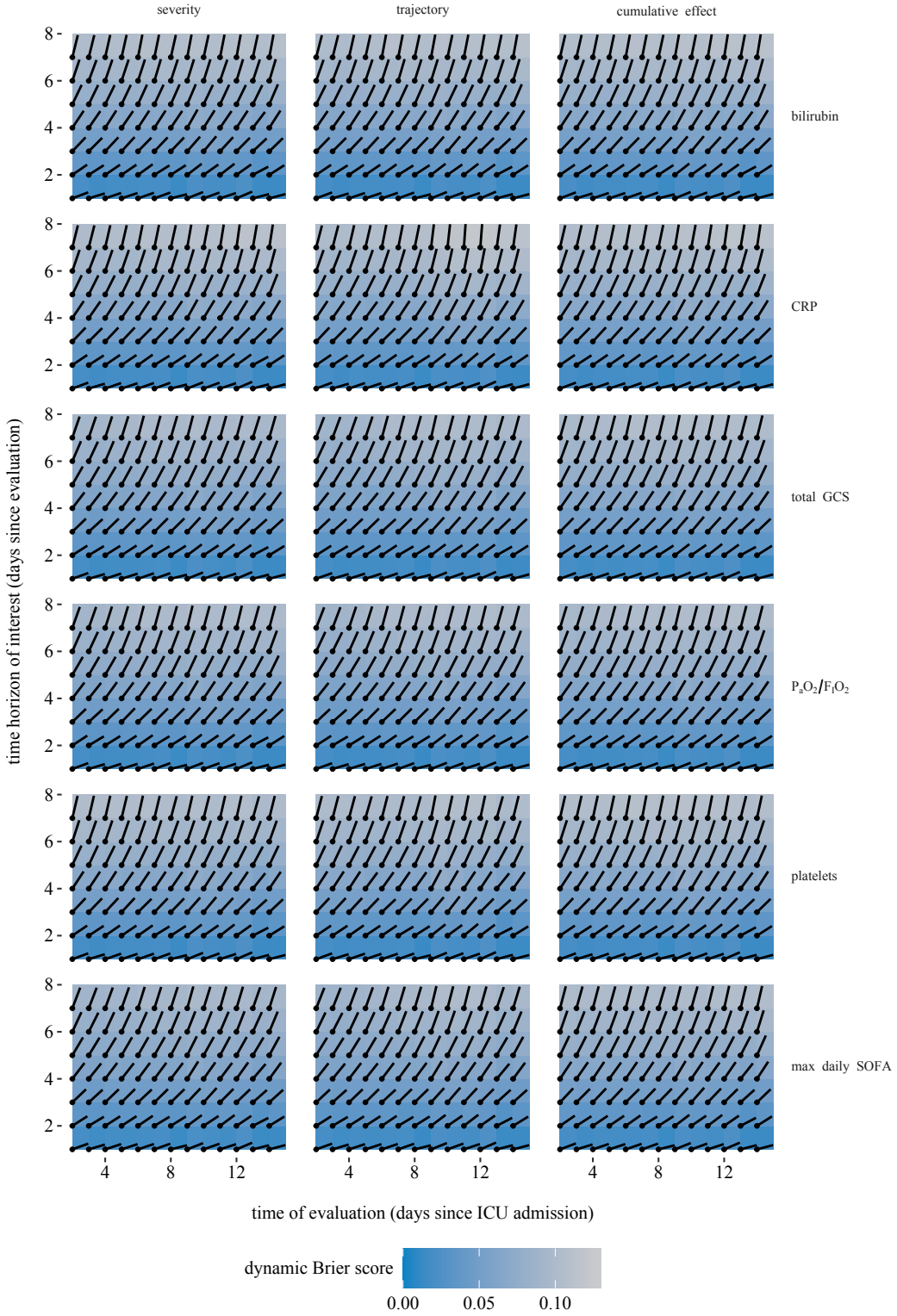


Figure 6.16: Dynamic Brier scores. The dynamic Brier score is shown for all biomarkers (rows) and morphologies (columns). The point of evaluation is shown on the x-axis from day 2-14. A time horizon of interest of up to 7 days is shown on the y-axis. Colour mapping is provided with deeper blue hue indicating better performance. Radial lines are drawn to emphasise this same effect with lines horizontal at 90° indicating better performance.

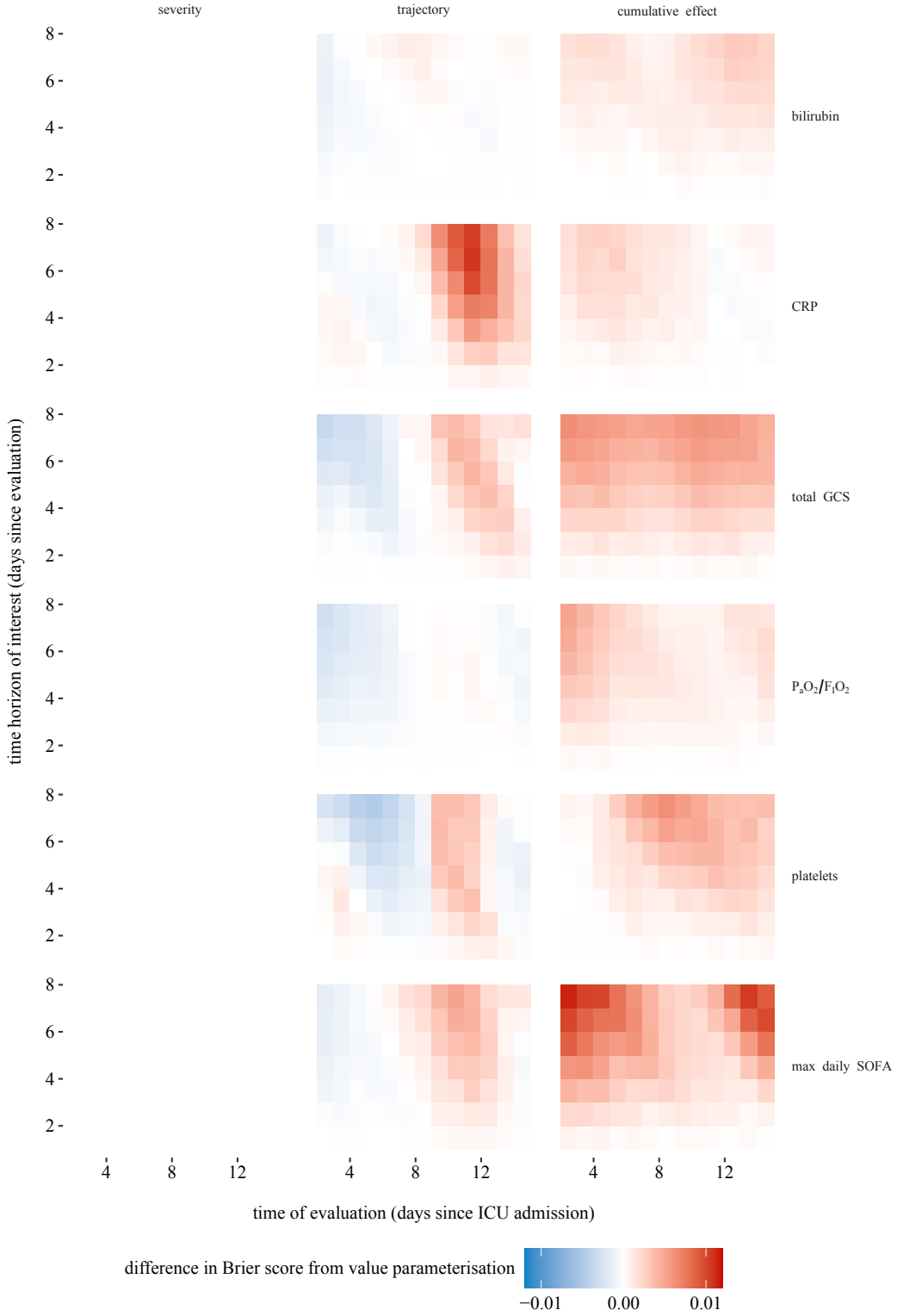


Figure 6.17: The setup of the plot is as for figure 6.16. The value in cell for the trajectory and cumulative effect models has been subtracted from the severity model. The result is displayed as a difference heatmaps showing the difference for each time point compared to the severity model, which is shown in white as the reference. Areas of red and blue indicate worse or better performance respectively.

6.5 Discussion

This study has provided encouraging evidence that biomarker morphologies in sepsis are differentially related to outcomes. What is most striking about these findings, is that regardless of the biomarker under investigation, the same pattern of findings was observed. Based on log likelihood and information criteria, the trajectory model provided a superior fit to these data than other morphologies. Further, the effect on outcome associated with an increase in the velocity of a biomarker was greater in all cases than the corresponding increase in severity. This effect size was greater than the typical effect sizes that were seen in any of the baseline patient characteristics, highlighting the relative importance of disease trajectories. In all cases, the cumulative effects parametrisation provided a very precise estimate for a negligible clinical effect on outcome. This raises questions over the transition between the shift from acute to chronic critical illness, and will be explored in this discussion.

The notion that different longitudinal morphologies of biomarkers could confer different relationships with outcomes in sepsis is not a new idea. Vincent *et al* developed the SOFA[10] score in 1996 [95]. In 1999, Meroeno *et al* [210] demonstrated that the change in SOFA score (ΔSOFA) correlated favourably with acute mortality. However, the performance of ΔSOFA fell short of the maximum day one SOFA score, even after adjustment for the SOFA score on arrival to ICU. These findings were corroborated shortly afterwords by Russell *et al* [211] and by Ferriera *et al* [212]. Using two time points (day zero and day three) Russell *et al* demonstrated an increased mortality in patients who showed an increase in organ dysfunction over this time. Ferreira *et al* demonstrated that increases in SOFA score within the first 48-96 hours of ICU were associated with a particularly high mortality (50% versus $< 35\%$). Their univariate analysis of 352 patients showed better discriminative performance in predicting outcome with the first day maximum SOFA score, rather than ΔSOFA, mirroring the earlier findings by Meroeno *et al* These earlier findings were corroborated by the larger SOAP observational study [213] in 2012.

In the intervening years, SOFA gained increasing popularity within the criti-

[10] At its introduction, the SOFA score was referred to as a the "sepsis related organ failure score" and has subsequently undergone a name change due to its broader application in critical care.

cal care community, culminating in its direct use in the sepsis-III definition [93]. Based on this prior research, the sepsis-3 definition is notable for its dynamic representation of the SOFA score to indicate salient organ dysfunction (a ΔSOFA of $\geq$ 2).

With increasing use not just as a target of inference in critical illness, but as an outcome in its own right, ΔSOFA was revisited by Degrooth *et al* in 2017 [214]. Their meta-analysis analysed 87 RCTs that reported SOFA score and mortality. Counter to some of the earlier evidence that shaped this field in the late 90s, Degrooth showed that although the SOFA score itself correlated with outcome, it explained only a small proportion of the outcome variance ($R^2 = 0.09$). Studies that reported ΔSOFA had a closer relationship with outcome, and explained a much higher proportion of the outcome variance ($R^2 = 0.32$). Though it is not entirely clear how the authors chose to include RCTs that did not make an account for informative censoring of SOFA.

Placing this current study in the context of this prior research, many of the comparable findings are in agreement. A consistent association was found between the trajectory of SOFA score and its components and mortality, with the velocity parameter showing a much larger effect on mortality than the value parameter. Many of the prior studies modelled ΔSOFA, over a single fixed timeframe (often 96 hours), in univariate analysis, and without consideration for informative censoring from death. This is analogous to modelling the velocity parameter in isolation, and at only a single point in time. After applying joint models to account for the informative censoring problem, the relationship becomes much clearer and is more in line with the findings from Degrooth *et al* [214].

Challenging this direct comparison, is the fact that I did not build any models that targeted the velocity parameter in isolation. A velocity only model would lack scientific plausibility and be challenging to interpret. To illustrate, we should reasonably expect that a patient moving from a SOFA score from 0 to 1 has a vastly better prognosis than a patient moving from 16 to 17. These patients demonstrate the same SOFA velocity, which must be contextualised with the severity of SOFA,

for this to be interpretable.

Three studies exploring this domain from the joint modelling perspective include the methodological studies from Deslandes and Chevret [135] and Musoro *et al* [136], and the clinical study from Harhay *et al* [215]. Deslandes and Chevret confirmed the relationship between SOFA severity and mortality [135]. Again this finding was corroborated by Musoro *et al* [136]. With direct clinical application Harhay *et al* re-analysed the data from an RCT that investigated the use of vasopressin in septic shock. None of these studies investigated the relationship between different longitudinal morphologies and outcome, which is to my knowledge a unique investigation of the subject matter.

A notable point of interest is the difference in findings when comparing models either based on likelihood or information criteria, and that with dynamic ROC and Brier methodologies. The trajectory models were unequivocally selected as the preferred model based on likelihood and information criteria methods. However, when examining model performance over a range of different time horizons, there was varied performance in the trajectory model. There are a number of reasonable explanations that may explain these findings. First, the power of the AUROC test as a means to compare models is lower than the likelihood approach, and so should be applied with caution for model selection. It may be that the model evaluations shown do not have the appropriate power to be used in the manner presented. Second, the performance indicated by the AUROC test is highly heterogeneous, with some regions performing better, and some worse. It may be, that on the whole, model performance is better with the trajectory model, but the peculiarities of this cohort render it less performant in certain places. It would be useful to validate these findings in an external cohort, to examine if these findings are replicable, or if they are specific to the cohort in hand.

6.5.1 Chronic Critical Illness

A prevailing sentiment in the critical care literature is that "chronic critical illness" is defined either arbitrarily at day 14, or by the point at which acute physiology is no longer predictive of a patient's outcomes [216, 217]. The findings here chal-

lenge this position, as there is little evidence that the acute physiology captured by the models was less predictive of outcome as time progressed. Indeed, the preference of the trajectory models, rather than the cumulative effects model suggests that outcomes are more closely related to acute physiology, and how that physiology is changing in any given moment, rather than the full history of how a patient arrived at that point in time.

The main evidence to counter this position is the cumulative effects CRP models. While the trajectory models were still selected as the best models for CRP, the AUROC evaluation does show there was better performance in the cumulative effects model beyond week 1. CRP is the only biomarker that captures inflammation. It is biologically plausible that the aggregation of inflammation could have a stronger relationship with outcomes, than any acute changes, particularly late in the clinical course. This aspect warrants further investigation.

6.6 Limitations

Efforts have been made to explicitly model the missing data mechanism, however I have not attempted to account for an informative sampling process. An example of an informative sampling process would be an increased sampling frequency of arterial blood gas samples, in a patient who is unwell. This may bias the system toward these more unwell cases and impact the study in two ways. First, there are scenarios in which the underlying data have been regularised to a fixed sampling schedule, for example, the SOFA models. In this situation, while the sampling frequency presented to the model has been regularised, the data that went into the SOFA calculation was not. This may have introduced bias into the SOFA calculation itself, of which we would be unaware. Second, there are also scenarios in which the underlying data has been modelled directly, for example, CRP. These biomarkers tend to have a natural daily cadence of measure in the ICU, but might be monitored more frequently in more severely unwell patients.

SOFA is an imperfect encapsulation of the treatment-physiology interaction of organ dysfunction, with numerous artefactual problems that may bias models that

use it to describe organ dysfunction trajectory. Its main limitations include:

- it is an ordinal scale, but most commonly modelled as if it were continuous (as has also been performed in this study) which is known to cause inferential issues [218].
- the categorisation of underlying continuous biomarkers reduces statistical efficiency.
- SOFA considers each organ system as independent, when they are demonstrably interconnected.

Despite these limitations, the SOFA model generally outperformed the other component models, particularly in the very short term.

The implementation of sepsis-3 in this study relies on potentially indirect evidence of infection through the use of ICNARC diagnostic codes. This limits the study to patients in whom sepsis is suspected on arrival into the ICU (which may itself be advantageous for the study) but does not permit the study of patients who go on to develop sepsis during their ICU episode. Suspicion of infection was not confirmed with the use of microbiology data, as the investigation of data quality outlined in chapter 4 does not provide enough confidence that this approach would be reliable.

Vasopressor and inotrope administration remains a point of concern. Usage of these drugs in the cohort was lower than would be clinically anticipated for patients who have sepsis in ICU. A possible explanation is that alternative forms of inotropy that are not captured by the CC-HIC data model are in use. For example, one site employs the use of glucose-insulin-potassium infusions that augment the cardiovascular system but do not feature in the CC-HIC data model. A more likely explanation is that data is missing due to translation errors at the point of local data extraction. In either case, the exploration of the cardiovascular system should proceed with caution, and the total SOFA score may have a systematic underestimation from a lack of cardiovascular data.

6.6.1 Cardiovascular System

The models investigating noradrenaline equivalents would not converge properly. While estimates were obtainable, valid standard errors could not be calculated, and so the estimates from these models should not be used without further investigation. It is possible that the heavy weighting of zeros in this variable proved challenging for convergence. A possible solution would be to perform a data reduction technique (such as PCA) over all cardiovascular data, and to regress against the principal components.

6.6.2 Other Forms of Censoring

This study has mainly concerned itself with informative right censoring. It should be noted that left censoring of biomarkers also features in this cohort. Patients are assumed to have had the onset of sepsis either at the very beginning of the ICU episode, or prior to arrival. As discussed, on the biological time scale, there is going to be a variable and fundamentally unknown period of time that patients are treated between the onset of sepsis and before arriving in ICU. This stands to maximally impact upon the cumulative effects parametrisation of the joint models, where this censoring may reduce any association seen in ICU.

6.6.3 Competing Risks

The joint models presented consider patient death as the only time-to-event outcome. There is, however, a competing risk present in this cohort contributed by patients leaving the ICU alive. Discharge from an ICU is on the whole predicated on the resolution of organ dysfunction. Once patients leave the ICU their physiological data is no longer submitted to the CC-HIC. Failing to account for this additional non-random loss of patient data from the cohort could potentially bias findings. The current approach was taken as a simplifying assumption, but it would be important to re-visit and address this in any future work. The use of a competing risks sub-model, in place of the currently used Cox or Weibull survival sub-models, would allow for the cause specific hazards from both death and discharge to be taken into consideration.

6.7 Conclusions

This study overcame a number of methodological and technical barriers including reduced data availability, and a lack of data harmonisation in key concepts. The same novel solutions as described in the hyperoxaemia study were implemented to recover sufficient data to conduct the primary research question. The study suggests that for all biomarkers investigated, biomarker trajectory was more strongly related to outcome than the value or cumulative effect of the biomarker. The consistent ability of these biomarkers to discriminate on outcome raises implications for the definition of chronic critical illness, implying that the acute phase of disease may last longer than previously described. The consistent evidence in support of the trajectory of biomarkers provides supporting evidence that risk prediction and prognostic trial enrichment should incorporate a dynamic function of treatment response, rather than just the level of organ dysfunction at a given time. The underlying biological drivers of these trajectories has not been explored here, because this does not form part of the routine data collection during the normal course of patient treatment.

Part III

Conclusions

Chapter 7

Conclusions

7.1 The CC-HIC

With its ancestry rooted in the ICNARC data model, the CC-HIC data model should have excelled in its role, supporting the secondary use of routinely collected critical care data for research. Perhaps unexpectedly, this was not the case. At the transition between a conventional disease registry and a full realisation of the EHR as a research platform for secondary use, the CC-HIC data model fails to provide necessary support. The CC-HIC data model is burdened in parts by its complexity and its "strong specification" where each data concept is hard coded into the schema directly. In other areas, complexity is lacking, limiting the potential utility of information as was demonstrated with representations of both microbiology data and medicines administration. An overall lack of consistency in approach has placed a large cognitive burden on the data engineers who export data into the CC-HIC data model, and the researchers who use the platform. This is perhaps most strongly highlighted by the lack of interoperability between data models used by different HIC themes, meaning that cross discipline research from within the same research platform would currently be challenging.

The data quality evaluation of the CC-HIC research database was written as a software contribution to this thesis as the R package inspectEHR [6]. inspectEHR [6] applied a standardised approach to data quality evaluation, necessitating the development of a novel accompanying data quality schema through which differ-

ent kinds of data quality issues could be presented alongside research data. As was the goal of the evaluation, numerous data quality deficiencies were discovered. This allowed contributing sites to make gradual improvements to data quality over time. This showcased a functional iterative data quality cycle for the duration of the project. Where data quality issues could not be corrected, limits could be set for inferences for the clinical research studies that followed. This was a vitally important step to ensure that research data were used responsibly by indicating where bias may be present from both the source EHR and the research pipeline itself.

The inherent challenges of sharing complex healthcare data from multiple NHS organisations have been clearly identified. Different NHS organisations store and access data in distinct ways and are able to contribute different levels of resource and technical expertise to data sharing projects. In such an environment the more that resources are allocated to bespoke endeavours—such as developing and implementing a bespoke data model for each theme of the HIC—the less that resources will be available for the routine maintenance of research data pipelines. Similarly, as research data pipelines become more technically specialist, the pool of trained individuals who can work with such resources is diminished. The penalty for building healthcare data models *de novo*—unless there is a genuine need—is therefore high.

Current conversations in this space often push toward the implementation of more advanced solutions to the interoperability and data sharing problem, for example Fast Healthcare Interoperability Resources (FHIR) [170] or openEHR [219, 220, 221]. Both these solutions are as comprehensive and elegant as they are complex. And while these solutions represent laudable goals, as we reflect on the research presented in this thesis, it is prudent to evoke the aphorism "perfect is the enemy of good". The technical specifications of these technologies are far in excess of the requirements for most typical observational research projects. Part of this desire to push for the latest and greatest technologies may arise from a perception that there is an unacceptable amount of data loss when transitioning into a certain clinical data model. However, clinical data models such as the OHDSI CDM

are well validated in this regard [17].

The learning that has taken place from both an applied use of the CC-HIC data model, coupled with these critical appraisals, led to a number of recommendations to be taken forward by the CC-HIC. Many of these recommendations set the tone for the next phase of the CC-HIC, with an understanding that quality and sustainability, are of paramount importance. In this spirit, the CC-HIC is currently implementing a modularised OHDSI data model. Tables are submitted one-by-one and populated as research requirements and data availability allow. The ambition is that this will allow for a tight feedback cycle between contributing sites and the central hub with a focus on regular updates to data and quality.

7.2 Exemplar Clinical Studies

Two clinical exemplar studies were conducted to evaluate areas of clinical importance using the unique feature of longitudinal data of the CC-HIC. Several problems arising within the research pipeline necessitated the development of solutions to ensure that research data were of a high enough standard to be used to address clinical research questions.

7.2.1 Cumulative Exposure to Oxygen

Oxygen administration at levels in excess of clinical need were readily demonstrated in the CC-HIC research cohort. Looking at just those samples with a P_aO_2 $\geq$13.3 kPa revealed a small but consistent association between excess exposure and a worse outcome. A dose dependency was not demonstrated, which—in accordance with the Bradford Hill criteria for causality—casts doubt over the potential for a causal relationship. One interpretation is that this might represent a ceiling effect, whereby a large, and unaccounted for, exposure to oxygen prior to admission to the ICU, renders the exposure seen in the ICU less relevant; has the damage already been done? Several small randomised controlled trials have failed to provide a definitive answer to this question, and much larger randomised controlled trials are now ongoing.

The modelling approach taken in the oxygen study reflects an earlier stage in

my statistical training. More appropriate methods, including the extended Cox or joint model, could have been employed to study this phenomenon.

7.2.2 Physiological Morphologies in Sepsis

The trajectory of organ dysfunction was consistently shown to be a better model fit, and associated with worse outcomes, than other biomarker morphologies. This finding is most striking for its consistency across all the different organ systems investigated. Previous criticisms have been made against SOFA for weighting each organ system equally, however the research presented here would argue strongly that this was in fact the correct approach.

These models show relatively little loss in discrimination or calibration over the short time horizons (in the order of a few days) under investigation, which may challenge the conventional wisdom on the definition of chronic critical illness. This is generally described as the point at which acute physiology fails to be predictive of outcome, and often described as being within two weeks of the onset of critical illness.

7.3 Future Work

This thesis is inherently cross discipline, positioning itself at the nexus of critical care medicine, healthcare data engineering, and applied statistics. There are a number of fertile areas to take forward in future work from across these domains.

7.3.1 Healthcare Data Engineering

The current data quality procedures have highlighted the importance of checking data quality as a routine component of secondary use research. The current procedures of inspectEHR [6], while extremely thorough, produce a large number of diagnostic plots for inspection and so this remains a manual and often overwhelming process. It has been difficult at times for contributing sites to identify and prioritise which changes should be made, unless directed towards assisting a specific research question. This issue can be resolved by integrating the evaluative process of inspectEHR [6] as a unit testing paradigm. This would involve specifying an acceptable error tolerance for each evaluative function. For example, records missing

outcome data could be set with a very low tolerance (e.g. "allow less than 1% of records to be missing outcome data"). This approach would allow the majority of data quality checks to proceed in an automatic fashion. The pointblank package for R is a contemporaneous example that successfully employs this approach for table validation [222]. The next steps therefore are to modify inspectEHR [6] to function against the OHDSI CDM backend, and extend the package to function in a unit testing paradigm. This would allow the package to support a range of applications, including both the CC-HIC and DECOVID.

7.3.2 DECOVID

In late 2019, a cluster of acute severe pneumonia cases presented in Wuhan, China. Shortly thereafter the beta-coronavirus SARS-CoV-2 was identified to cause the clinical syndrome of COVID-19. With no existing community resistance, COVID-19 proliferated across the globe with the WHO declaring a pandemic on 11th March 2020. There was a global call to the scientific community to address the deficits in knowledge to tackle this novel infection.

Although the CC-HIC was already established at the time of the COVID-19 pandemic, several limitations meant that it was not suitable to support research into COVID-19. Secondary to the "strong specification" issue, the existing data model proved too rigid to extend to include new data concepts (for example, COVID-19 status). Doing so would have required a fundamental re-write to core infrastructure at each contributing site. COVID-19 changed the way critical care is delivered in the UK. Patients who would normally be admitted to an ICU were treated outside the ICU owing to capacity issues. Temporary ICUs were created dynamically in response to extreme bed pressures. The CC-HIC data model did not support these developments, and so the complex patterns of patient movement within the hospital were not visible.

In response to the scientific call, and the limitations of existing data sharing platforms, the DECOVID consortium was established. DECOVID is a multi-centre research consortium between University Hospitals Birmingham (UHB), University of Birmingham, University College London Hospitals (UCLH), University College

London (UCL) and The Alan Turing Institute. DECOVID presented a unique opportunity to apply lessons from the lived experience of the CC-HIC. Many of the design decisions taken in DECOVID were aimed at removing limitations previously encountered with the CC-HIC.

Raw data were extracted from the local EHR at UCLH and UHB, and transformed into the OHDSI CDM (version 5.3.1). This data model undergoes some minor alterations, jointly developed by Roma Klapaukh and myself, to render it maximally useful in a UK environment, while meeting the security requirements required by DECOVID. Data are securely transferred to the UCL DSH, where dedicated analytic teams can work directly with the data using methods to which they are accustomed.

The DECOVID data specification is comprehensive, including many thousands of discrete concepts represented at the natural resolution of the source EHR, including:

- patient characteristics:
  - date of birth.
  - sex.
  - ethnicity.
- social factors:
  - smoking status and history.
- full resolution longitudinal laboratory findings:
  - diagnostic labels.
  - vital signs.
  - biochemistry.
  - near patient sampling.
  - sampling and reporting times.
- longitudinal treatments:
  - all pharmacology.
  - non-pharmacological (e.g. comprehensive ventilation parameters).

  - treatment limitation orders.

- patient outcomes:
  - mortality at any point of hospitalisation.
- patient movements:
  - clinically salient movement within the hospital.

DECOVID is now entering a testing phase, making available a large quantity of routinely collected data for secondary use research. My current involvement is both to help direct data quality evaluation procedures, and as a principal investigator for a research question that aims to identify an optimal ventilation strategy in severe COVID-19 pneumonitis. This ventilation research question is being undertaken in partnership with the London School of Hygiene and Tropical Medicine (LSHTM), applying the target trial approach [223] and a marginal structural modelling approach, with inverse probability weighting, to ascertain estimates for dynamic treatment regimens [224].

7.3.3 Physiological Morphologies and Risk Communication

A major component of the clinical research aspect of this thesis has been to understand the connection between longitudinal morphologies of physiology and outcomes. The goals of this thesis have largely been inferential in nature, though the joint models applied are equally powerful in the task of dynamic prediction; the process of predicting in light of updated information [160]. This has been used successfully in adjacent fields to, for example, provide patient specific surveillance schedules in cancer [225]. Current risk models in critical care are typically based upon the patients physiology from the first 24 hours following admission and baseline factors [29]. From a clinical perspective, it can be challenging to convey ongoing risk to a patient and their relatives. Using a validated risk prediction model that has the ability to dynamically update based upon the latest physiological information, would be extremely useful in terms of risk communication. In particular, the visualisation of a prediction interval would help patients and relatives understand

the high levels of uncertainty over making predictions for the days and weeks that lie ahead in critical care.

Prior to developing these models as risk prediction tools, some of the limitations previously discussed will need to be addressed, most notably accounting for the competing risks problem.

The current research design focused on the morphologies of individual biomarkers and their relationship with outcome. It would be important to look at this problem in the multivariate context, in order to understand how the different organ systems interact with each other. The SOFA score currently considers each organ system in isolation. Further research into this area to help quantify the relative weighting and interaction of different organ systems could provide the foundation for a revision to the SOFA score.

Treatment effects were not explored with the joint modelling paradigm. It would be prudent to align the research on organ dysfunction morphologies and exposure to hyperoxaemia together under one framework. This would serve three main goals. First, this would help bring the work on exposure to hyperoxaemia into a more methodologically grounded framework. Second, this would be a suitable means through which to explore how the proposed association with mortality is mediated. The effects of hyperoxaemia should be mediated through deteriorations in cardiovascular and respiratory function, and so recasting the question in the joint modelling paradigm would allow a deeper investigation into how these effects are mediated. Last, as a useful proof of principle for handling longitudinal exposures in general which are common place in critical care, including exposure to antibiotics, vasoactive substances and invasive devices such as ventilators.

7.4 Closing Remarks

The critical care community—and medicine more broadly—has in recent years seen a significant increase in the number of published studies that rely on large secondary use cohorts. With the rise in EHR data availability, a large mass of patient data is now readily available for study. Previously, retrospective studies of this nature would require a significant effort in terms of manual data collection and curation. This enforced a first pass of data—typically by a domain expert—to facilitate the production of a "clean" dataset that was fit for research use. The relative ease with which data *can* be extracted from an EHR has permitted easy access to healthcare data that has often undergone little clinical scrutiny. The NHS is ideally placed to create secondary use research platforms. It is the clear finding of this thesis that such platforms must have data quality as a foundational element. This requires a shift in focus from data volume to data quality, and a commensurate shift in funding from establishing the platform to maintaining it.

This research would not have been possible without the titan efforts of the CC-HIC who wrote the data model and bravely shared granular patient data across institutes. The difficulty of these endeavours should not be under-estimated, and this research would not have been possible without this vital foundation. It is incumbent on us to ensure that these cohorts are maintained to a high degree, commensurate to the effort that was expended in making them.

Part IV

Appendices

Appendix A

Tables

			limits		
NIHR HIC code	concept name	inspectEHR class	min	max	units
NIHR-HIC-ICU-0001	PAS number	string-1d	-	-	-
NIHR-HIC-ICU-0002	Site code (ICNARC CMP number)	string-1d	-	-	-
NIHR-HIC-ICU-0003	Code of GP	string-1d	-	-	-
NIHR-HIC-ICU-0004	Treatment function code	string-1d	-	-	-
NIHR-HIC-ICU-0005	Critical care local identifier / ICNARC admission number	integer-1d	-	-	-
NIHR-HIC-ICU-0006	CCU bed configuration 02	string-1d	-	-	-
NIHR-HIC-ICU-0007	Level 2 (HDU) days	integer-1d	0	365	days
NIHR-HIC-ICU-0008	Level 3 (ICU) days	integer-1d	0	365	days
NIHR-HIC-ICU-0009	Organ support maximum	integer-1d	0	11	systems
NIHR-HIC-ICU-0010	Acute myeloid/lymphocytic leukaemia or myeloma	integer-1d	0	1	-
NIHR-HIC-ICU-0011	Admission for pre-surgical preparation	integer-1d	0	1	-
NIHR-HIC-ICU-0013	Adult ICU/HDU within your critical care transfer group (in)	integer-1d	0	1	-
NIHR-HIC-ICU-0015	Antimicrobial use after 48 hours in your unit	integer-1d	0	1	-

NIHR-HIC-ICU-0016	Biopsy proven cirrhosis	integer-1d	0	1	-
NIHR-HIC-ICU-0017	Height	real-1d	1	3	m
NIHR-HIC-ICU-0018	Height (Source)	string-1d	-	-	-
NIHR-HIC-ICU-0019	Weight	real-1d	20	300	kg
NIHR-HIC-ICU-0020	Weight (Source)	string-1d	-	-	-
NIHR-HIC-ICU-0021	Cardiopulmonary resuscitation within 24 hours prior to admission to unit	string-1d	-	-	-
NIHR-HIC-ICU-0022	Basic Cardiovascular support days	integer-1d	0	365	days
NIHR-HIC-ICU-0023	Advanced Cardiovascular support days	integer-1d	0	365	days
NIHR-HIC-ICU-0024	Chemotherapy (within the last 6months) steroids alone excluded	integer-1d	0	1	-
NIHR-HIC-ICU-0025	Chronic myelogenous /lymphocytic leukaemia	integer-1d	0	1	-
NIHR-HIC-ICU-0026	Chronic renal replacement therapy	integer-1d	0	1	-
NIHR-HIC-ICU-0027	classification of surgery	string-1d	-	-	-
NIHR-HIC-ICU-0029	Congenital immunohumoral or cellular immune deficiency state	integer-1d	0	1	-
NIHR-HIC-ICU-0030	Critical care visit post-discharge from your unit	string-1d	-	-	-

NIHR-HIC-ICU-0031	Critical care visit prior to this admission to your unit	string-1d	-	-	-
NIHR-HIC-ICU-0032	date of admission to your hospital	date-1d	-	-	calendar date
NIHR-HIC-ICU-0033	Date of birth	date-1d	-	-	calendar date
NIHR-HIC-ICU-0034	Date of last critical care visit prior to this admission to your unit	date-1d	-	-	calendar date
NIHR-HIC-ICU-0035	Date of original admission to/attendance at acute hospital	date-1d	-	-	calendar date
NIHR-HIC-ICU-0036	Date of original admission to ICU/HDU	date-1d	-	-	calendar date
NIHR-HIC-ICU-0037	Date of ultimate discharge from ICU/HDU	date-1d	-	-	calendar date
NIHR-HIC-ICU-0038	Date body removed from your unit	date-1d	-	-	calendar date
NIHR-HIC-ICU-0039	Time body removed from your unit	time-1d	-	-	time
NIHR-HIC-ICU-0042	Date of death on yout unit	date-1d	-	-	calendar date
NIHR-HIC-ICU-0043	Time of death on your unit	time-1d	-	-	time
NIHR-HIC-ICU-0044	Date of declaration of brain stem death	date-1d	-	-	calendar date
NIHR-HIC-ICU-0045	Time of declaration of brain stem death	time-1d	-	-	time
NIHR-HIC-ICU-0048	Date treatment first withdrawn	date-1d	-	-	calendar date
NIHR-HIC-ICU-0049	Time treatment first withdrawn	time-1d	-	-	time

NIHR-HIC-ICU-0050	Date fully ready for discharge	date-1d	-	-	calendar date
NIHR-HIC-ICU-0051	Time fully ready for discharge	time-1d	-	-	time
NIHR-HIC-ICU-0053	Delayed admission	string-1d	-	-	-
NIHR-HIC-ICU-0054	Delay	real-1d	0	100	hours
NIHR-HIC-ICU-0055	Dependency prior to admission	string-1d	-	-	-
NIHR-HIC-ICU-0056	Dermatological support days	integer-1d	0	365	days
NIHR-HIC-ICU-0058	Ethnicity	string-1d	-	-	-
NIHR-HIC-ICU-0059	Gastrointestinal support days	integer-1d	0	365	days
NIHR-HIC-ICU-0060	Hepatic encephalopathy	integer-1d	0	1	-
NIHR-HIC-ICU-0062	HIV/AIDS	integer-1d	0	1	-
NIHR-HIC-ICU-0063	Home ventilation	integer-1d	0	1	-
NIHR-HIC-ICU-0065	Hospital housing location (in)	string-1d	-	-	-
NIHR-HIC-ICU-0066	Level of care at discharge from your unit	integer-1d	0	3	level
NIHR-HIC-ICU-0067	Liver support days	integer-1d	0	365	days
NIHR-HIC-ICU-0068	Location (in)	string-1d	-	-	-
NIHR-HIC-ICU-0069	Discharge location (location out)	string-1d	-	-	-
NIHR-HIC-ICU-0070	Lymphoma	integer-1d	0	1	-
NIHR-HIC-ICU-0071	Metastatic disease	integer-1d	0	1	-
NIHR-HIC-ICU-0072	Neurological support days	integer-1d	0	365	days
NIHR-HIC-ICU-0073	NHS number	string-1d	-	-	-

NIHR-HIC-ICU-0074	Other condition in past medical history	string-1d	-	-	-
NIHR-HIC-ICU-0075	Portal hypertension	integer-1d	0	1	-
NIHR-HIC-ICU-0076	Postcode	string-1d	-	-	-
NIHR-HIC-ICU-0080	Radiotherapy	integer-1d	0	1	-
NIHR-HIC-ICU-0081	Discharge status (Reason for discharge from your unit)	string-1d	-	-	-
NIHR-HIC-ICU-0082	Referred for solid organ or tissue donation	integer-1d	0	1	-
NIHR-HIC-ICU-0083	Renal support days	integer-1d	0	365	days
NIHR-HIC-ICU-0084	Residence post discharge from acute hospital	string-1d	-	-	-
NIHR-HIC-ICU-0085	residence prior to admission to acute hospital	string-1d	-	-	-
NIHR-HIC-ICU-0086	Basic respiratory support days	integer-1d	0	365	days
NIHR-HIC-ICU-0087	Advanced respiratory support days	integer-1d	0	365	days
NIHR-HIC-ICU-0088	Secondary reasons for admission to your unit	string-1d	0	1	-
NIHR-HIC-ICU-0092	Severe respiratory disease	integer-1d	0	1	-
NIHR-HIC-ICU-0093	Sex	string-1d	-	-	-
NIHR-HIC-ICU-0094	Solid organ or tissue donor	string-1d	-	-	-

NIHR-HIC-ICU-0095	Status at discharge from your hospital	string-1d	-	-	-
NIHR-HIC-ICU-0097	Dead or alive on discharge	string-1d	-	-	-
NIHR-HIC-ICU-0098	Status at ultimate discharge from hospital	string-1d	-	-	-
NIHR-HIC-ICU-0099	Steroid treatment	integer-1d	0	1	-
NIHR-HIC-ICU-0100	Transferring unit admission number	integer-1d	-	-	-
NIHR-HIC-ICU-0101	Transferring unit identifier (in)	string-1d	-	-	-
NIHR-HIC-ICU-0103	Treatment withheld/withdrawn	string-1d	-	-	-
NIHR-HIC-ICU-0104	Type of adult ICU/HDU (in)	string-1d	-	-	-
NIHR-HIC-ICU-0107	Very severe cardiovascular disease	integer-1d	0	1	-
NIHR-HIC-ICU-0108	Heart rate	integer-2d	0	300	bpm
NIHR-HIC-ICU-0109	Heart rhythm	integer-2d	1	31	-
NIHR-HIC-ICU-0110	Mean arterial blood pressure - Art BPMean arterial blood pressure	integer-2d	0	266	mmHg
NIHR-HIC-ICU-0111	Mean arterial blood pressure - NBPMean arterial blood pressure	integer-2d	0	266	mmHg
NIHR-HIC-ICU-0112	Systolic Arterial blood pressure - Art BPSystolic Arterial blood pressure	integer-2d	0	400	mmHg

NIHR-HIC-ICU-0113	Systolic Arterial blood pressure - NBPSystolic Arterial blood pressure	integer-2d	0	400	mmHg
NIHR-HIC-ICU-0114	Diastolic arterial blood pressure - Art BPDiastolic arterial blood pressure	integer-2d	0	200	mmHg
NIHR-HIC-ICU-0115	Diastolic arterial blood pressure - NBPDiastolic arterial blood pressure	integer-2d	0	200	mmHg
NIHR-HIC-ICU-0116	Central venous pressure	real-2d	−25	50	mmHg
NIHR-HIC-ICU-0117	Cardiac output - LiDCO Plus	real-2d	0	35	L/min
NIHR-HIC-ICU-0118	Cardiac output - LiDCO Rapid	real-2d	0	35	L/min
NIHR-HIC-ICU-0119	Cardiac output - PICCO	real-2d	0	35	L/min
NIHR-HIC-ICU-0120	Cardiac output - PA Catheter	real-2d	0	35	L/min
NIHR-HIC-ICU-0121	Cardiac output - Doppler	real-2d	0	35	L/min
NIHR-HIC-ICU-0122	Lactate - ABG	real-2d	0	40	mmol/L
NIHR-HIC-ICU-0123	Lactate - Lab	real-2d	0	40	mmol/L
NIHR-HIC-ICU-0125	Central venous saturation	real-2d	0	100	%
NIHR-HIC-ICU-0126	Airway	string-2d	-	-	-
NIHR-HIC-ICU-0129	SpO2	integer-2d	0	100	%
NIHR-HIC-ICU-0130	SaO2 - ABG	integer-2d	0	100	%
NIHR-HIC-ICU-0132	PaO2 - ABG	real-2d	0	90	kPa

NIHR-HIC-ICU-0134	PaCO2 - ABG	real-2d	0	30	kPa
NIHR-HIC-ICU-0136	pH - ABG / VBG	real-2d	6	8	-log[mmol/L]
NIHR-HIC-ICU-0138	HCO3 - ABG / VBG	real-2d	0	60	mmol/L
NIHR-HIC-ICU-0141	Temperature - Central	real-2d	15	45	degrees Celsius
NIHR-HIC-ICU-0142	Temperature - Non-central	real-2d	15	45	degrees Celsius
NIHR-HIC-ICU-0143	Position	integer-2d	1	7	-
NIHR-HIC-ICU-0144	Invasive or non-invasive (ventilation)	integer-2d	1	2	-
NIHR-HIC-ICU-0145	Total respiratory rate (monitor)	integer-2d	0	180	cycles/min
NIHR-HIC-ICU-0146	Total respiratory rate (ventilator)	integer-2d	0	180	cycles/min
NIHR-HIC-ICU-0147	Mandatory Respiratory Rate	integer-2d	0	60	cycles/min
NIHR-HIC-ICU-0148	Minute volume	real-2d	0	40	L/min
NIHR-HIC-ICU-0149	Peak airway pressure	integer-2d	0	80	cmH2O
NIHR-HIC-ICU-0150	Inspired fraction of oxygen	real-2d	0	1	-
NIHR-HIC-ICU-0151	Positive End Expiratory Pressure	real-2d	0	60	cmH2O
NIHR-HIC-ICU-0152	Airway pressure	integer-2d	0	80	cmH2O
NIHR-HIC-ICU-0153	Frequency (Hz)	integer-2d	0	30	Hz
NIHR-HIC-ICU-0154	Cycle Volume	integer-2d	0	1,000	mL
NIHR-HIC-ICU-0155	Base flow	integer-2d	0	1,000	L/min
NIHR-HIC-ICU-0156	GCS - total	integer-2d	3	15	-
NIHR-HIC-ICU-0157	GCS - motor component	integer-2d	1	6	-
NIHR-HIC-ICU-0158	GCS - eye component	integer-2d	1	4	-

NIHR-HIC-ICU-0159	GCS - verbal component	integer-2d	1	5	-
NIHR-HIC-ICU-0160	Sedation score (hourly)	integer-2d	−6	4	-
NIHR-HIC-ICU-0161	Renal replacement mode	integer-2d	1	2	-
NIHR-HIC-ICU-0162	Urine output	integer-2d	0	2,500	mL
NIHR-HIC-ICU-0164	Urea	real-2d	1	60	mmol/L
NIHR-HIC-ICU-0166	Creatinine	real-2d	1	2,000	micromol/L
NIHR-HIC-ICU-0168	Sodium	real-2d	110	170	mmol/L
NIHR-HIC-ICU-0169	Sodium ABG/VBG	real-2d	110	170	mmol/L
NIHR-HIC-ICU-0171	Potassium	real-2d	2	12	mmol/L
NIHR-HIC-ICU-0172	Potassium ABG/VBG	real-2d	2	12	mmol/L
NIHR-HIC-ICU-0174	Bilirubin	real-2d	0	500	mmol/L
NIHR-HIC-ICU-0175	Glucose ABG/VBG	real-2d	0	60	mmol/L
NIHR-HIC-ICU-0176	Glucose bedside test	real-2d	0	60	mmol/L
NIHR-HIC-ICU-0178	Haemoglobin ABG/VBG	real-2d	0	200	g/L
NIHR-HIC-ICU-0179	Haemoglobin	real-2d	0	200	g/L
NIHR-HIC-ICU-0182	White cell count	real-2d	0	200	cells x 10^9/L
NIHR-HIC-ICU-0183	Neutrophil count	real-2d	0	150	cells x 10^9/L
NIHR-HIC-ICU-0184	Platelets	real-2d	0	1,500	cells x 10^3/L
NIHR-HIC-ICU-0187	Organism	string-2d	-	-	-
NIHR-HIC-ICU-0242	Fentanyl	real-2d	0	1,000	micrograms/hour
NIHR-HIC-ICU-0252	milrinone	real-2d	0	20	-
NIHR-HIC-ICU-0395	CCU bed configuration 03	string-1d	-	-	-

NIHR-HIC-ICU-0396	CCU bed configuration 05	string-1d	-	-	-
NIHR-HIC-ICU-0397	CCU bed configuration 90	string-1d	-	-	-
NIHR-HIC-ICU-0398	Admission type	string-1d	-	-	-
NIHR-HIC-ICU-0399	Primary reason for admission to your unit	string-1d	-	-	-
NIHR-HIC-ICU-0400	Brain stem death declared	integer-1d	0	1	-
NIHR-HIC-ICU-0405	Timeliness of discharge from your unit	string-1d	-	-	-
NIHR-HIC-ICU-0406	Date of discharge from your hospital	date-1d	-	-	-
NIHR-HIC-ICU-0407	Date of first critical care post-discharge from your unit	date-1d	-	-	-
NIHR-HIC-ICU-0408	Date of ultimate discharge from your hospital	date-1d	-	-	-
NIHR-HIC-ICU-0409	APACHE II Score	integer-1d	0	71	-
NIHR-HIC-ICU-0410	APACHE II Probability	real-1d	0	100	-
NIHR-HIC-ICU-0411	Date & Time of admission to your unit	datetime-1d	-	-	-
NIHR-HIC-ICU-0412	Date & Time of discharge from your unit	datetime-1d	-	-	-
NIHR-HIC-ICU-0413	Fluid Balance (hourly)	integer-2d	$-2,500$	$10,000$	mL/hour
NIHR-HIC-ICU-0414	Amikacin	integer-2d	0	$4,500$	mg

NIHR-HIC-ICU-0415	Amoxicillin	integer-2d	0	2,000	mg
NIHR-HIC-ICU-0416	Azithromycin	real-2d	0	2,000	mg
NIHR-HIC-ICU-0417	Benzylpenicillin	integer-2d	0	3,000	mg
NIHR-HIC-ICU-0418	Cefotaxime	real-2d	0	4,000	mg
NIHR-HIC-ICU-0419	Ceftazidime	integer-2d	0	4,000	mg
NIHR-HIC-ICU-0420	Ceftriaxone	real-2d	0	4,000	mg
NIHR-HIC-ICU-0421	Cefuroxime	integer-2d	0	3,000	mg
NIHR-HIC-ICU-0422	Chloramphenicol	real-2d	0	5,000	mg
NIHR-HIC-ICU-0423	Ciprofloxacin	integer-2d	0	3,000	mg
NIHR-HIC-ICU-0424	Clarithromycin	integer-2d	0	4,000	mg
NIHR-HIC-ICU-0425	Clindamycin	integer-2d	0	4,000	mg
NIHR-HIC-ICU-0426	Co-Amoxiclav	integer-2d	0	2,400	mg
NIHR-HIC-ICU-0427	Colistin	real-2d	0	10	Millions of units
NIHR-HIC-ICU-0428	Co-Trimoxazole	integer-2d	0	18,000	mg
NIHR-HIC-ICU-0429	Demeclocycline HCL	integer-2d	0	5,000	mg
NIHR-HIC-ICU-0430	Doxycycline	integer-2d	0	800	mg
NIHR-HIC-ICU-0432	Ertapenem	real-2d	0	2,000	mg
NIHR-HIC-ICU-0433	Erythromycin	integer-2d	0	1,000	mg
NIHR-HIC-ICU-0434	Ethambutal HCL	integer-2d	0	5,000	mg
NIHR-HIC-ICU-0435	Flucloxacillin	integer-2d	0	2,000	mg
NIHR-HIC-ICU-0436	Fuscidic acid	integer-2d	0	2,000	mg
NIHR-HIC-ICU-0437	Gentamicin	integer-2d	0	1,000	mg

NIHR-HIC-ICU-0438	Isoniazid	integer-2d	0	900	mg
NIHR-HIC-ICU-0439	Levofloxacin	integer-2d	0	1,000	mg
NIHR-HIC-ICU-0440	Linezolid	integer-2d	0	1,200	mg
NIHR-HIC-ICU-0441	Meropenem	real-2d	0	4,000	mg
NIHR-HIC-ICU-0442	Metronidazole	integer-2d	0	2,000	mg
NIHR-HIC-ICU-0443	Moxifloxacin	integer-2d	0	800	mg
NIHR-HIC-ICU-0444	Neomycin	real-2d	0	3,000	mg
NIHR-HIC-ICU-0445	Nitrofurantion	integer-2d	0	400	mg
NIHR-HIC-ICU-0446	Ofloxacin	integer-2d	0	800	mg
NIHR-HIC-ICU-0447	Pentamidine	integer-2d	0	600	mg
NIHR-HIC-ICU-0448	Phenoxymethylpenicillin	integer-2d	0	1,000	mg
NIHR-HIC-ICU-0449	Piperacillin/Tazobactam	real-2d	0	9	g
NIHR-HIC-ICU-0450	Pyrazinamide	integer-2d	0	5,000	mg
NIHR-HIC-ICU-0452	Rifampacin	integer-2d	0	1,800	mg
NIHR-HIC-ICU-0453	Rifater	integer-2d	0	-	tablets
NIHR-HIC-ICU-0454	Rifinah	integer-2d	0	-	tablets
NIHR-HIC-ICU-0456	Sodium Fusidate	integer-2d	0	2,000	mg
NIHR-HIC-ICU-0457	Teicoplanin	integer-2d	0	2,000	mg
NIHR-HIC-ICU-0458	Tigecycline	integer-2d	0	200	mg
NIHR-HIC-ICU-0459	Tobramycin	integer-2d	0	900	mg
NIHR-HIC-ICU-0460	Trimethoprim	integer-2d	0	1,500	mg
NIHR-HIC-ICU-0461	Vancomycin	real-2d	0	3,000	mg

NIHR-HIC-ICU-0462	Propofol	real-2d	0	1,000	mg/hour
NIHR-HIC-ICU-0463	Midazolam	real-2d	0	30	mg/hour
NIHR-HIC-ICU-0464	Remifentanil	real-2d	0	2	micrograms/Kg/hour
NIHR-HIC-ICU-0465	Adrenaline	real-2d	0	4	micrograms/Kg/min
NIHR-HIC-ICU-0466	Dobutamine	real-2d	0	20	micrograms/Kg/min
NIHR-HIC-ICU-0467	Dopamine	real-2d	0	200	micrograms/hour
NIHR-HIC-ICU-0468	Enoximone	real-2d	0	30	micrograms/hour
NIHR-HIC-ICU-0469	Levosimendan	real-2d	0	0	micrograms/hour
NIHR-HIC-ICU-0470	Noradrenaline	real-2d	0	4	micrograms/Kg/min
NIHR-HIC-ICU-0471	Vasopressin	real-2d	0	40	micrograms/hour
NIHR-HIC-ICU-0549	Spontaneous Respiratory Rate	integer-2d	0	60	-
NIHR-HIC-ICU-0550	Tidal volume	integer-2d	0	5,000	mL
NIHR-HIC-ICU-0552	Duration of therapy (hours per day)	real-2d	0	24	hours/day
NIHR-HIC-ICU-0553	Total effluent per day	integer-2d	0	1,000	L/day
NIHR-HIC-ICU-0554	Dialysate	integer-2d	0	200,000	L
NIHR-HIC-ICU-0555	Replacement fluid during RRT	integer-2d	-	-	-
NIHR-HIC-ICU-0556	Type of anticoagulation	integer-2d	1	4	-
NIHR-HIC-ICU-0557	C reactive protein	real-2d	0	1,000	mg/L
NIHR-HIC-ICU-0558	Thiopentone / Thiopental	real-2d	0	1,000	mg/hour
NIHR-HIC-ICU-0559	Clonidine	real-2d	0	200	micrograms/hour
NIHR-HIC-ICU-0560	Dexmedetomidine	real-2d	0	1,000	micrograms/hour
NIHR-HIC-ICU-0561	Ketamine	real-2d	0	300	mg

NIHR-HIC-ICU-0563	Morphine	real-2d	0	45	mg
NIHR-HIC-ICU-0564	dopexamine	real-2d	0	1,000	-
NIHR-HIC-ICU-0565	Terlipressin	real-2d	0	1,000	-
NIHR-HIC-ICU-0573	Destination post discharge within your hospital	string-1d	-	-	-
NIHR-HIC-ICU-0906	Esmolol	real-2d	0	1,000	-
NIHR-HIC-ICU-0907	Metoprolol	real-2d	0	1,000	-
NIHR-HIC-ICU-0908	Dexamethasone	real-2d	0	100	-
NIHR-HIC-ICU-0909	Hydrocortisone	real-2d	0	2,000	-
NIHR-HIC-ICU-0910	Methylprednisolone	real-2d	0	1,000	-
NIHR-HIC-ICU-0911	Sedation yes/no	integer-2d	0	1	-
NIHR-HIC-ICU-0912	Ultimate primary reason for admission to unit	string-1d	-	-	-
NIHR-HIC-ICU-0913	PaO2/FiO2 ratio	real-2d	0	100	kPa
NIHR-HIC-ICU-0915	Fluid Balance (daily)	integer-2d	−20	20	L
NIHR-HIC-ICU-0918	Glucose (laboratory)	real-2d	0	60	mmol/L
NIHR-HIC-ICU-0930	Dead or alive on discharge	string-1d	-	-	-
NIHR-HIC-ICU-0931	Advanced respiratory support	integer-2d	0	1	-
NIHR-HIC-ICU-0932	Basic respiratory support	integer-2d	0	1	-
NIHR-HIC-ICU-0933	Advanced Cardiovascular support	integer-2d	0	1	-
NIHR-HIC-ICU-0934	Basic Cardiovascular support	integer-2d	0	1	-
NIHR-HIC-ICU-0935	Renal support	integer-2d	0	1	-

NIHR-HIC-ICU-0936	Neurological support	integer-2d	0	1	-
NIHR-HIC-ICU-0937	Liver support	integer-2d	0	1	-
NIHR-HIC-ICU-0938	Dermatological support	integer-2d	0	1	-
NIHR-HIC-ICU-0939	Gastrointestinal support	integer-2d	0	1	-

Table A.1: CC-HIC data specification

	log hazard ratio			hazard ratio				
parameter	Coef	L95	U95	Coef	L95	U95	z	p
bilirubin - severity								
age	0.43	0.34	0.51	1.53	1.41	1.67	9.82	9.36×10^{-23}
male sex	−0.14	−0.29	0.01	0.87	0.75	1.01	−1.85	6.41×10^{-2}
weight	−0.06	−0.14	0.01	0.94	0.87	1.01	−1.59	1.12×10^{-1}
CPR	0.16	−0.14	0.46	1.17	0.87	1.58	1.05	2.92×10^{-1}
cormorbidities	0.64	0.49	0.79	1.90	1.64	2.20	8.56	1.16×10^{-17}
dependencies	0.42	0.27	0.57	1.52	1.31	1.77	5.42	5.89×10^{-8}
baseline SOFA	0.20	0.12	0.27	1.22	1.12	1.31	4.87	1.09×10^{-6}
severity	0.10	0.08	0.12	1.11	1.09	1.12	12.37	3.72×10^{-35}
bilirubin - trajectory								
age	0.42	0.34	0.51	1.53	1.40	1.66	9.68	3.70×10^{-22}
male sex	−0.13	−0.28	0.01	0.87	0.75	1.01	−1.77	7.61×10^{-2}
weight	−0.07	−0.15	0.01	0.93	0.86	1.01	−1.80	7.24×10^{-2}
CPR	0.10	−0.20	0.40	1.11	0.82	1.49	0.67	5.05×10^{-1}
cormorbidities	0.61	0.47	0.76	1.85	1.59	2.14	8.15	3.57×10^{-16}
dependencies	0.42	0.27	0.57	1.52	1.31	1.77	5.44	5.21×10^{-8}
baseline SOFA	0.18	0.11	0.26	1.20	1.11	1.30	4.59	4.49×10^{-6}
severity	0.09	0.07	0.10	1.09	1.07	1.11	9.97	1.97×10^{-23}
velocity	0.73	0.48	0.98	2.08	1.62	2.67	5.77	8.01×10^{-9}
bilirubin - cumulative effect								
age	0.38	0.30	0.46	1.46	1.34	1.59	8.91	4.93×10^{-19}
male sex	−0.15	−0.30	−0.01	0.86	0.74	0.99	−2.07	3.85×10^{-2}
weight	−0.07	−0.14	0.01	0.94	0.87	1.01	−1.65	9.88×10^{-2}
CPR	0.16	−0.14	0.45	1.17	0.87	1.57	1.03	3.04×10^{-1}
cormorbidities	0.70	0.56	0.85	2.02	1.75	2.33	9.57	1.11×10^{-21}
dependencies	0.38	0.23	0.53	1.46	1.25	1.69	4.91	8.94×10^{-7}
baseline SOFA	0.28	0.20	0.35	1.32	1.22	1.43	7.20	6.00×10^{-13}
cumulative effect	0.01	0.00	0.01	1.01	1.00	1.01	7.23	4.68×10^{-13}
CRP - severity								
age	0.35	0.27	0.43	1.42	1.30	1.54	8.24	1.68×10^{-16}
male sex	−0.15	−0.30	−0.01	0.86	0.74	0.99	−2.05	4.05×10^{-2}

weight	−0.10	−0.18	−0.03	0.90	0.83	0.97	−2.59	9.64×10^{-3}
CPR	0.17	−0.12	0.47	1.19	0.89	1.60	1.15	2.50×10^{-1}
cormorbidities	0.73	0.58	0.87	2.07	1.79	2.39	9.86	6.09×10^{-23}
dependencies	0.43	0.28	0.58	1.53	1.32	1.78	5.55	2.84×10^{-8}
baseline SOFA	0.29	0.22	0.37	1.34	1.24	1.44	7.73	1.07×10^{-14}
severity	0.09	0.07	0.11	1.09	1.07	1.12	9.03	1.71×10^{-19}
CRP - trajectory								
age	0.35	0.27	0.44	1.42	1.31	1.55	8.28	1.26×10^{-16}
male sex	−0.17	−0.32	−0.02	0.84	0.73	0.98	−2.22	2.64×10^{-2}
weight	−0.10	−0.18	−0.02	0.90	0.83	0.98	−2.48	1.31×10^{-2}
CPR	0.11	−0.19	0.41	1.12	0.83	1.51	0.72	4.74×10^{-1}
cormorbidities	0.65	0.50	0.80	1.92	1.66	2.22	8.66	4.83×10^{-18}
dependencies	0.46	0.30	0.61	1.58	1.35	1.84	5.83	5.62×10^{-9}
baseline SOFA	0.26	0.19	0.34	1.30	1.21	1.40	6.90	5.33×10^{-12}
severity	0.11	0.09	0.13	1.12	1.09	1.14	11.10	1.20×10^{-28}
velocity	0.85	0.67	1.03	2.34	1.96	2.79	9.40	5.58×10^{-21}
CRP - cumulative effect								
age	0.35	0.26	0.43	1.41	1.30	1.54	8.26	1.49×10^{-16}
male sex	−0.18	−0.33	−0.03	0.83	0.72	0.97	−2.43	1.51×10^{-2}
weight	−0.08	−0.16	−0.01	0.92	0.85	0.99	−2.10	3.53×10^{-2}
CPR	0.16	−0.14	0.46	1.17	0.87	1.58	1.05	2.94×10^{-1}
cormorbidities	0.73	0.58	0.87	2.07	1.79	2.39	9.95	2.64×10^{-23}
dependencies	0.34	0.19	0.49	1.41	1.21	1.64	4.45	8.75×10^{-6}
baseline SOFA	0.32	0.24	0.39	1.37	1.28	1.48	8.44	3.27×10^{-17}
cumulative effect	0.00	0.00	0.01	1.00	1.00	1.01	3.25	1.14×10^{-3}
GCS - severity								
age	0.41	0.33	0.50	1.51	1.39	1.65	9.50	2.01×10^{-21}
male sex	−0.08	−0.23	0.07	0.92	0.79	1.07	−1.10	2.70×10^{-1}
weight	−0.10	−0.18	−0.02	0.91	0.84	0.99	−2.32	2.01×10^{-2}
CPR	−0.07	−0.38	0.23	0.93	0.69	1.26	−0.47	6.37×10^{-1}
cormorbidities	0.74	0.60	0.89	2.10	1.82	2.44	9.96	2.37×10^{-23}
dependencies	0.37	0.22	0.52	1.44	1.24	1.68	4.74	2.19×10^{-6}
baseline SOFA	0.06	−0.02	0.15	1.07	0.98	1.16	1.54	1.25×10^{-1}
severity	−0.16	−0.18	−0.14	0.85	0.83	0.87	−14.37	8.19×10^{-47}

GCS - trajectory								
age	0.42	0.33	0.50	1.52	1.39	1.66	9.22	2.91×10^{-20}
male sex	−0.09	−0.25	0.06	0.91	0.78	1.06	−1.19	2.32×10^{-1}
weight	−0.09	−0.17	−0.00	0.92	0.84	1.00	−2.02	4.33×10^{-2}
CPR	−0.03	−0.34	0.29	0.97	0.71	1.33	−0.17	8.63×10^{-1}
cormorbidities	0.51	0.35	0.67	1.67	1.42	1.96	6.35	2.16×10^{-10}
dependencies	0.35	0.19	0.51	1.41	1.20	1.66	4.25	2.13×10^{-5}
baseline SOFA	0.12	0.04	0.20	1.13	1.04	1.23	2.78	5.42×10^{-3}
severity	−0.17	−0.20	−0.15	0.84	0.82	0.86	−15.04	4.27×10^{-51}
velocity	−1.30	−1.50	−1.11	0.27	0.22	0.33	−13.32	1.73×10^{-40}
GCS - cumulative effect								
age	0.39	0.31	0.48	1.48	1.36	1.61	9.10	8.76×10^{-20}
male sex	−0.12	−0.27	0.02	0.88	0.76	1.02	−1.65	9.98×10^{-2}
weight	−0.09	−0.17	−0.01	0.91	0.84	0.99	−2.31	2.07×10^{-2}
CPR	0.07	−0.23	0.37	1.07	0.79	1.45	0.45	6.52×10^{-1}
cormorbidities	0.74	0.60	0.89	2.10	1.81	2.43	9.99	1.65×10^{-23}
dependencies	0.34	0.19	0.49	1.40	1.20	1.63	4.35	1.34×10^{-5}
baseline SOFA	0.24	0.16	0.32	1.27	1.17	1.37	5.96	2.50×10^{-9}
cumulative effect	−0.01	−0.01	−0.01	0.99	0.99	0.99	−8.01	1.11×10^{-15}
PF - severity								
age	0.35	0.27	0.44	1.43	1.31	1.56	8.00	1.23×10^{-15}
male sex	−0.15	−0.30	0.01	0.86	0.74	1.01	−1.89	5.82×10^{-2}
weight	−0.09	−0.17	−0.01	0.91	0.84	0.99	−2.21	2.69×10^{-2}
CPR	0.28	−0.02	0.59	1.33	0.98	1.80	1.83	6.77×10^{-2}
cormorbidities	0.68	0.53	0.83	1.97	1.69	2.28	8.89	6.24×10^{-19}
dependencies	0.43	0.27	0.58	1.53	1.32	1.79	5.44	5.21×10^{-8}
baseline SOFA	0.32	0.24	0.39	1.38	1.28	1.48	8.39	4.73×10^{-17}
severity	−0.94	−1.07	−0.82	0.39	0.34	0.44	−14.72	4.48×10^{-49}
PF - trajectory								
age	0.36	0.27	0.45	1.43	1.31	1.56	7.75	9.51×10^{-15}
male sex	−0.15	−0.31	0.01	0.86	0.74	1.01	−1.82	6.87×10^{-2}
weight	−0.08	−0.16	0.01	0.93	0.85	1.01	−1.78	7.47×10^{-2}
CPR	0.40	0.08	0.72	1.50	1.09	2.06	2.48	1.33×10^{-2}
cormorbidities	0.64	0.48	0.80	1.90	1.62	2.22	7.94	2.08×10^{-15}

dependencies	0.41	0.25	0.57	1.51	1.28	1.78	4.96	6.88×10^{-7}
baseline SOFA	0.32	0.24	0.40	1.37	1.27	1.49	7.91	2.64×10^{-15}
severity	−1.01	−1.14	−0.88	0.36	0.32	0.41	−15.30	7.75×10^{-53}
velocity	−6.44	−7.91	−4.96	0.00	0.00	0.01	−8.56	1.12×10^{-17}
PF - cumulative effect								
age	0.35	0.27	0.44	1.43	1.31	1.55	8.11	4.93×10^{-16}
male sex	−0.15	−0.30	0.00	0.86	0.74	1.00	−1.94	5.23×10^{-2}
weight	−0.10	−0.18	−0.02	0.90	0.83	0.98	−2.47	1.35×10^{-2}
CPR	0.23	−0.08	0.53	1.25	0.93	1.70	1.47	1.43×10^{-1}
cormorbidities	0.72	0.57	0.87	2.05	1.77	2.38	9.55	1.25×10^{-21}
dependencies	0.40	0.24	0.55	1.49	1.28	1.73	5.06	4.23×10^{-7}
baseline SOFA	0.35	0.28	0.43	1.42	1.32	1.53	9.17	4.59×10^{-20}
cumulative effect	−0.08	−0.09	−0.07	0.92	0.91	0.94	−11.53	9.65×10^{-31}
platelets - severity								
age	0.41	0.33	0.50	1.51	1.39	1.64	9.59	9.11×10^{-22}
male sex	−0.10	−0.24	0.05	0.91	0.78	1.05	−1.28	2.01×10^{-1}
weight	−0.06	−0.14	0.02	0.94	0.87	1.02	−1.53	1.26×10^{-1}
CPR	0.27	−0.03	0.56	1.30	0.97	1.75	1.78	7.53×10^{-2}
cormorbidities	0.37	0.22	0.53	1.45	1.24	1.70	4.70	2.65×10^{-6}
dependencies	0.40	0.25	0.56	1.50	1.29	1.74	5.29	1.25×10^{-7}
baseline SOFA	0.16	0.08	0.23	1.17	1.08	1.26	3.93	8.50×10^{-5}
severity	−0.11	−0.13	−0.10	0.89	0.88	0.91	−13.04	7.66×10^{-39}
platelets - trajectory								
age	0.40	0.32	0.49	1.50	1.37	1.63	9.16	5.42×10^{-20}
male sex	−0.04	−0.19	0.11	0.96	0.83	1.12	−0.53	5.94×10^{-1}
weight	−0.06	−0.14	0.02	0.94	0.87	1.02	−1.46	1.43×10^{-1}
CPR	0.29	−0.01	0.59	1.34	0.99	1.80	1.89	5.83×10^{-2}
cormorbidities	0.26	0.10	0.42	1.29	1.10	1.52	3.17	1.54×10^{-3}
dependencies	0.41	0.26	0.56	1.51	1.29	1.75	5.26	1.46×10^{-7}
baseline SOFA	0.14	0.06	0.22	1.15	1.06	1.24	3.50	4.66×10^{-4}
severity	−0.12	−0.14	−0.11	0.88	0.87	0.90	−13.72	7.46×10^{-43}
velocity	−1.01	−1.20	−0.83	0.36	0.30	0.43	−10.97	5.51×10^{-28}
platelets - cumulative effect								
age	0.38	0.30	0.46	1.46	1.35	1.59	8.95	3.68×10^{-19}

male sex	−0.13	−0.28	0.01	0.88	0.76	1.01	−1.78	7.48×10^{-2}
weight	−0.07	−0.15	0.01	0.93	0.86	1.01	−1.74	8.10×10^{-2}
CPR	0.21	−0.08	0.51	1.24	0.92	1.66	1.43	1.54×10^{-1}
cormorbidities	0.56	0.42	0.71	1.76	1.51	2.04	7.43	1.11×10^{-13}
dependencies	0.35	0.20	0.50	1.42	1.22	1.65	4.60	4.23×10^{-6}
baseline SOFA	0.27	0.19	0.34	1.31	1.21	1.41	7.00	2.52×10^{-12}
cumulative effect	−0.01	−0.01	−0.00	0.99	0.99	1.00	−8.53	1.47×10^{-17}
SOFA - severity								
age	0.44	0.35	0.52	1.55	1.42	1.68	9.95	2.40×10^{-23}
male sex	−0.18	−0.32	−0.03	0.84	0.72	0.97	−2.38	1.72×10^{-2}
weight	−0.11	−0.19	−0.03	0.90	0.83	0.97	−2.75	6.04×10^{-3}
CPR	0.02	−0.28	0.31	1.02	0.76	1.36	0.12	9.08×10^{-1}
cormorbidities	0.48	0.33	0.63	1.61	1.39	1.87	6.36	2.07×10^{-10}
dependencies	0.45	0.30	0.60	1.56	1.35	1.81	5.85	4.82×10^{-9}
severity	0.21	0.19	0.23	1.24	1.21	1.26	21.69	2.66×10^{-104}
SOFA - trajectory								
age	0.45	0.36	0.54	1.57	1.43	1.71	10.00	1.58×10^{-23}
male sex	−0.17	−0.32	−0.02	0.85	0.73	0.98	−2.17	3.00×10^{-2}
weight	−0.10	−0.18	−0.02	0.90	0.83	0.98	−2.50	1.24×10^{-2}
CPR	0.05	−0.25	0.35	1.05	0.78	1.42	0.33	7.41×10^{-1}
cormorbidities	0.37	0.22	0.53	1.45	1.25	1.69	4.78	1.73×10^{-6}
dependencies	0.48	0.33	0.63	1.62	1.39	1.88	6.12	9.42×10^{-10}
severity	0.21	0.19	0.23	1.23	1.21	1.25	20.75	1.34×10^{-95}
velocity	0.95	0.71	1.19	2.60	2.04	3.30	7.80	5.96×10^{-15}
SOFA - cumulative effect								
age	0.36	0.28	0.44	1.43	1.32	1.55	8.56	1.13×10^{-17}
male sex	−0.14	−0.28	0.01	0.87	0.75	1.01	−1.88	5.98×10^{-2}
weight	−0.08	−0.16	−0.00	0.92	0.85	1.00	−2.04	4.11×10^{-2}
CPR	0.22	−0.07	0.52	1.25	0.93	1.68	1.50	1.32×10^{-1}
cormorbidities	0.66	0.52	0.80	1.93	1.68	2.23	9.01	1.99×10^{-19}
dependencies	0.37	0.22	0.52	1.45	1.25	1.68	4.90	9.43×10^{-7}
cumulative effect	0.01	0.01	0.01	1.01	1.01	1.01	13.39	6.86×10^{-41}

Table A.2: All coefficients for univariate joint models. Findings are grouped by biomarker and morphological parameterisation.

Appendix B

Search Terms for Literature Review

The following search parameters were used when performing a literature review (note, all parameters for a given topic are used with boolean "OR" clauses):

- Sepsis:
  - sepsis[Title]
  - septic[Title]
  - Sepsis[MeSH]
  - septic shock[Title]
- Hyperoxaemia:
  - hyperoxia[Title/Abstract]
  - hyperoxemic[Title/Abstract]
  - excessive[Title/Abstract]
  - excess[Title/Abstract]
  - unnecessary[Title/Abstract]
- Outcomes:
  - outcome(s)[Title/Abstract]
  - morbidity[Title/Abstract]
  - mortality[Title/Abstract]
  - harm[Title/Abstract]
- Organ dysfunction:

  - sofa[Title]
  - organ failure assessment[Title/Abstract]
  - organ failure[Title]

- Critical Care:
  - critical care[MeSH Term]
  - critical care[Title/Abstract]
  - intensive care[All fields]
  - critical illness[MeSH Term]
  - critical illness[Title/Abstract]
  - intensive care units[MeSH Term])

- Trajectories:
  - trajectory[Title/Abstract]
  - course[Title/Abstract]
  - natural history[Title]
  - time series[Title]
  - longitudinal[Title]
  - profile[Title]

- Phenotype:
  - phenotype[Title/Abstract]
  - phenotype[MeSH Terms]
  - endotype[Title/Abstract]
  - subgroup[Title]
  - subtype[Title]
  - classification[Title]
  - latent[Title]
  - heterogeneity[Title/Abstract]
  - cluster[Title]

- Joint Models:
  - joint model[Title/Abstract]

  - joint modelling[Title/Abstract]
  - shared parameter[Title/Abstract]

Initial searches were conducted in pubmed and google scholar on 31st October 2018 and updated on 23rd March 2020. Searches were restricted to adult humans with the full text available in English. References of relevant papers were searched for any missed papers. Google citation alerts were set up for key authors in the relevant fields.

Initial search strategies:

- "oxygen" AND hyperoxaemia AND critical care AND outcomes
- Sepsis AND organ failure AND trajectories
- sepsis AND trajectories AND phenotype
- joint models AND (sepsis OR critical care)

Appendix C

Software Vignettes

C.1 Data Quality Evaluation with inspectEHR

inspectEHR [6] applies the Kahn data quality evaluation framework to the CC-HIC research database and persists the findings as meta-data alongside the primary research data. It also produces a large number of diagnostic plots to explore each contributed data concept, which are stored in a user defined location external to the research database. inspectEHR [6] follows an extract, evaluate and export paradigm. The basic approach is that data concepts are extracted from the CC-HIC database, evaluated for particular qualities (for example, that they are within appropriate reference ranges) and any violations are converted to a predefined format, and exported to the database. After performing the full evaluation, any data concept that *does not* have a corresponding row in a data quality table, can therefore be safely assumed to have passed all data quality evaluation procedures.

Usage

```
library(inspectEHR)


# Establish a database connection
db_pth <- system.file("path_to_database")
ctn <- DBI::dbConnect(RSQLite::SQLite(), db_pth)
```

Optionally, you may wish to obscure the names of contributing sites by providing translation lookups.

```
translate_site <- tibble::tibble(
  site = c("St. Elsewhere", "Royal Other"),
  translation = LETTERS[1:2])
```

Choose an output folder. This will be where plots are exported for inspection. Then you can run inspectEHR.

```
output_folder <- "~/documents/cchic/eval/"


perform_evaluation(connection = ctn,
                   output_folder = output_folder,
                   translate_site = translate_site,
                   verbose = TRUE)
```

C.2 Extracting data with wranglEHR

A common statistical work flow (and the one adopted in this thesis) is to represent data in the so-called "tidy" format [185], that is, a rectangular format with the following specification:

- one row per statistical unit
- one column per unique variable

In our case, the statistical unit is either an ICU episode (where time invariant data are concerned), or a period of time for each patient (often 30 minutes or an hour)[1]. wranglEHR exposes two main functions to the end user:

1. `extract_demographics()`
2. `extract_timevarying()`

Both allow flexible data extraction from the CC-HIC database, and reconcile the data into the appropriate format for analysis. Any accompanying meta-data are identified and arranged into appropriately labelled columns. The extraction process can be customised to suite a specific case use, including:

[1]The underlying EHR rarely stores data more often than at 5 minute intervals, and so this is a reasonable expected upper boundary. At higher temporal resolutions, the methods described here are likely to fail

- Setting the desired temporal cadence of the table (i.e. one row per hour, versus one row per day)
- Defining a custom, possibly user specified, action if the data storage resolution is higher than the target row cadence

Installation

```
# install directly from github with
remotes::install_github("DocEd/wranglEHR")
```

A copy should already be installed into the group library for the CC-HIC team inside the UCL safe haven. If you are having problems with this, please contact me directly. The package is now in a stable state, and so changes to the interface are unlikely. Please do ensure you capture the version number or git commit hash to ensure reproducibility of your pipeline.

Usage

```
library(wranglEHR)


# Establish a database connection
db_pth <- system.file("path_to_database")
ctn <- DBI::dbConnect(RSQLite::SQLite(), db_pth)
```

We can extract demographic data by specifying the `code_name` of interest.

```
# Extract static variables. Rename on the fly.
dtb <- extract_demographics(
  connection = ctn,
  episode_ids = 13639:13643,
  code_names = c("NIHR_HIC_ICU_0017", "NIHR_HIC_ICU_0019"),
  rename = c("height", "weight")
)
```

Flexible extraction of longitudinal data is possible.

```
# Extract time varying variables. Rename on the fly.
ltb <- extract_timevarying(
  connection = ctn,
  code_names = "NIHR_HIC_ICU_0108",
  rename = "hr")
```

We can perform for complex and parameterised extractions:

- Set the base cadence for 2 hours
- Automatically handle metadata
- Limit the time boundaries of the cohort
- Supply a user defined summary function to handle realignment of the data to the new cadence:
  - Must return a vector of length 1
  - Must return in consistent data type
  - Must be able to handle the variable appearance of NAs

```
summary_mean <- function(x) {
  if (all(is.na(x))) {
    return(x[1])
  } else {
    mean(x, na.rm = TRUE)
  }
}


ltb_2 <- extract_timevarying(
  connection = ctn,
  code_names = "NIHR_HIC_ICU_0108", "NIHR_HIC_ICU_0116",
  rename = "hr", "cvp",
  cadence = 2, # 1 row every 2 hours
  coalesce_rows = summary_mean,
  time_boundaries = c(0, 6) # first 6 hours only
```

```
)

DBI::dbDisconnect(ctn)
```

Getting help

If you find a bug, please file a minimal reproducible example on github at https://github.com/DocEd/wranglEHR/issues

Appendix D

Colophon

This document was set in the Times Roman typeface using LaTeX and BibTeX. It was composed with the Atom and R studio integrated developer environment. Plotting was performed with the GGplot2 package for R.

Bibliography

[1] The CC-HIC. The Critical Care Health Informatics Collaborative. https://hic.nihr.ac.uk/critical+care. Last accessed 23-06-2021, 2016.

[2] The DECOVID Consortium. DECOVID. https://www.decovid.org. Last Accessed 23-06-2021., 2020.

[3] Matt Hancock. Driving digital in the NHS. https://www.gov.uk/government/speeches/driving-digital-in-the-nhs. Accessed, March 2021.

[4] Edward Palmer, Benjamin Post, Roman Klapaukh, Giampiero Marra, Niall S. MacCallum, David Brealey, Ari Ercole, Andrew Jones, Simon Ashworth, Peter Watkinson, Richard Beale, Stephen J Brett, J. Duncan Young, Claire Black, Aasiyah Rashan, Daniel Martin, Mervyn Singer, and Steve Harris. The Association Between Supra-Physiologic Arterial Oxygen Levels and Mortality in Critically Ill Patients: A Multi-Centre Observational Cohort Study. *American Journal of Respiratory and Critical Care Medicine*, pages rccm.201904–0849OC, September 2019.

[5] Edward Palmer. wranglEHR: Standardised data extraction for CC-HIC. www.github.com/DocEd/wranglEHR. Version 1., 2020.

[6] Edward Palmer. inspectEHR: Standardised data quality evaluation for CC-HIC. www.github.com/DocEd/inspectEHR. Version 1., 2020.

[7] MIT Critical Data. *Secondary Analysis of Electronic Health Records*. Springer International Publishing : Imprint: Springer, Cham, 1st ed. 2016 edition, 2016.

[8] Richard Williams, Evangelos Kontopantelis, Iain Buchan, and Niels Peek. Clinical code set engineering for reusing EHR data for research: A review. *Journal of Biomedical Informatics*, 70:1–13, June 2017.

[9] George Hripcsak, Charles Knirsch, Li Zhou, Adam Wilcox, and Genevieve Melton. Bias Associated with Mining Electronic Health Records. *Journal of Biomedical Discovery and Collaboration*, 6:48–52, 2011.

[10] Lee Jacobs. Interview with Lawrence Weed, MD– The Father of the Problem-Oriented Medical Record Looks Ahead. *The Permanente Journal*, 13(3), July 2009.

[11] David Harrison. Case Mix Programme dataset update. *ICNARC*, page 26, 2019.

[12] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, May 2016.

[13] Alistair E.W. Johnson, David J. Stone, Leo A. Celi, and Tom J. Pollard. The MIMIC Code Repository: Enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018.

[14] Steve Harris, Sinan Shi, David Brealey, Niall S. MacCallum, Spiros Denaxas, David Perez-Suarez, Ari Ercole, Peter Watkinson, Andrew Jones, Simon Ashworth, Richard Beale, Duncan Young, Stephen Brett, and Mervyn Singer. Critical Care Health Informatics Collaborative (CCHIC): Data, tools and methods for reproducible research: A multi-centre UK intensive care

database. *International Journal of Medical Informatics*, 112:82–89, April 2018.

[15] Jeffrey G. Klann, Aaron Abend, Vijay A. Raghavan, Kenneth D. Mandl, and Shawn N. Murphy. Data interchange using i2b2. *Journal of the American Medical Informatics Association: JAMIA*, 23(5):909–915, September 2016.

[16] S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. C. Chueh, S. Churchill, and I. Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130, March 2010.

[17] J. Marc Overhage, Patrick B. Ryan, Christian G. Reich, Abraham G. Hartzema, and Paul E. Stang. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association: JAMIA*, 19(1):54–60, 2012 Jan-Feb.

[18] Erica A. Voss, Rupa Makadia, Amy Matcho, Qianli Ma, Chris Knoll, Martijn Schuemie, Frank J. DeFalco, Ajit Londhe, Vivienne Zhu, and Patrick B. Ryan. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association: JAMIA*, 22(3):553–564, May 2015.

[19] Bruce M. Psaty and Alasdair M. Breckenridge. Mini-Sentinel and regulatory science–big data rendered fit and functional. *The New England Journal of Medicine*, 370(23):2165–2167, June 2014.

[20] Lesley H. Curtis, Mark G. Weiner, Denise M. Boudreau, William O. Cooper, Gregory W. Daniel, Vinit P. Nair, Marsha A. Raebel, Nicolas U. Beaulieu, Robert Rosofsky, Tiffany S. Woodworth, and Jeffrey S. Brown. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiology and Drug Safety*, 21 Suppl 1:23–31, January 2012.

[21] Rachael L. Fleurence, Lesley H. Curtis, Robert M. Califf, Richard Platt, Joe V. Selby, and Jeffrey S. Brown. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association: JAMIA*, 21(4):578–582, 2014 Jul-Aug.

[22] Mark D. Danese, Marc Halperin, Jennifer Duryea, and Ryan Duryea. The Generalized Data Model for clinical research. *BMC medical informatics and decision making*, 19(1):117, June 2019.

[23] NIHR HIC. The NIHR Health Informatics Collaborative. https://hic.nihr.ac.uk/about?page_id=24. Last Accessed 23-06-2021, July 2020.

[24] David a Harrison, Anthony R Brady, and Kathy Rowan. Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: The Intensive Care National Audit & Research Centre Case Mix Programme Database. *Critical care (London, England)*, 8(2):R99–111, 2004.

[25] K. Rowan. Are scoring systems adequate indicators for quality and performance of the ICU? *Anasthesiologie, Intensivmedizin, Notfallmedizin, Schmerztherapie: AINS*, 33(1):52–55, January 1998.

[26] N. F. de Keizer, G. J. Bonsel, C. Goldfad, and K. M. Rowan. The added value that increasing levels of diagnostic information provide in prognostic models to estimate hospital mortality for adult intensive care patients. *Intensive Care Medicine*, 26(5):577–584, May 2000.

[27] Jonathan A. Hyam, Catherine A. Welch, David A. Harrison, and David K. Menon. Case mix, outcomes and comparison of risk prediction models for admissions to adult, general and specialist critical care units for head injury: A secondary analysis of the ICNARC Case Mix Programme Database. *Critical Care (London, England)*, 10 Suppl 2:S2, 2006.

[28] David A Harrison, Gareth J Parry, James R Carpenter, Alasdair Short, and Kathy Rowan. A new risk prediction model for critical care: The Intensive Care National Audit & Research Centre (ICNARC) model. *Critical care medicine*, 35(4):1091–8, April 2007.

[29] Paloma Ferrando-Vivas, Andrew Jones, Kathryn M. Rowan, and David A. Harrison. Development and validation of the new ICNARC model for prediction of acute hospital mortality in adult critical care. *Journal of Critical Care*, 38:335–339, 2017.

[30] ISO. ISO/IEC 27001 Information Security Management. https://www.iso.org/isoiec-27001-information-security.html. Last Accessed 23-06-2021.

[31] NHS. Data Security and Protection Toolkit. https://www.dsptoolkit.nhs.uk. Last Accessed 23-06-2021.

[32] Annie Herbert, Linda Wijlaars, Ania Zylbersztejn, David Cromwell, and Pia Hardelid. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *International journal of epidemiology*, 46(4):1093–1093i, August 2017.

[33] University College London Hospitals NHS Foundation Trust. National Institute for Health Research, health Informatics Collaborative, Critical care. https://www.youtube.com/watch?v=NjE9VQo-nP4&t=23s, June 2015.

[34] Pamela M Vacek. Assessing the Effect of Intensity When Exposure Varies Over Time. page 9, 1997.

[35] Derek K Chu, Lisa H-y Kim, Paul J Young, Nima Zamiri, Saleh A Almenawer, Roman Jaeschke, Wojciech Szczeklik, and Holger J Schünemann. Mortality and morbidity in acutely ill adults treated with liberal versus conservative oxygen therapy (IOTA): A systematic review and meta-analysis. *Lancet*, 391(10131):1693–1705, 2018.

[36] Benjamin Post, Edward Palmer, Steve Harris, Mervyn Singer, and Daniel Martin. Oxygenation of the critically ill in selected intensive care units in the UK: Are we usual? *British Journal of Anaesthesia*, 125(3):e277–e279, September 2020.

[37] Antoine Lavoisier. *Histoire de La Société Royale de Médecine. Années 1782 et 1783. Avec Les Mémoires de Médecine et de Physique Médicale, Pour Les Mêmes Années, Tirés Des Registres de Cette Société.* First edition, 1787.

[38] J. S. Haldane. The Therapeutic Administration of Oxygen. *British Medical Journal*, 1(2928):181–183, February 1917.

[39] L. Frank, J. R. Bucher, and R. J. Roberts. Oxygen toxicity in neonatal and adult animals of various species. *Journal of Applied Physiology: Respiratory, Environmental and Exercise Physiology*, 45(5):699–704, November 1978.

[40] R. E. Barber and W. K. Hamilton. Oxygen toxicity in man. A prospective study in patients with irreversible brain damage. *The New England Journal of Medicine*, 283(27):1478–1484, December 1970.

[41] W B Davis, S I Rennard, P B Bitterman, and R G Crystal. Pulmonary oxygen toxicity. Early reversible changes in human alveolar structures induced by hyperoxia. *The New England journal of medicine*, 309(15):878–883, 1983.

[42] Jérôme Aboab, Bjorn Jonson, Achille Kouatchet, Solenne Taille, Lisbet Niklason, and Laurent Brochard. Effect of inspired oxygen fraction on alveolar derecruitment in acute respiratory distress syndrome. *Intensive Care Medicine*, 32(12):1979–1986, December 2006.

[43] Hendrik J. F. Helmerhorst, Rob B. P. de Wilde, Dae Hyun Lee, Meindert Palmen, Jos R. C. Jansen, David J. van Westerloo, and Evert de Jonge. Hemodynamic effects of short-term hyperoxia after coronary artery bypass grafting. *Annals of Intensive Care*, 7(1):20, December 2017.

[44] Z. Bak, F. Sjöberg, A. Rousseau, I. Steinvall, and B. Janerot-Sjoberg. Human cardiovascular dose-response to supplemental oxygen. *Acta Physiologica (Oxford, England)*, 191(1):15–24, September 2007.

[45] Patrick H. McNulty, Bryan J. Robertson, Mark A. Tulli, Joshua Hess, Lisa A. Harach, Sofia Scott, and Lawrence I. Sinoway. Effect of hyperoxia and vitamin C on coronary blood flow in patients with ischemic heart disease. *Journal of Applied Physiology*, 102(5):2040–2045, May 2007.

[46] C. T. Dollery, D. W. Hill, C. M. Mailer, and P. S. Ramalho. High Oxygen Pressure and the Retinal Blood Vessels. *Lancet*, 2(7354):291–292, August 1964.

[47] W. A. Haque, J. Boehmer, B. S. Clemson, U. A. Leuenberger, D. H. Silber, and L. I. Sinoway. Hemodynamic effects of supplemental oxygen administration in congestive heart failure. *Journal of the American College of Cardiology*, 27(2):353–357, February 1996.

[48] S D Milone, G E Newton, and J D Parker. Hemodynamic and biochemical effects of 100% oxygen breathing in humans. *Canadian journal of physiology and pharmacology*, 77(2):124–130, 1999.

[49] W. Ganz, R. Donoso, H. Marcus, and H. J. Swan. Coronary hemodynamics and myocardial oxygen metabolism during oxygen breathing in patients with and without coronary artery disease. *Circulation*, 45(4):763–768, April 1972.

[50] Ichiroh Shimada, Ayumi Kubota, Masataka Katoh, and Fumiko Suzuki. Hyperoxia causes diffuse alveolar damage through mechanisms involving upregulation of c-Myc/Bax and enhanced production of reactive oxygen species. *Respiratory Investigation*, 54(1):59–68, January 2016.

[51] Pierre Asfar, Frédérique Schortgen, Julie Boisramé-Helms, Julien Charpentier, Emmanuel Guérot, Bruno Megarbane, David Grimaldi, Fabien Grelon, Nadia Anguel, et al for the HYPER2S Investigators REVA research network, Sigismond Lasocki, Matthieu Henry-Lagarrigue, Frédéric

Gonzalez, François Legay, Christophe Guitton, Maleka Schenck, Jean Marc Doise, Jérôme Devaquet, Thierry Van Der Linden, Delphine Chatellier, Jean Philippe Rigaud, Jean Dellamonica, Fabienne Tamion, Ferhat Meziani, Alain Mercat, Didier Dreyfuss, Valérie Seegers, and Peter Radermacher. Hyperoxia and hypertonic saline in patients with septic shock (HYPER2S): A two-by-two factorial, multicentre, randomised, clinical trial. *Lancet Respiratory Medicine*, 5(3):180–190, March 2017.

[52] The ICU-ROX Investigators and the Australian and New Zealand Intensive Care Society Clinical Trials Group. Conservative Oxygen Therapy during Mechanical Ventilation in the ICU. *New England Journal of Medicine*, page NEJMoa1903297, October 2019.

[53] Paul Young, Diane Mackle, Rinaldo Bellomo, Michael Bailey, Richard Beasley, Adam Deane, Glenn Eastwood, Simon Finfer, Ross Freebairn, Victoria King, Natalie Linke, Edward Litton, Colin McArthur, Shay McGuinness, Rakshit Panwar, and ICU-ROX Investigators the Australian New Zealand Intensive Care Society Clinical Trials Group. Conservative oxygen therapy for mechanically ventilated adults with sepsis: A post hoc analysis of data from the intensive care unit randomized trial comparing two approaches to oxygen therapy (ICU-ROX). *Intensive Care Medicine*, 46(1):17–26, January 2020.

[54] Dion Stub, Karen Smith, Stephen Bernard, Ziad Nehme, Michael Stephenson, Janet E. Bray, Peter Cameron, Bill Barger, Andris H. Ellims et al, for the AVOID Investigators, Andrew J. Taylor, Ian T. Meredith, and David M. Kaye. Air versus oxygen in ST-segment-elevation myocardial infarction. *Circulation*, 131(24):2143–2150, June 2015.

[55] M. Wijesinghe, K. Perrin, A. Ranchord, M. Simmonds, M. Weatherall, and R. Beasley. Routine use of oxygen in the treatment of myocardial infarction: Systematic review. *Heart*, 95(3):198–202, March 2009.

[56] Robin Hofmann, Stefan K. James, Tomas Jernberg, Bertil Lindahl, David Erlinge, Nils Witt, Gabriel Arefalk, Mats Frick, Joakim Alfredsson, et al for the DETO2X–SWEDEHEART Investigators, Lennart Nilsson, Annica Ravn-Fischer, Elmir Omerovic, Thomas Kellerth, David Sparv, Ulf Ekelund, Rickard Linder, Mattias Ekström, Jörg Lauermann, Urban Haaga, John Pernow, Ollie Östlund, Johan Herlitz, and Leif Svensson. Oxygen Therapy in Suspected Acute Myocardial Infarction. *The New England Journal of Medicine*, 377(13):1240–1249, September 2017.

[57] Elisa Damiani, Erica Adrario, Massimo Girardis, Rocco Romano, Paolo Pelaia, Mervyn Singer, and Abele Donati. Arterial hyperoxia and mortality in critically ill patients: A systematic review and meta-analysis. *Critical Care*, 18(6):711, December 2014.

[58] Glenn M. Eastwood, Aiko Tanaka, Emilo Daniel Valenzuela Espinoza, Leah Peck, Helen Young, Johan Mårtensson, Ling Zhang, Neil J. Glassford, Yu Feng Frank Hsiao, Satoshi Suzuki, and Rinaldo Bellomo. Conservative oxygen therapy in mechanically ventilated patients following cardiac arrest: A retrospective nested cohort study. *Resuscitation*, 101:108–114, 2016.

[59] Satoshi Suzuki, Glenn M. Eastwood, Neil J. Glassford, Leah Peck, Helen Young, Mercedes Garcia-Alvarez, Antoine G. Schneider, and Rinaldo Bellomo. Conservative Oxygen Therapy in Mechanically Ventilated Patients: A Pilot Before-and-After Trial. *Critical Care Medicine*, 42(6):1414–1422, June 2014.

[60] Hendrik J. F. Helmerhorst, Marcus J. Schultz, Peter H. J. van der Voort, Robert J. Bosman, Nicole P. Juffermans, Rob B. P. de Wilde, M. Elske van den Akker-van Marle, Leti van Bodegom-Vos, Marieke de Vries, Saeid Eslami, Nicolette F. de Keizer, Ameen Abu-Hanna, David J. van Westerloo, and Evert de Jonge. Effectiveness and Clinical Outcomes of a Two-Step Implementation of Conservative Oxygenation Targets in Critically Ill Patients: A Before and After Trial. *Critical Care Medicine*, 44(3):554–563, 2016.

[61] Massimo Girardis, Stefano Busani, Elisa Damiani, Abele Donati, Laura Rinaldi, Andrea Marudi, Andrea Morelli, Massimo Antonelli, and Mervyn Singer. Effect of Conservative vs Conventional Oxygen Therapy on Mortality Among Patients in an Intensive Care Unit. *JAMA*, 316(15):1583, 2016.

[62] Hendrik J. F. Helmerhorst, Marie-José Roos-Blom, David J. van Westerloo, and Evert de Jonge. Association Between Arterial Hyperoxia and Outcome in Subsets of Critical Illness. *Critical Care Medicine*, 43(7):1508–1519, 2015.

[63] J. Hope Kilgannon, Alan E. Jones, Nathan I. Shapiro, Mark G. Angelos, Barry Milcarek, Krystal Hunter, Joseph E. Parrillo, Stephen Trzeciak, and for the Emergency Medicine Shock Research Network (EMShockNet) Investigators. Association between arterial hyperoxia following resuscitation from cardiac arrest and in-hospital mortality. *JAMA*, 303(21):2165–2171, June 2010.

[64] J. Hope Kilgannon, Alan E. Jones, Joseph E. Parrillo, R. Phillip Dellinger, Barry Milcarek, Krystal Hunter, Nathan I. Shapiro, and Stephen Trzeciak. Relationship between supranormal oxygen tension and outcome after resuscitation from cardiac arrest. *Circulation*, 123(23):2717–2722, 2011.

[65] Brian W. Roberts, J. Hope Kilgannon, Benton R. Hunter, Michael A. Puskarich, Lisa Pierce, Michael Donnino, Marion Leary, Jeffrey A. Kline, Alan E. Jones, Nathan I. Shapiro, Benjamin S. Abella, and Stephen Trzeciak. Association Between Early Hyperoxia Exposure After Resuscitation From Cardiac Arrest and Neurological Disability: Prospective Multicenter Protocol-Directed Cohort Study. *Circulation*, 137(20):2114–2124, May 2018.

[66] Renate Stolmeijer, Jan C. ter Maaten, Jan G. Zijlstra, and Jack J. M. Ligtenberg. Oxygen therapy for sepsis patients in the emergency department: A little less? *European Journal of Emergency Medicine: Official Journal of the European Society for Emergency Medicine*, 21(3):233–235, June 2014.

[67] Julien Demiselle, Martin Wepler, Clair Hartmann, Peter Radermacher, Frédérique Schortgen, Ferhat Meziani, Mervyn Singer, Valérie Seegers, and Pierre Asfar. Hyperoxia toxicity in septic shock patients according to the Sepsis-3 criteria: A post hoc analysis of the HYPER2S trial. *Annals of Intensive Care*, 8:90, September 2018.

[68] Rakshit Panwar, Miranda Hardie, Rinaldo Bellomo, Loïc Barrot, Glenn M Eastwood, Paul J Young, Gilles Capellier, Peter W J Harrigan, and Michael Bailey. Conservative versus Liberal Oxygenation Targets for Mechanically Ventilated Patients. A Pilot Multicenter Randomized Controlled Trial. *American Journal of Respiratory and Critical Care Medicine*, 193(1):43–51, January 2016.

[69] F. Sjöberg and M. Singer. The medical use of oxygen: A time for critical reappraisal. *Journal of Internal Medicine*, 274(6):505–528, 2013.

[70] Evert de Jonge, Linda Peelen, Peter J Keijzers, Hans Joore, Dylan de Lange, Peter HJ J van der Voort, Robert J Bosman, Ruud AL L de Waal, Ronald Wesselink, and Nicolette F de Keizer. Association between administered oxygen, arterial partial oxygen pressure and mortality in mechanically ventilated intensive care unit patients. *Critical Care*, 12(6):R156, December 2008.

[71] Hendrik J. F. Helmerhorst, Derk L. Arts, Marcus J. Schultz, Peter H. J. van der Voort, Ameen Abu-Hanna, Evert de Jonge, and David J. van Westerloo. Metrics of Arterial Hyperoxia and Associated Outcomes in Critical Care*:. *Critical Care Medicine*, 45(2):187–195, February 2017.

[72] Glenn Eastwood, Rinaldo Bellomo, Michael Bailey, Gopal Taori, David Pilcher, Paul Young, and Richard Beasley. Arterial oxygen tension and mortality in mechanically ventilated patients. *Intensive Care Medicine*, 38(1):91–98, 2012.

[73] William Osler. *The Principles and Practice of Medicine*. Third edition, 1898.

[74] Helen McKenna. *The Bioenergetic and Redox Phenotype in Human Critical Illness*. PhD thesis, University College London, May 2020.

[75] Jeremy Stoller, Laura Halpin, Matthew Weis, Brett Aplin, Weikai Qu, Claudiu Georgescu, and Munier Nazzal. Epidemiology of severe sepsis: 2008-2012. *Journal of Critical Care*, 31(1):58–62, February 2016.

[76] Margaret Jean Hall, Sonja N. Williams, Carol J. DeFrances, and Aleksandr Golosinskiy. Inpatient care for septicemia or sepsis: A challenge for patients and hospitals. *NCHS data brief*, (62):1–8, June 2011.

[77] Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, Carolin Fleischmann-Struzek, Flavia R Machado, Konrad K Reinhart, Kathryn Rowan, Christopher W Seymour, R Scott Watson, T Eoin West, Fatima Marinho, Simon I Hay, Rafael Lozano, Alan D Lopez, Derek C Angus, Christopher J L Murray, and Mohsen Naghavi. Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study. *The Lancet*, 395(10219):200–211, January 2020.

[78] M. Shankar-Hari, D.A. Harrison, G.D. Rubenfeld, and K. Rowan. Epidemiology of sepsis and septic shock in critical care units: Comparison between sepsis-2 and sepsis-3 populations using a national critical care database. *British Journal of Anaesthesia*, 119(4):626–636, October 2017.

[79] J. C. Marshall. SIRS and MODS: What is their relevance to the science and practice of intensive care? *Shock (Augusta, Ga.)*, 14(6):586–589, December 2000.

[80] John C. Marshall. Sepsis Definitions: A Work in Progress. *Critical Care Clinics*, 34(1):1–14, 2018.

[81] Alan Aderem and Richard J. Ulevitch. Toll-like receptors in the induction of the innate immune response. *Nature*, 406(6797):782–787, August 2000.

[82] Ruslan Medzhitov and Charles A Janeway Jr. Innate immune recognition and control of adaptive immune responses. *Seminars in Immunology*, 10(5):351–353, October 1998.

[83] P. Matzinger. Tolerance, danger, and the extended family. *Annual Review of Immunology*, 12:991–1045, 1994.

[84] Liliana Schaefer. Complexity of Danger: The Diverse Nature of Damage-associated Molecular Patterns. *Journal of Biological Chemistry*, 289(51):35237–35245, December 2014.

[85] Qin Zhang, Mustafa Raoof, Yu Chen, Yuka Sumi, Tolga Sursal, Wolfgang Junger, Karim Brohi, Kiyoshi Itagaki, and Carl J. Hauser. Circulating mitochondrial DAMPs cause inflammatory responses to injury. *Nature*, 464(7285):104–107, March 2010.

[86] Mervyn Singer. Critical illness and flat batteries. *Critical Care (London, England)*, 21(Suppl 3):309, December 2017.

[87] Rachel Pool, Hernando Gomez, and John A. Kellum. Mechanisms of Organ Dysfunction in Sepsis. *Critical Care Clinics*, 34(1):63–80, 2018.

[88] Andrew Conway-Morris, Julie Wilson, and Manu Shankar-Hari. Immune Activation in Sepsis. *Critical Care Clinics*, 34(1):29–42, 2018.

[89] Mervyn Singer. The role of mitochondrial dysfunction in sepsis-induced multi-organ failure. *Virulence*, 5(1):66–72, January 2014.

[90] David Brealey, Michael Brand, Iain Hargreaves, Simon Heales, John Land, Ryszard Smolenski, Nathan A Davies, Chris E Cooper, and Mervyn Singer. Association between mitochondrial dysfunction and severity and outcome of septic shock. *The Lancet*, 360(9328):219–223, July 2002.

[91] David Brealey, Sekhar Karyampudi, Thomas S. Jacques, Marco Novelli, Ray Stidwill, Val Taylor, Ryszard T. Smolenski, and Mervyn Singer. Mitochondrial dysfunction in a long-term rodent model of sepsis and organ failure.

American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 286(3):R491–R497, March 2004.

[92] Alain Rudiger and Mervyn Singer. Mechanisms of sepsis-induced cardiac dysfunction:. *Critical Care Medicine*, 35(6):1599–1608, June 2007.

[93] Mervyn Singer, Clifford S. Deutschman, Christopherwarren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tomvan Der Poll, Jean Louis Vincent, and Derek C. Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA - Journal of the American Medical Association*, 315(8):801–810, 2016.

[94] Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):762–74, February 2016.

[95] J L Vincent, R Moreno, J Takala, S Willatts, A De Mendonca, H Bruining, C K Reinhart, P M Suter, and L G Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis- Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med*, 22(7):707–10., 1996.

[96] J R Le Gall, J Klar, S Lemeshow, F Saulnier, C Alberti, a Artigas, and D Teres. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. *JAMA : the journal of the American Medical Association*, 1996.

[97] ICNARC. Definition: Sepsis (Sepsis-3) during the first 24 hours following admission to the critical care unit, November 2018.

[98] J.D. Young, C. Goldfrad, and K. Rowan. Development and testing of a hierarchical method to code the reason for admission to intensive care units: The ICNARC Coding Method. *British Journal of Anaesthesia*, 87(4):543–548, October 2001.

[99] Derek C. Angus. The Search for Effective Therapy for Sepsis: Back to the Drawing Board? *JAMA*, 306(23):2614, December 2011.

[100] Arturo J. Martí-Carvajal, Ivan Solà, Dimitrios Lathyris, and Andrés Felipe Cardona. Human recombinant activated protein C for severe sepsis. *The Cochrane Database of Systematic Reviews*, (3):CD004388, March 2012.

[101] John C. Marshall. The staging of sepsis: Understanding heterogeneity in treatment efficacy. *Critical Care*, 9(6):626–628, 2005.

[102] Jonathan Cohen, Jean-Louis Vincent, Neill K. J. Adhikari, Flavia R. Machado, Derek C. Angus, Thierry Calandra, Katia Jaton, Stefano Giulieri, Julie Delaloye, Steven Opal, Kevin Tracey, Tom van der Poll, and Eric Pelfrene. Sepsis: A roadmap for future research. *The Lancet. Infectious Diseases*, 15(5):581–614, May 2015.

[103] Hallie C Prescott, John J. Osterholzer, Kenneth M. Langa, Derek C. Angus, and Theodore J Iwashyna. Late mortality after sepsis: Propensity matched cohort study. *BMJ (Clinical research ed.)*, 353:i2375, May 2016.

[104] Alain Rudiger, Alex Dyson, Karen Felsmann, Jane E. Carré, Valerie Taylor, Sian Hughes, Innes Clatworthy, Alessandro Protti, Denis Pellerin, Jana Lemm, Ralf A. Claus, Michael Bauer, and Mervyn Singer. Early functional and transcriptomic changes in the myocardium predict outcome in a long-term rat model of sepsis. *Clinical Science*, 124(6):391–401, March 2013.

[105] Christopher W. Seymour, Jeremy M. Kahn, Christian Martin-Gill, Clifton W. Callaway, Donald M. Yealy, Damon Scales, and Derek C. Angus. Delays from first medical contact to antibiotic administration for sepsis. *Critical Care Medicine*, 45(5):759–765, 2017.

[106] Manu Shankar-Hari, David A. Harrison, Kathryn M. Rowan, and Gordon D. Rubenfeld. Estimating attributable fraction of mortality from sepsis to inform clinical trials. *Journal of Critical Care*, 45:33–39, June 2018.

[107] Mervyn Singer, Matt Inada-Kim, and Manu Shankar-Hari. Sepsis hysteria: Excess hype and unrealistic expectations. *The Lancet*, 394(10208):1513–1514, October 2019.

[108] Emma E. Davenport, Katie L. Burnham, Jayachandran Radhakrishnan, Peter Humburg, Paula Hutton, Tara C. Mills, Anna Rautanen, Anthony C. Gordon, Christopher Garrard, Adrian V. S. Hill, Charles J. Hinds, and Julian C. Knight. Genomic landscape of the individual host response and outcomes in sepsis: A prospective cohort study. *The Lancet. Respiratory Medicine*, 4(4):259–271, April 2016.

[109] Brendon P. Scicluna, Lonneke A. van Vught, Aeilko H. Zwinderman, Maryse A. Wiewel, Emma E. Davenport, Katie L. Burnham, Peter Nürnberg, Marcus J. Schultz, Janneke Horn, Olaf L. Cremer, Marc J. Bonten, Charles J. Hinds, Hector R. Wong, Julian C. Knight, Tom van der Poll, Friso M. de Beer, Lieuwe D.J. Bos, Jos F. Frencken, Maria E. Koster-Brouwer, Kirsten van de Groep, Diana M. Verboom, Gerie J. Glas, Roosmarijn T.M. van Hooijdonk, Arie J. Hoogendijk, Mischa A. Huson, Peter M.Klein Klouwenberg, David S.Y. Ong, Laura R.A. Schouten, Marleen Straat, Esther Witteveen, and Luuk Wieske. Classification of patients with sepsis according to blood genomic endotype: A prospective cohort study. *The Lancet Respiratory Medicine*, 5(10):816–826, 2017.

[110] Raymond J. Langley, Ephraim L. Tsalik, Jennifer C. van Velkinburgh, Seth W. Glickman, Brandon J. Rice, Chunping Wang, Bo Chen, Lawrence Carin, Arturo Suarez, Robert P. Mohney, Debra H. Freeman, Mu Wang, Jinsam You, Jacob Wulff, J. Will Thompson, M. Arthur Moseley, Stephanie Reisinger, Brian T. Edmonds, Brian Grinnell, David R. Nelson, Darrell L. Dinwiddie, Neil A. Miller, Carol J. Saunders, Sarah S. Soden, Angela J. Rogers, Lee Gazourian, Laura E. Fredenburgh, Anthony F. Massaro, Rebecca M. Baron, Augustine M. K. Choi, G. Ralph Corey, Geoffrey S. Ginsburg, Charles B. Cairns, Ronny M. Otero, Vance G. Fowler, Emanuel P. Rivers, Christopher W. Woods, and Stephen F. Kingsmore. An integrated clinico-metabolomic model improves prediction of death in sepsis. *Science Translational Medicine*, 5(195):195ra95, July 2013.

[111] Sarah J. Beesley, Emily L. Wilson, Michael J. Lanspa, Colin K. Grissom, Sajid Shahul, Daniel Talmor, and Samuel M. Brown. Relative Bradycardia in Patients With Septic Shock Requiring Vasopressor Therapy. *Critical Care Medicine*, 45(2):225–233, February 2017.

[112] Hendrik Schmidt, Ursula Müller-Werdan, Thomas Hoffmann, Darrel P. Francis, Massimo F. Piepoli, Mathias Rauchhaus, Roland Prondzinsky, Harald Loppnow, Michael Buerke, Dirk Hoyer, and Karl Werdan. Autonomic dysfunction predicts mortality in patients with multiple organ dysfunction syndrome of different age groups. *Critical Care Medicine*, 33(9):1994–2002, September 2005.

[113] Daniel B. Knox, Michael J. Lanspa, Kathryn G. Kuttler, Simon C. Brewer, and Samuel M. Brown. Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome. *Intensive Care Medicine*, 41(5):814–822, 2015.

[114] Djordje Gligorijevic, Jelena Stojanovic, and Zoran Obradovic. Disease types discovery from a large database of inpatient records: A sepsis study. *Methods (San Diego, Calif.)*, 111:45–55, December 2016.

[115] Brett K. Beaulieu-Jones and Casey S. Greene. Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of Biomedical Informatics*, 64:168–178, December 2016.

[116] Rimma Pivovarov, Adler J. Perotte, Edouard Grave, John Angiolillo, Chris H. Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous EHR data. *Journal of biomedical informatics*, 58:156–165, December 2015.

[117] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[118] Michael B. Mayhew, Brenden K. Petersen, Ana Paula Sales, John D. Greene, Vincent X. Liu, and Todd S. Wasson. Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models. *Journal of biomedical informatics*, 78(May 2017):33–42, February 2018.

[119] P. M. Rothwell. Can overall results of clinical trials be applied to all patients? *Lancet (London, England)*, 345(8965):1616–1619, June 1995.

[120] Armand R. J. Girbes and Harm-Jan de Grooth. Time to stop randomized and large pragmatic trials for intensive care medicine syndromes: The case of sepsis and acute respiratory distress syndrome. *Journal of Thoracic Disease*, 12(S1):S101–S109, February 2020.

[121] Edna Schechtman. Odds Ratio, Relative Risk, Absolute Risk Reduction, and the Number Needed to Treat—Which of These Should We Use? *Value in Health*, 5(5):431–436, September 2002.

[122] Mohammed Nabhan, Tarig Elraiyah, Daniel R. Brown, James Dilling, Annie LeBlanc, Victor M. Montori, Timothy Morgenthaler, James Naessens, Larry Prokop, Veronique Roger, Stephen Swensen, Rodney L. Thompson, and M. Hassan Murad. What is preventable harm in healthcare? A systematic review of definitions. *BMC health services research*, 12:128, May 2012.

[123] T. M. Cook, N. Woodall, J. Harper, and J. Benger. Major complications of airway management in the UK: Results of the Fourth National Audit Project of the Royal College of Anaesthetists and the Difficult Airway Society. Part 2: Intensive care and emergency departments. *British Journal of Anaesthesia*, 106(5):632–642, 2011.

[124] Theodore J. Iwashyna, James F. Burke, Jeremy B. Sussman, Hallie C. Prescott, Rodney A. Hayward, and Derek C. Angus. Implications of Heterogeneity of Treatment Effect for Reporting and Analysis of Randomized Trials in Critical Care. *American Journal of Respiratory and Critical Care Medicine*, 192(9):1045–1051, November 2015.

[125] Charles Warlow. MRC European Carotid Surgery Trial: Interim results for symptomatic patients with severe (70-99%) or with mild (0-29%) carotid stenosis. European Carotid Surgery Trialists' Collaborative Group. *Lancet (London, England)*, 337(8752):1235–1243, May 1991.

[126] B. Farrell, J. Godwin, S. Richards, and C. Warlow. The United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: Final results. *Journal of Neurology, Neurosurgery, and Psychiatry*, 54(12):1044–1054, December 1991.

[127] The RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19 — Preliminary Report. *New England Journal of Medicine*, page NEJMoa2021436, July 2020.

[128] Ming-Ju Tsai, Kuang-Yao Yang, Ming-Cheng Chan, Kuo-Chin Kao, Hao-Chien Wang, Wann-Cherng Perng, Chieh-Liang Wu, Shinn-Jye Liang, Wen-Feng Fang, Jong-Rung Tsai, Wei-An Chang, Ying-Chun Chien, Wei-Chih Chen, Han-Chung Hu, Chiung-Yu Lin, Wen-Cheng Chao, Chau-Chyun Sheu, and for Taiwan Severe Influenza Research Consortium (TSIRC) Investigators. Impact of corticosteroid treatment on clinical outcomes of influenza-associated ARDS: A nationwide multicenter study. *Annals of Intensive Care*, 10(1):26, February 2020.

[129] ARDSnet. Ventilation with Lower Tidal Volumes as Compared with Traditional Tidal Volumes for Acute Lung Injury and the Acute Respiratory Distress Syndrome. *The New England Journal of Medicine*, page 8, 2000.

[130] Steve Harris. *Timing Delivery of Critical Care*. PhD thesis, London School of Hygiene and Tropical Medicine, January 2014.

[131] Tudor Toma, Ameen Abu-Hanna, and Robert-Jan Bosman. Discovery and inclusion of SOFA score episodes in mortality prediction. *Journal of Biomedical Informatics*, 40(6):649–660, December 2007.

[132] Tudor Toma, Ameen Abu-Hanna, and Robert-Jan Bosman. Discovery and integration of univariate patterns from daily individual organ-failure scores for intensive care mortality prediction. *Artificial Intelligence in Medicine*, 43(1):47–60, May 2008.

[133] Andre L. Holder, Elizabeth Overton, Peter Lyu, Jordan A. Kempker, Shamim Nemati, Fereshteh Razmi, Greg S. Martin, Timothy G. Buchman, and David J. Murphy. Serial Daily Organ Failure Assessment Beyond ICU Day 5 Does Not Independently Add Precision to ICU Risk-of-Death Prediction:. *Critical Care Medicine*, 45(12):2014–2022, December 2017.

[134] Omar Badawi, Xinggang Liu, Erkan Hassan, Pamela J. Amelung, and Sunil Swami. Evaluation of ICU Risk Models Adapted for Use as Continuous Markers of Severity of Illness Throughout the ICU Stay*:. *Critical Care Medicine*, 46(3):361–367, March 2018.

[135] Emmanuelle Deslandes and Sylvie Chevret. Joint modeling of multivariate longitudinal data and the dropout process in a competing risk setting: Application to ICU data. *BMC medical research methodology*, 10:69, July 2010.

[136] Jammbe Z Musoro, Aeilko H Zwinderman, Ameen Abu-Hanna, Rob Bosman, and Ronald B Geskus. Dynamic prediction of mortality among patients in intensive care using the sequential organ failure assessment (SOFA)

score: A joint competing risk survival and longitudinal modeling approach: Dynamic prediction of mortality using SOFA scores. *Statistica Neerlandica*, 72(1):34–47, February 2018.

[137] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019.

[138] Sam Brilleman. Simulate Joint Longitudinal and Survival Data. 2018.

[139] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, December 1982.

[140] D Cox and D Oakes. *Analysis of Survival Data*. Chapman & Hall/CRC, February 2018.

[141] Kate Bull and David J Spiegelhalter. Tutorial in Biostatistics Survival Analysis in Observational Studies. *STAT. MED.*, 16:34, 1997.

[142] Stef van Buuren. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC Interdisciplinary Statistics Series. CRC Press, Taylor and Francis Group, Boca Raton, second edition edition, 2018.

[143] Alessandro Gasparini, Keith R. Abrams, Jessica K. Barrett, Rupert W. Major, Michael J. Sweeting, Nigel J. Brunskill, and Michael J. Crowther. Mixed effects models for healthcare longitudinal data with an informative visiting process: A Monte Carlo simulation study. *Statistica Neerlandica*, page stan.12188, September 2019.

[144] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Texts in Statistical Science. Taylor and Francis, CRC Press, Boca Raton, second edition, 2020.

[145] Dimitris Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Number 6 in Chapman & Hall/CRC Biostatistics Series. CRC Press, Boca Raton, 2012.

[146] Dimitris Rizopoulos. Joint Models for Longitudinal and Survival Data [ESP72]. Statistical training programme. Rotterdam, Netherlands. https://erasmussummerprogramme.nl/summer-programme-courses/?ct=ESP72&pg=Courses, August 2019.

[147] Cheryl L. Faucett and Duncan C. Thomas. Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach. *Statistics in Medicine*, 15(15):1663–1685, August 1996.

[148] Michael S. Wulfsohn and Anastasios A. Tsiatis. A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, 53(1):330, March 1997.

[149] Katya Mauff, Ewout Steyerberg, Isabella Kardys, Eric Boersma, and Dimitris Rizopoulos. Joint models with multiple longitudinal outcomes and a time-to-event outcome: A corrected two-stage approach. *Statistics and Computing*, 30(4):999–1014, July 2020.

[150] Marcel Wolbers, Abdel Babiker, Caroline Sabin, Jim Young, Maria Dorrucci, Geneviève Chêne, Cristina Mussini, Kholoud Porter, and Heiner C. Bucher. Pretreatment CD4 cell slope and progression to AIDS or death in HIV-infected patients initiating antiretroviral therapy - The CASCADE collaboration: A collaboration of 23 cohort studies. *PLoS Medicine*, 7(2):1–9, 2010.

[151] S. Fieuws, G. Verbeke, B. Maes, and Y. Vanrenterghem. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics*, 9(3):419–431, July 2008.

[152] Felipe A. Medeiros. Biomarkers and surrogate endpoints in glaucoma clinical trials. *The British Journal of Ophthalmology*, 99(5):599–603, May 2015.

[153] Dimitris Rizopoulos. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010.

[154] Michael J. Crowther. Extended multivariate generalised linear and non-linear mixed effects models. *arXiv:1710.02223 [stat]*, October 2017.

[155] Graeme L. Hickey, Pete Philipson, and Andrea Jorgensen and. joineRML: A joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Medical Research Methodology*, 18(1), 2018.

[156] Dimitris Rizopoulos. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, 72(7):1–45, 2016.

[157] Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. Rstanarm: Bayesian applied regression modeling via Stan. 2020.

[158] SL Brilleman, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. Joint longitudinal and time-to-event models via Stan. 2018.

[159] Anastasios A Tsiatis and Marie Davidian. Joint Modeling of Longitudinal and Time-to-Event Data: An Overview. page 27, 2004.

[160] Dimitris Rizopoulos. Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics*, 67(3):819–829, September 2011.

[161] Daniel L Moody and Graeme G Shanks. Improving the quality of data models: Empirical validation of a quality management framework. *Information Systems*, 28(6):619–650, September 2003.

[162] Michael G. Kahn, Deborah Batson, and Lisa M. Schilling. Data Model Considerations for Clinical Effectiveness Researchers:. *Medical Care*, 50:S60–S67, July 2012.

[163] Katie Harron, Ruth Gilbert, David Cromwell, and Jan Van Der Meulen. Linking data for mothers and babies in de-identified electronic health data. *PLoS ONE*, 11(10):1–18, 2016.

[164] Gareth Hagger-Johnson, Katie Harron, Arturo Gonzalez-Izquierdo, Mario Cortina-Borja, Nirupa Dattani, Berit Muller-Pebody, Roger Parslow, Ruth Gilbert, and Harvey Goldstein. Identifying possible false matches in anonymized hospital administrative data without patient identifiers. *Health Services Research*, 50(4):1162–1178, 2015.

[165] Gareth Hagger-Johnson, Katie Harron, Tom Fleming, Ruth Gilbert, Harvey Goldstein, Rebecca Landy, and Roger C Parslow. Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ Open*, 5(8):e008118, 2015.

[166] Katie Harron, Angie Wade, Ruth Gilbert, Berit Muller-Pebody, and Harvey Goldstein. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Medical Research Methodology*, 14(1):36, 2014.

[167] G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, January 2013.

[168] John Mylopoulos and Michael L. Brodie, editors. *Readings in Artificial Intelligence and Databases*. M. Kaufmann, San Mateo, Calif, 1989.

[169] Michelle M.Y. Wong, Keith P. McCullough, Brian A. Bieber, Juergen Bommer, Manfred Hecking, Nathan W. Levin, William M. McClellan, Ronald L. Pisoni, Rajiv Saran, Francesca Tentori, Tadashi Tomo, Friedrich K. Port, and Bruce M. Robinson. Interdialytic Weight Gain: Trends, Predictors, and Associated Outcomes in the International Dialysis Outcomes and Practice Patterns Study (DOPPS). *American Journal of Kidney Diseases*, 69(3):367–379, March 2017.

[170] Tim Benson and Grahame Grieve. *Principles of Health Interoperability: SNOMED CT, HL7 and FHIR*. Health Information Technology Standards. Springer International Publishing, Cham, 2016.

[171] Ronald Cornet and Nicolette de Keizer. Forty years of SNOMED: A literature review. *BMC Medical Informatics and Decision Making*, 8(S1):S2, October 2008.

[172] D.J. Berndt, J.W. Fisher, A.R. Hevner, and J. Studnicki. Healthcare data warehousing and quality assurance. *Computer*, 34(12):56–65, Dec./2001.

[173] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *BMJ*, page k1479, April 2018.

[174] Julie K. Bower, Sejal Patel, Joyce E. Rudy, and Ashley S. Felix. Addressing Bias in Electronic Health Record-based Surveillance of Cardiovascular Disease Risk: Finding the Signal Through the Noise. *Current Epidemiology Reports*, 4(4):346–352, December 2017.

[175] Michael G. Kahn, Tiffany J. Callahan, Juliana Barnard, Alan E. Bauck, Jeff Brown, Bruce N. Davidson, Hossein Estiri, Carsten Goerg, Erin Holve, Steven G. Johnson, Siaw-Teng Liaw, Marianne Hamilton-Lopez, Daniella Meeker, Toan C. Ong, Patrick Ryan, Ning Shang, Nicole G. Weiskopf, Chunhua Weng, Meredith N. Zozus, and Lisa Schilling. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 4(1):18, September 2016.

[176] Michael G. Kahn, Jeffrey S. Brown, Alein T. Chun, Bruce N. Davidson, Daniella Meeker, Patrick B. Ryan, Lisa M. Schilling, Nicole G. Weiskopf, Andrew E. Williams, and Meredith Nahm Zozus. Transparent Reporting of Data Quality in Distributed Data Networks. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 3(1):7, 2015.

[177] Vojtech Huser, Frank J. DeFalco, Martijn Schuemie, Patrick B. Ryan, Ning Shang, Mark Velez, Rae Woong Park, Richard D. Boyce, Jon Duke, Ritu Khare, Levon Utidjian, and Charles Bailey. Multisite Evaluation of a Data

Quality Tool for Patient-Level Clinical Datasets. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 4(1):24, November 2016.

[178] The Turing Way Community, Becky Arnold, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, Anna Krystalli, Alexander Morley, Martin O'Reilly, and Kirstie Whitaker. The Turing Way: A Handbook for Reproducible Data Science. Zenodo, March 2019.

[179] Hadley Wickham. *R Packages*. O'Reilly Media, Sebastopol, CA, first edition edition, 2015.

[180] Hadley Wickham. *Advanced R*. CRC Press/Taylor & Francis Group, Boca Raton, second edition edition, 2019.

[181] Jon Loeliger and Matthew McCullough. *Version Control with Git*. O'Reilly, Beijing, second edition edition, 2012.

[182] Greg Miller. A Scientist's Nightmare: Software Problem Leads to Five Retractions. *Science*, 314(5807):1856–1857, December 2006.

[183] Hadley Wickham. Testthat: Get started with testing. *The R Journal*, 3:5–10, 2011.

[184] Adam Brinckman, Kyle Chard, Niall Gaffney, Mihael Hategan, Matthew B. Jones, Kacper Kowalik, Sivakumar Kulasekaran, Bertram Ludäscher, Bryce D. Mecum, Jarek Nabrzyski, Victoria Stodden, Ian J. Taylor, Matthew J. Turk, and Kandace Turner. Computing environments for reproducibility: Capturing the "Whole Tale". *Future Generation Computer Systems*, 94:854–867, May 2019.

[185] Hadley Wickham. Tidy Data. *Journal of Statistical Software*, 59(10), 2014.

[186] Robert W. Aldridge, Kunju Shaji, Andrew C. Hayward, and Ibrahim Abubakar. Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies. *PLOS ONE*, 10(8):e0136179, August 2015.

[187] MIT Critical Data. MIMIC-OMOP. https://github.com/MIT-LCP/mimic-omop. Last Accessed 23-06-2021.

[188] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019.

[189] Alistair E. W. Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. The MIMIC-CXR Database, 2019.

[190] Matthias Templ, Alexander Kowarik, and Bernhard Meindl. Statistical Disclosure Control for Micro-Data Using the *R* Package **sdcMicro**. *Journal of Statistical Software*, 67(4), 2015.

[191] B R O'Driscoll, L S Howard, J Earis, and V Mak. British Thoracic Society guideline for oxygen use in adults in healthcare and emergency settings. *Thorax*, 72(Suppl 1):ii1–90, June 2017.

[192] Sonal Rachmale, Guangxi Li, Gregory Wilson, Michael Malinchoc, and Ognjen Gajic. Practice of excessive FIO2 and effect on pulmonary outcomes in mechanically ventilated patients with acute lung injury. *Respiratory Care*, 67(11):1887–1893, 2012.

[193] Luigi Pisani, Jan-Paul Roozeman, Fabienne D. Simonis, Antonio Giangregorio, Sophia M. van der Hoeven, Laura R. Schouten, Janneke Horn, Ary Serpa Neto, Emir Festic, Arjen M. Dondorp, Salvatore Grasso, Lieuwe D. Bos, and Marcus J. Schultz. Risk stratification using SpO2/FiO2 and PEEP at initial ARDS diagnosis and after 24 h in patients with moderate or severe ARDS. *Annals of Intensive Care*, 7(108), October 2017.

[194] Jason Y. Adams, Angela Rogers, Alejandro Schuler, Gregory P. Marelich, Jennifer M. Fresco, Sandra L. Taylor, Albert W. Riedl, Jennifer M. Baker,

Gabriel J. Escobar, and Vincent Liu. The association between SpO2/FiO2 ratio time-at-risk and hospital mortality in mechanically ventilated patients. In *C23. Critical Care: What Can Be Measured Can Be Improved - Investigating the Epidemiology and Outcomes of Patients with Acute Critical Illness*, American Thoracic Society International Conference Abstracts, pages A5029–A5029. American Thoracic Society, May 2017.

[195] Michael W. Sjoding, Robert P. Dickson, Theodore J. Iwashyna, Steven E. Gay, and Thomas S. Valley. Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine*, 383(25):2477–2478, December 2020.

[196] Martin Britos, Elizabeth Smoot, Kathleen D. Liu, B. Taylor Thompson, William Checkley, Roy G. Brower, and National Institutes of Health Acute Respiratory Distress Syndrome Network Investigators. The value of positive end-expiratory pressure and Fio_2 criteria in the definition of the acute respiratory distress syndrome. *Critical Care Medicine*, 39(9):2025–2030, September 2011.

[197] Daniel S Martin, Denny Z H Levett, Mike P W Grocott, and Hugh E Montgomery. Variation in human performance in the hypoxic mountain environment. *Exp Physiol*, 95(3):463–470, 2010.

[198] Nicholas J. Johnson, Kalani Dodampahala, Babette Rosselot, Sarah M. Perman, Mark E. Mikkelsen, Munish Goyal, David F. Gaieski, and Anne V. Grossestreuer. The Association Between Arterial Oxygen Tension and Neurological Outcome After Cardiac Arrest. *Therapeutic Hypothermia and Temperature Management*, 7(1):36–41, March 2017.

[199] Jean-François Llitjos, Jean-Paul Mira, Jacques Duranteau, and Alain Cariou. Hyperoxia toxicity after cardiac arrest: What is the evidence? *Annals of Intensive Care*, 6(1):23, 2016.

[200] Peter A. Stewart, John A. Kellum, Paul W. G. Elbers, and Peter A. Stewart. *Stewart's Textbook of Acid-Base*. AcidBase.org, Amsterdam, 2nd ed edition, 2009.

[201] Faculty of Intensive Care Medicine and Intensive Care Society. Guidelines for the Provision of Intensive Care Services (GPICS), October 2018.

[202] Loic Barrot, Pierre Asfar, Frederic Mauny, Hadrien Winiszewski, Florent Montini, Julio Badie, Jean-Pierre Quenot, Sebastien Pili-Floury, Belaid Bouhemad, Guillaume Louis, Bertrand Souweine, Olivier Collange, Julien Pottecher, Bruno Levy, Marc Puyraveau, Lucie Vettoretti, Jean-Michel Constantin, and Gilles Capellier. Liberal or Conservative Oxygen Therapy for Acute Respiratory Distress Syndrome. *New England Journal of Medicine*, 382(11):999–1008, March 2020.

[203] Patrick Royston, Willi Sauerbrei, and Heiko Becher. Modelling continuous exposures with a 'spike' at zero: A new procedure based on fractional polynomials. *Statistics in Medicine*, pages n/a–n/a, 2010.

[204] Mathilde Ruggiu, Nadia Aissaoui, Julien Nael, Caroline Haw-Berlemont, Bertrand Herrmann, Jean-Loup Augy, Sofia Ortuno, Damien Vimpère, Jean-Luc Diehl, Clotilde Bailleul, and Emmanuel Guerot. Hyperoxia effects on intensive care unit mortality: A retrospective pragmatic cohort study. *Critical Care*, 22(1), December 2018.

[205] Ursula Beckmann, Donna M. Gillies, Sean M. Berenholtz, Albert W. Wu, and Peter Pronovost. Incidents relating to the intra-hospital transfer of critically ill patients. An analysis of the reports submitted to the Australian Incident Monitoring Study in Intensive Care. *Intensive Care Medicine*, 30(8):1579–1585, August 2004.

[206] Jonathan P. N. Papson, Kassandra L. Russell, and David McD Taylor. Unexpected Events during the Intrahospital Transport of Critically Ill Patients. *Academic Emergency Medicine*, 14(6):574–577, June 2007.

[207] David Page, Enyo Ablordeppey, Brian T. Wessman, Nicholas M. Mohr, Stephen Trzeciak, Marin H. Kollef, Brian W. Roberts, and Brian M. Fuller. Emergency department hyperoxia is associated with increased mortality in mechanically ventilated patients: A cohort study. *Critical Care*, 22, January 2018.

[208] Bob Gray. *Cmprsk: Subdistribution Analysis of Competing Risks*, 2020.

[209] Stephanie J Reisinger, Patrick B Ryan, Donald J O'Hara, Gregory E Powell, Jeffery L Painter, Edward N Pattishall, and Jonathan A Morris. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *Journal of the American Medical Informatics Association*, 17(6):652–662, November 2010.

[210] R. Moreno, J. L. Vincent, R. Matos, A. Mendonça, F. Cantraine, L. Thijs, J. Takala, C. Sprung, M. Antonelli, H. Bruining, and S. Willatts. The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study. Working Group on Sepsis related Problems of the ESICM. *Intensive Care Medicine*, 25(7):686–696, July 1999.

[211] J. A. Russell, J. Singer, G. R. Bernard, A. Wheeler, W. Fulkerson, L. Hudson, R. Schein, W. Summer, P. Wright, and K. R. Walley. Changing pattern of organ dysfunction in early human sepsis is related to mortality. *Critical Care Medicine*, 28(10):3405–3411, October 2000.

[212] F. L. Ferreira, D. P. Bota, A. Bross, C. Mélot, and J. L. Vincent. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA*, 286(14):1754–1758, October 2001.

[213] Yasser Sakr, Suzana M. Lobo, Rui P. Moreno, Herwig Gerlach, V. Marco Ranieri, Argyris Michalopoulos, Jean-Louis Vincent, and SOAP Investigators. Patterns and early evolution of organ failure in the intensive care unit

and their relation to outcome. *Critical Care (London, England)*, 16(6):R222, November 2012.

[214] Harm-Jan de Grooth, Irma L. Geenen, Armand R. Girbes, Jean-Louis Vincent, Jean-Jacques Parienti, and Heleen M. Oudemans-van Straaten. SOFA and mortality endpoints in randomized controlled trials: A systematic review and meta-regression analysis. *Critical Care (London, England)*, 21(1):38, February 2017.

[215] Michael O. Harhay, Alessandro Gasparini, Allan J. Walkey, Gary E. Weissman, Michael J. Crowther, Sarah J. Ratcliffe, and James A. Russell. Assessing the Course of Organ Dysfunction Using Joint Longitudinal and Time-to-Event Modeling in the Vasopressin and Septic Shock Trial. *Critical Care Explorations*, 2(4):e0104, April 2020.

[216] Theodore J Iwashyna, Carol L. Hodgson, David Pilcher, Michael Bailey, Allison van Lint, Shaila Chavan, and Rinaldo Bellomo. Timing of onset and burden of persistent critical illness in Australia and New Zealand: A retrospective, population-based, observational study. *The Lancet Respiratory Medicine*, 4(7):566–573, 2016.

[217] Sean M. Bagshaw, Henry T. Stelfox, Theodore J. Iwashyna, Rinaldo Bellomo, Dan Zuege, and Xioaming Wang. Timing of onset of persistent critical illness: A multi-centre retrospective cohort study. *Intensive Care Medicine*, 44(12):2134–2144, December 2018.

[218] Torrin M. Liddell and John K. Kruschke. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328–348, November 2018.

[219] Thomas Beale and Sam Heard. An ontology-based model of clinical information. *Studies in health technology and informatics*, 129(Pt 1):760–4, 2007.

[220] Dipak Kalra, Thomas Beale, and Sam Heard. The openEHR Foundation. *Studies in health technology and informatics*, 115:153–73, 2005.

[221] Chunlan Ma, Heath Frankel, Thomas Beale, and Sam Heard. EHR query language (EQL)–a query language for archetype-based health records. *Studies in health technology and informatics*, 129(Pt 1):397–401, 2007.

[222] Richard Iannone and Mauricio Vargas. *Pointblank: Validation of Local and Remote Data Tables*, 2021.

[223] Miguel A. Hernán and James M. Robins. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *American Journal of Epidemiology*, 183(8):758–764, April 2016.

[224] Bibhas Chakraborty and Erica E. M. Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Statistics for Biology and Health. Springer, New York, NY, 2013.

[225] Anirudh Tomer, Daan Nieboer, Monique J. Roobol, Ewout W. Steyerberg, and Dimitris Rizopoulos. Personalized schedules for surveillance of low-risk prostate cancer patients. *Biometrics*, 75(1):153–162, March 2019.