Use of NHS Digital datasets as trial data in the UK: a position paper

Version 2.0, 11-Feb-2022

Macey L Murray[1,2,3]
Heather Pinches[3]
Marion Mafham[2,5]
Suzanne Hartley[3,6]
James Carpenter[1,2,4]
Martin Landray[2,3,5]
Sharon B Love[1,2]
Mahesh K B Parmar[1,2]
Matthew R Sydes[1,2,7]

(The Healthcare Systems Data for Clinical Trials Group)

1: MRC Clinical Trials Unit at UCL
2: Health Data Research UK
3: NHS DigiTrials Programme, NHS Digital
4: London School of Hygiene and Tropical Medicine, University of London
5: University of Oxford
6: University of Leeds
7: BHF Data Science Centre

CONTENTS

1 EXECUTIVE SUMMARY

Background: Clinical trial teams increasingly want to make use of data from healthcare systems (“healthcare data”), particularly to enhance recruitment and follow-up of participants, to reduce time and cost, and to stop the duplication of effort. However, there is continued uncertainty of how regulators regard healthcare data used for trial purposes, in terms of provenance, quality and reliability.

Objectives: There were two key objectives: First, to demonstrate the data integrity of two datasets held by NHS Digital (NHSD) that are most requested by trial teams; and second, to set out an approach by which any other healthcare systems datasets can be similarly evaluated.

Method: The data lifecycles of the datasets were carefully documented, mapping the flow of data from the originating healthcare provider’s databases to NHSD warehouses and onwards to clinical trials teams. These were assessed for evidence of whether the datasets are accurate, reliable, complete, contemporaneous, and well-governed.

Result: The assessment method was applied to (a) the Hospital Episode Statistics Admitted Patient Care (HES APC) dataset and (b) the Civil Registration of Deaths (CRD) dataset. This paper clearly demonstrates that their collection and management through NHSD systems ensure their integrity and reliability. The datasets are accurate representations of the data held by the originating providers (acute NHS trusts and local registrars).

Conclusion: Based on these findings, the HES APC and CRD datasets satisfy the assessment criteria that demonstrate they are reliable transcribed copies of the original source data.

Implications: First, these datasets can be used directly for clinical trial data, with trial teams focusing on the accuracy of algorithms and processes to identify particular outcomes rather than on the integrity of the data flow. Second, this assessment approach should be used to assess whether other healthcare systems datasets are ready to be used as transcribed copies of source data, and for data providers to take appropriate steps to redress this matter if they are not.

2 PURPOSE

We aimed to show that:

1. Two sets of healthcare data are suitable for trial use, as they have integrity and are transcribed directly from original source records (HES APC and CRD from NHSD),
2. Our approach can be similarly applied to other healthcare data to ascertain data integrity and reliability.

To achieve these aims, the paper is structured as follows. **Section 0** provides the background on healthcare data use in clinical trials and a summary of the regulatory guidance on trial data management and data integrity. **Section 0** states the assessment process, and **Section 5** describes the two selected healthcare datasets in line with this process, followed by three case studies of healthcare data use in clinical trials in **Section 6**. The paper concludes in **Section 7** with key recommendations for the use of these datasets and for the characterisation of other healthcare data.

3 BACKGROUND AND INTRODUCTION

This position paper is a summary of findings from a collaborative project between MRC Clinical Trials Unit at UCL, NHS DigiTrials, and University of Oxford. Clinical trial sponsors must have oversight of all trial data to comply with regulatory requirements, so they need assurances about the management and lineage of healthcare datasets that they use for trial purposes.

3.1 Increasing use of routinely collected healthcare data in clinical trials

Trial teams running trials in the UK are requesting to access healthcare systems data (also referred to as routinely-collected health data, registry data, and "real world data") to aid recruitment and to supplement or replace trial-collected information on treatment, procedures and clinical outcomes.[1] A 2020 review by Lensen *et al*. showed that, between 2013 and 2018, fewer than 5% of UK-based randomised clinical trials (RCTs) obtained routine data, with CRD and HES being the most frequently accessed, often to inform researchers about participant outcomes.[2] This proportion is expected to grow with better documentation of, and access routes to, these datasets. For example, McKay *et al*. found that, somewhat conversely to Lensen et al, 50% of NIHR-funded trials (to 2019) were *planning* to access and use healthcare data.[3] So there is an intention by trialists to make use of healthcare data for study design and recruitment through to outcome ascertainment and post-trial follow-up,[1, 3, 4] although there are some challenges that may limit their application.[5] NHSD is a major custodian in England of national-level healthcare systems data. Given the widespread use of NHSD datasets in healthcare research outside of clinical trials, there is an expectation that these datasets should also be suitable for use in clinical trials.[6-8]

The UK regulatory authority, the Medicines and Healthcare products Regulatory Agency (MHRA) has recognised the value of "real-world data" (RWD; which includes healthcare systems data) in generating "real-world evidence". By publishing draft guidance on RWD, MHRA is encouraging sponsors to utilise these datasets in their trials to support regulatory decisions. This guidance is currently under revision after public consultation in December 2020, and is not intended to provide details on how sponsors demonstrate the RWD source is of sufficient quality (point 40 of guidance).[9]

Clinical trials that have successfully integrated healthcare data include the Salford Lung Study,[10, 11] the Platform Randomised trial of Interventions against COVID-19 in older people (PRINCIPLE)[12], the Randomised Evaluation of COVID-19 Therapy trial (RECOVERY),[4, 13-15] the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS),[16, 17] and A Study of Cardiovascular Events in Diabetes (ASCEND).[18, 19] These pioneering trials show the

potential value of healthcare systems data, especially in the UK; the latter three are presented as case studies to demonstrate the important role of NHSD datasets in delivering timely and efficient trials (**Section 6**).

The main advantages of healthcare data being considered acceptable as high-quality data for use in clinical trials, recognised as a transcribed copy of the original source data, are:

- For sponsors: Healthcare systems data can be used as trial data, in compliance with Clinical Trials regulations, with data collection becoming simpler, potentially more complete, more efficient, and less costly, especially for large multicentre trials and trials with long follow-up periods.
- For investigators: The burden on NHS site staff is reduced so they can focus on patient care and collection of data that are unavailable from healthcare data, such as patient reported outcomes.
- For data providers: They can provide details of the provenance, lineage and integrity of their datasets enabling their use in clinical trials.
- For public benefit: More trials can be run efficiently because of the use of centralised national datasets, to support innovation, research and development of better, and potentially cheaper treatments, which can be made available to the NHS in a timely manner.

3.2 Selected NHS Digital datasets

NHS Digital (NHSD) has a statutory role to collect, analyse and publish health data and to provide technical infrastructure to support clinicians at work, help patients get the best care, and use data to improve health and care.[20] It hosts over 200 data assets, many of which are used or have potential use to supplement data collection in clinical trials. As of June 2021, 37 national datasets were available to successful applicants through NHSD's Data Access Request Service (DARS), and more datasets will be made available in the near future.[21]

This paper selected two datasets for initial focus:

1. HES Admitted Patient Care (HES APC)
2. Civil Registration of Deaths (CRD)

HES APC is a mature and complete NHSD dataset, populated from the Secondary Uses Service (SUS+) database comprising Commissioning Data Sets (CDS) centrally gathered from NHS trusts in England for payment and monitoring purposes.[22-25] The data fields included in HES APC are mandatorily submitted by trusts, such as diagnostic information coded as ICD-10 and OPCS-4,[26] in accordance to an Information Standard and NHS

Business rules.[27, 28] HES APC along with its live precursor database, SUS+, is one of the most impactful data assets for improving health and care.[6]

The **CRD** is based on the death registration database held by the Office for National Statistics (ONS) which captures all deaths formally registered and medically certified in England and Wales. Deaths should be statutorily registered within five days of the date of death, so the ONS database is contemporaneous and complete, and therefore a reliable source of mortality data. For decades, trialists have used the ONS or NHSD death registrations (formerly the Medical Research Information Service) for patient outcomes where the primary endpoint is mortality (disease-specific or all-cause).[13, 17, 29-32]

HES APC and CRD were chosen for focus in this paper because they are the two of the most requested datasets by trialists.[2]

3.3 Regulatory definition of "source data"

Source data is defined in ICH GCP E6 (R2) as "original records or a certified copy of original, necessary for evaluation and reconstruction of a trial".[33] ICH GCP also says source data should follow the ALCOA principles of being: Accurate, Legible, Contemporaneous, Original, Attributable, Complete (ICH E6 R2 section 4.9.0). Source data are used by investigators for accurate and prompt recording of data which are then used to complete case report forms (CRFs). Sponsors review source data for quality control and to verify the reliability and accuracy of reported trial data (ICH E6 R2 section 2.10, 2.13, 4.9, 5.18).

The regulators, European Medicines Agency (EMA) and the Food and Drug Administration (FDA) use the ***"ALCOA+ principles"*** to describe electronic source data with 12 specific requirements for assessing systems that create and/or capture electronic clinical data based on the Clinical Data Interchange Standards Consortium (CDISC) standard.[34, 35] These have been tabulated in **Appendix 1**.

We note that we initially sought to identify how and whether HES APC can fulfil the ALCOA+ dimensions. However, the CDISC requirements were devised for the qualitative assessment of electronic patient administration systems (PAS) as source data, and therefore are not suitable for assessing centrally curated datasets such as those held by NHSD. We found that HES APC could not be defined as electronic source data or a verified copy of source for several reasons, for example:

1) the Information Standard and Business rules used to abstract CDS (forming the basis of HES) leave its creation and modification audit trail within the PAS; and

2) centralised processing of HES by NHSD is away from the investigator's control.

Further reasons are given in **Appendix 1**.

3.4 Revision of ICH GCP and MHRA's GXP data integrity guidance

The original version of ICH's GCP E6(R1) guidance in 1996 was finalised when most patient healthcare records were paper-based, and all trials were similarly limited to paper records. The 2016 revision to ICH E6(R2) focused on the protection of trial participants, central monitoring and data integrity through the validation of computerised systems.[33] The ICH Expert Working Group for the forthcoming third revision recognises that E6(R2) did not go far enough to address new technological advances, the digital age of trials, and innovative trial designs.[36] Consequently the vision for ICH E6(R3) is for the guidance to be flexible to allow for and to encourage innovation while helping to ensure the protection of trial participants and the reliability of trial results.[37] The draft principles of E6(R3) were published in April 2021, and the guideline intends to be "media neutral to enable the use of different technologies for the purposes of documentation", and acknowledges "the use of a variety of relevant data sources in clinical trials"[38], including patient/disease registries and electronic medical records.[38]

Section 10 of the draft revised ICH E6(R3) guidance describes the principles for ensuring trials generate reliable results.[39] Sponsors will need to demonstrate the integrity and reliability of the trial data, by providing assurances regarding compliance with data protection requirements, the control of processes and fitness of systems that manage the data, as well as data lineage (traceability) and security arrangements. This potential understanding of "trial data" is a paradigm shift from the definition of previous E6 versions, and would represent a move towards the risk-based approach of the MHRA's GXP data integrity guidance.[40] These principles apply to healthcare data used within trials, so aligned guidance on data integrity documentation would be helpful for sponsors, investigators, and data providers. Further elucidation on the secondary use of healthcare data is expected in Annex 2 of E6(R3), which is currently planned for release by the ICH Expert Working Group in late 2022.[40] In the meantime, sponsors, investigators and data providers need clarity on what is required by regulators to show that a specific healthcare dataset is "relevant" and of "sufficient quality" for clinical trials, particularly if it supports a regulatory submission.

The MHRA's GXP data integrity guidance is helpful as it adopts a risk-based approach to data management, including data integrity risk, criticality and lifecycle[1].[41] It states that organisations need to take responsibility for their systems and their data, and are expected to "*implement, design and operate a documented system that provides an acceptable state of control based on the data integrity risk and supporting rationale*". The example given is a data integrity risk assessment which maps out data flows and formats, their controls, and documents criticality and inherent risks. In that guidance, "data" is defined using ALCOA+, but the guidance also emphasises the importance of data governance measures to ensure data are complete, consistent, enduring and available through the lifecycle.

3.4.1 Governance of trial data: essential role of the data provider

The role of the data provider also requires further consideration as sponsors and investigators design more trials that use centrally curated healthcare data; this issue is **outside** the scope of this paper, but is important for clinical trials to document given the following:

- Data providers, such as NHSD, ensure their datasets are of high-quality and accessed securely and in accordance with data protection legislation.
- ICH E6(R2) states that the local investigator is responsible for accuracy, completeness, legibility, timeliness of CRF data, and must sign and date completed CRFs and any corrections.[33] If routine data is accepted as a form of CRF data, then further consideration of how it becomes "validated" is needed.
- ICH E6(R2) is unclear regarding who is responsible for data collated centrally by data providers when it is no longer under the control of the local investigator.

[1] Lifecycle is defined as "*all phases in the life of the data from generation and recording through processing (including analysis, transformation or migration), use, data retention, archive/retrieval and destruction."*

4 ASSESSMENT PROCESS

The ICH E6 principles (R2 and draft R3) and data integrity guidance require sponsors to demonstrate the integrity and reliability of trial data, including healthcare data used for trial purposes. This requires documentation of the tools and systems used, showing efficient and controlled processes for managing data, data lineage and access arrangements. There was no established procedure for assessing healthcare datasets as "accurate or transcribed copies of the original source data", we present the available evidence on the systems that manage the data to demonstrate the integrity, robustness, and reliability of the datasets, and therefore their "suitability of use" in clinical trials. The following section focuses on three key stages in the data lifecycle:

1) collection and transfer of data from healthcare systems to NHSD's systems,
2) centralised processing and curation to form the validated dataset, and
3) linkage and extraction for the end user.

The governance of data within NHSD will also be briefly described.

Since the initiation of this paper, the Food and Drug Administration of the USA has published draft guidance on assessing RWD to support regulatory decision-making. It states that sponsors must ensure the relevance and reliability of RWD used in clinical studies: the selection of data sources, the validation of definitions for study elements such as exposure and outcomes, and the data provenance and quality (accrual, curation, and transformation). Our assessment process aligns closely with this last aspect.[42]

4.1 Out of scope

The assessment of the quality of data recorded within the healthcare systems before the collection stage of the data lifecycle is out of the scope of this paper. In terms of data integrity, NHSD is responsible for the systems and the data they generate, but not responsible for the NHS Trusts' data and systems (i.e., original PAS data/patient notes) because this is not in their jurisdiction. They are not responsible for ensuring health professionals are entering data consistently, accurately, and completely. However, NHSD has a statutory role to assess the extent that collected data meets defined national standards and to publish results of these assessments in the form of the Data Quality Maturity Index (DQMI; see **Section 5.1.1** for further details).[43]

5 DATA INTEGRITY OF SELECTED NHS DIGITAL DATASETS

For brevity, acronyms are used throughout this section. They are expanded at first use and in the **Glossary of terms**.

5.1 Evidence for Hospital Episode Statistics Admitted Patient Care dataset

In this section, as an example to demonstrate how this approach can be applied in practice, ***we provide evidence and documentation that Hospital Episode Statistics Admitted Patient Care (HES APC) is of sufficient integrity to be of use directly as data for clinical trials***. The APC dataset of SUS+ is the live precursor to HES APC, so the evidence also covers this dataset.

The management of HES datasets within the NHS Digital Data Warehouse is similar to that of trial datasets by trial teams (**Figure 1**). The data are abstracted from the originating healthcare provider, then brought into a central database where they undergo validation and quality checks; and then to resolve anomalies, data queries are generated which are sent to the originating healthcare provider. Consequently, amended data may be submitted to the database, and checks are run again until the data are considered “clean” or no further resolution is possible. At set time points, the data become “fixed”, similar to a data freeze or database lock in trial databases so that the dataset can be extracted and analysed.

At the time of this paper’s publication, these processes are documented and published on the NHS Digital website. However, in 2021, NHSD developed a central metastore using the Collibra tool[2] where the complete lineage of their current data assets including permissions, governance, data quality rules, process flow and storage are recorded and managed. This metastore for data management can provide information on data integrity and will allow trial sponsors to see complete visual representations of how these datasets are abstracted, curated, and linked for data access. This work is ongoing, but examples of how Collibra visualises data lineage and validation of HES APC are provided in **Appendix 3** and **Appendix 4**.

[2] https://www.collibra.com

Figure 1: Schematic of data flows in a clinical trial using both trial-specific data collection and NHS Digital (NHSD) Hospital Episode Statistics (HES) Admitted Patient Care (APC) datasets[3]

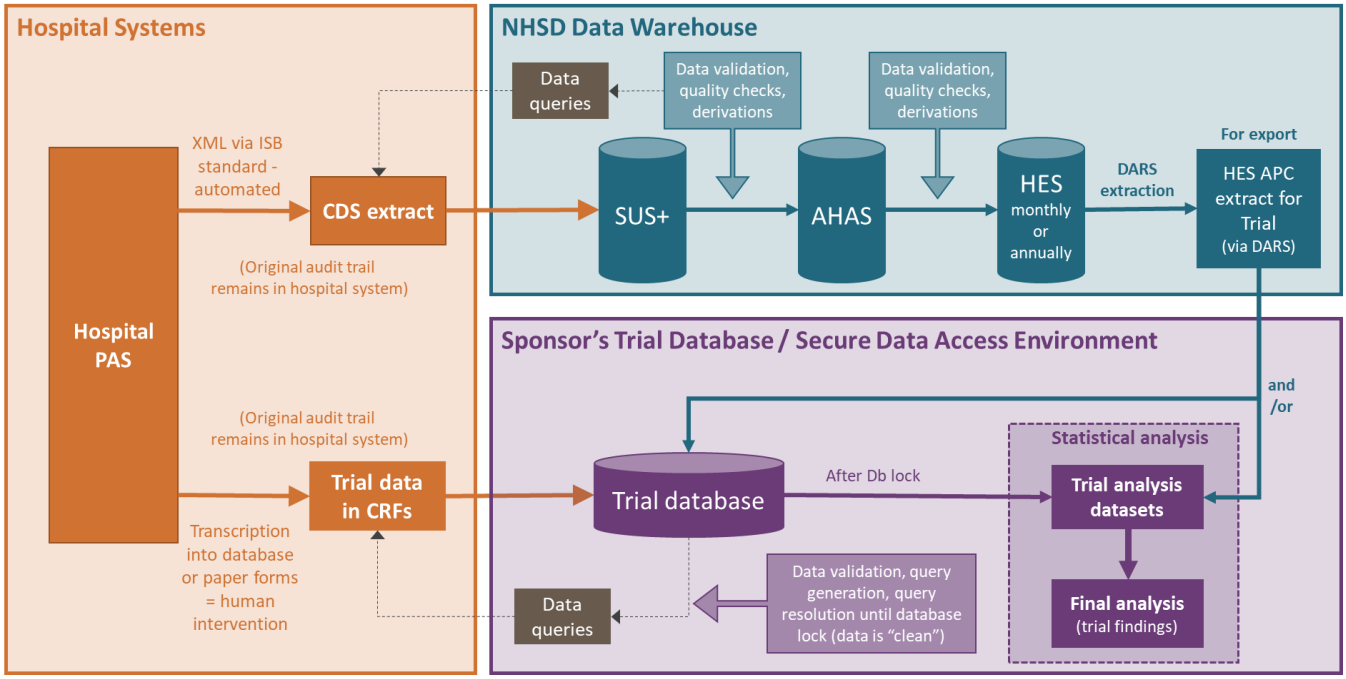
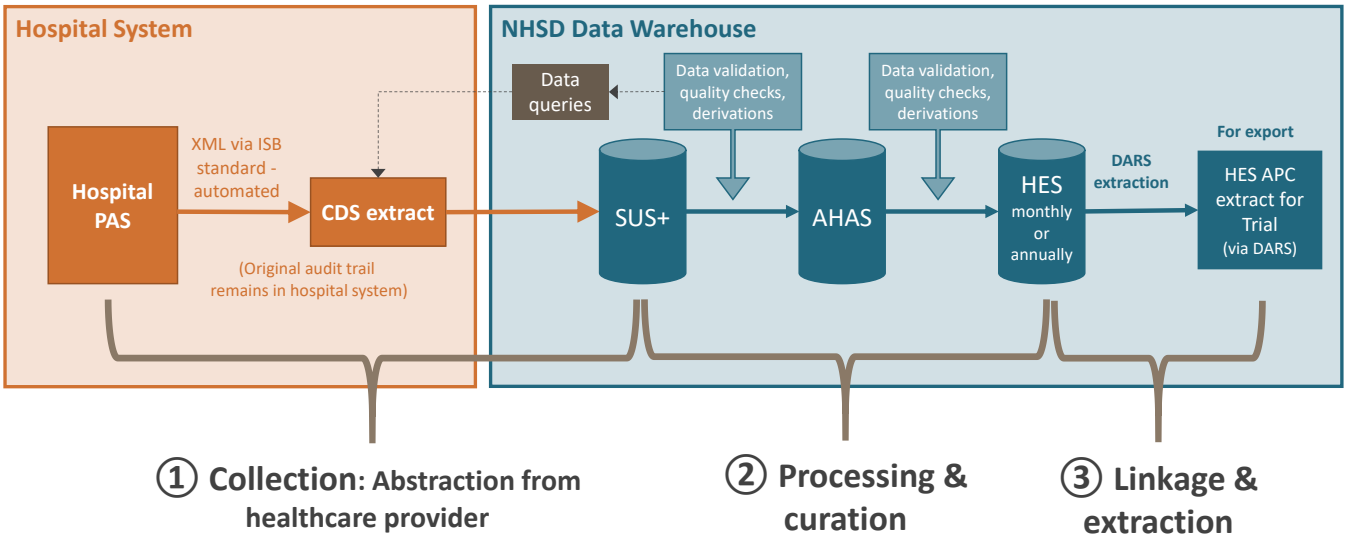


Figure 2: Data management stages in Hospital Episode Statistics (HES)[4]



[3] NHSD Data Warehouse is not presented in full here; only the relevant aspects concerning HES data flow are included. AHAS: Acute Hospital Analytical System; CDS: Commissioning Data Sets; CRF: case report form; DARS: Data Access Request Service; ISB: Information Standards Board for Health and Social Care; PAS: Patient Administration System; SUS+: Secondary Uses Service.

[4] Same abbreviations as in footnote (d).

5.1.1 Stage 1: Collection from acute NHS Trusts

The data that form HES APC are submitted from hospital-based PAS typically at least twice monthly as CDS into SUS+ (a live precursor to HES), using an XML schema adhering to an Information sStandard (ISB 0092) and CDS Business Rules comprising rules for SUS+ and NHS Data Standards (**Figure 2**).[20, 24, 27,28] Therefore CDS is likely to be highly accurate, and is a copy of the same data fields as that in the PAS.[24, 27, 28] Data are coded with ICD-10 and OPCS-4 (according to published coding standards) before submission to NHSD. Up-to-date technical details of CDS submissions by healthcare providers to SUS+ are available from the NHSD website, along with the submission timetable.[24, 44, 45] SUS+ has capacity to handle as many as 12 million CDS records submitted daily.

CDS (current version 6.2) are condensed datasets which comprise specific data fields, so not all fields available in the PAS records are abstracted. There are different PAS software used by acute NHS Trusts, such as Lorenzo, Cerner, and EPIC, so only the necessary and common fields are abstracted as defined by the Information standard.[27] In accordance with CDS Business Rules, information on the creation and modification of records within the PAS (the hospital audit trail) is not submitted with CDS.

CDS is used for statutory administrative purposes, including reimbursement of healthcare providers via Payment by Results (PbR).[46] PbR uses a data assurance framework at acute NHS Trust level to audit clinical coding for accuracy and reliability. This involved a comparison of SUS+ data with source documentation at Trusts.[47] There have been notable improvements in recording at acute Trusts over time, such as clinical coding, and this framework provides further evidence that the CDS data submitted to SUS+ are reliable and complete.[48] The quality of patient records created by healthcare providers is out of this project's scope, however NHS staff adhere to the NHSD's Provider Data Quality Assurance Framework and the NHSX code of records management.[49-51] Providers are also financially incentivised by the Commissioning for Quality and Innovation scheme which promotes robust and timely recording of clinical activity.[52] It should be acknowledged that the primary purpose of hospital PAS is for patient administration, not research purposes. In accordance to the Care Act 2014,[53] healthcare providers have a legal obligation to supply accurate information, and in its statutory role, NHSD has to assess the extent that collected data meets applicable published standards, and to publish these assessments (in the form of the DQMI).[20, 50] The DQMI is published by the Data Quality team at NHSD, and it is based on the completeness and validity of core data items including NHS number, date of birth, gender, postcode, speciality, and consultant. The methodology of the index along with a full list of core data items in CDS is available on the NHSD website.[43] If poor quality data is identified

by the DQMI, the Data Quality and Data Liaison teams work with healthcare providers to encourage greater attention to data collection and submission.[54]

5.1.2 Stage 2: Processing and curation

There are two flows of data from SUS+ to HES APC; one is through a legacy system (from 2014) in an Oracle data warehouse, decommissioned by NHSD in July 2021;[51] the second data flow is through the cloud-based platform, Data Processing Service (DPS), initiated in 2018 and fully implemented in January 2020 (**Appendix 2**). The processing and data quality rules are applied equally in both systems.

5.1.2.1 Legacy system

The processing of SUS+ to form static HES data are described in several documents from the HES Data Quality (DQ) Team published in 2016 (and still applicable), covering processing, data quality and removal of duplicate records.[55-57] The HES Autocleans dictionary describes in detail the rules for data cleaning and derivation of HES data from SUS+;[40] each rule is numbered so they can be referenced in the HES Data Dictionary.[23, 54] Provisional monthly HES datasets are generated and there is one annual update (when records become fixed), usually in March each year; it takes NHSD approximately 40-45 days to generate HES output from submitted hospital data in the form of CDS. The HES DQ team has information to show monthly data do not change significantly.[58] In addition, HES Data Quality Notes[59] and the DQMI[43] are published monthly to highlight specific known issues which should be considered when using the data, and it also provides detailed information on organisation coverage and completeness of data fields. This is comprehensive approach to assessing data quality, and users of these datasets can be aware of their strengths and limitations.

5.1.2.2 Data Processing Service (DPS) platform

DPS standardises and automates the processing of health and care data at NHSD, and is used for collection, person matching, controlled linkage, de-identification and re-identification. It can also be remotely accessed by authorised personnel. This platform improves data integrity and allows for better governance of the datasets held.

DPS core processing assesses the quality of data submitted to SUS+ from healthcare providers, through the Acute Hospital Analytical System (AHAS)[5]. AHAS is populated cumulatively from SUS+ and was developed so daily live feeds of data could be made

[5] AHAS currently holds records from October 2018 as this is newly developed; historic data abstracted between 2016 and September 2018 are held in SUS+. SUS+ data from earlier than 2016 are archived as .csv files. The Oracle Warehouse is only place where historical SUS+ data pre-2016 exists for querying, and there are plans to upload legacy data into DPS.

available. SUS+ data are passed to AHAS, and records are person-matched via the Master Person Service (MPS). The daily delta feed[6] of AHAS is available in less than 16 hours after receipt from hospital submissions (as CDS data) since AHAS does not contain all of the processing and derivations that are otherwise required for the static HES dataset. Therefore, AHAS is contemporaneous. [Note that if accessed by users, AHAS requires careful handling because it is a live cumulative database: users need to filter out original records that have a subsequent correction.]

Linkage via MPS within the DPS platform is reassuringly accurate. MPS is an enhanced four-stage algorithm for matching NHSD records to the right person.[56, 57] This automatically matches 99% of patients, and unmatched records are passed to MPS for further processing using a more complex algorithm.[60] It was developed so a common identifier (MPS person identifier; MPS Person ID) can be used across all national patient level datasets to increase the amount of linkable data available, increase efficiency through automation, and improve data quality and timeliness. Although neither are perfect, the MPS algorithm is significantly better at matching patients than the HES identifier (HESID) algorithm. From April 2021, MPS Person ID replaced the HESID as the patient identifier in HES datasets.[60] The Personal Demographics Service (PDS) database, used by healthcare professionals to identify patients, is updated daily, and MPS cross-checks NHS number and associated demographics (such as age, gender, postcode) with the PDS database to find the best record match. NHSD describes how the PDS National Back Office manages and resolves data quality issues to safeguard patients; issues include duplicate NHS numbers, confused records, and incorrect demographic data or death status.[58]

HES datasets including APC are generated from AHAS monthly followed by an annual refresh. As the methodology for processing HES has changed, a publication describing the transition and an analysis of the impact of the changes was published in December 2020.[60] Overall, these changes automate and simplify processing, so larger volumes of data can be handled faster, and data are more accurate.

5.1.3 Stage 3: Linkage and extraction

Methods of linking trial participants' identifiers to their NHSD records continue to improve over time to excellent levels, due to the MPS algorithm in DPS.[60] Legacy HES data contains

[6] The daily delta feed is the mechanism for sending differences between sequential data to AHAS (so which records are new or changed since the previous version), to update the database.

the HESID, which is used to person-match records via a three-step process[7]; this is less accurate than the MPS algorithm.[60]

Data can be accessed in three ways after successful application to DARS: through the Data Access Environment to DPS; as data extracts that are produced by the data production team in NHS Digital; or through the newly developing Trusted Research Environment (TRE) at NHSD.

Data such as AHAS, are stored in a DPS Hive, which can then be extracted to an SQL server database so external users can access it via the Data Access Environment. A similar arrangement will be available through the TRE to interrogate and access data in DPS.

Data extracts are securely transferred in and out of NHSD using SEFT (secure electronic file transfer system), with 256-bit AES encryption. The trial sponsor submits participant identifiers (usually NHS number and date of birth) in a file through SEFT, which is then used by the Data Production team at NHSD for record linkage to the requested datasets.

This is outside of this paper's scope, but sponsors need to check that the participant records of HES APC data extracts are correctly matched; this might require screening of records for known events, e.g. specific cancer diagnosis for participants of a cancer treatment trial. Although cohort validation is routinely done by the NHSD data production team prior to linkage, recording errors by research staff in participants' NHS numbers or demographics may result in linkage failure (incorrect or no matches). And so trial-specific procedures must also be applied at trial recruitment to ensure accuracy in the collection of personal identifiers for linkage by local investigating sites and/or the sponsor.

5.2 Evidence for Civil Registration of Deaths dataset (CRD)

Here, ***we provide evidence that CRD dataset is integral and reliable, and it can be used directly as mortality data in clinical trials***. NHSD published their Mortality Data Review in June 2020 which provides an overview of the data sources and systems used to record deaths in England including those operated by ONS, NHSD, NHS England, and the former Public Health England (PHE).[61] Enhancements were made to death reporting as a consequence of the COVID-19 pandemic (Coronavirus Act 2020).[62] The review describes the full reporting landscape of deaths in England: the statutory reporting process (civil registration), and the health and care system of reporting. Together, these processes

[7] Step 1 matches sex, date of birth (partial match), and NHS number. Step 2 matches sex, date of birth (partial match), postcode and provider code plus local patient identifier within provider. Step 3 attempts to match exactly sex, date of birth and postcode.

generate 12 mortality datasets, including the CRD dataset held by NHSD. The available data fields in CRD are listed in **Appendix 5**.

5.2.1 Stage 1: Collection from informants and abstraction from General Register Office

The collection and abstraction of death certification data is described in the NHSD mortality data review (**Appendix 6**) and the ONS mortality statistics guide.[61, 63] Informants provide details of the medical certificate of cause of death to the General Register Office via registrars at the local registration service. Data are captured in the Registration Online (RON) system in a structured manner with built-in validations (range, data type, logical consistency) to ensure the data are entered by registrars correctly; RON replaced the 1998 Registration Service Software in July 2009.[63] Information is also captured about the informant such as name, address and relationship to the deceased. Falsely supplied information will render informants liable to prosecution for perjury, so the provided information is widely considered correct.[64]

5.2.2 Stage 2: Processing and curation

The processing, curation and data flow from ONS to NHSD is described in the NHSD Mortality Data Review (**Appendix 6**).[61] ONS provides an overview of the data flow of death certification from the General Register Office to and through ONS.[63, 65] Detailed quality and methodology information describes the processing and curation of the data, including quality checks to look for missingness and duplication, automatic coding of geography, occupation and cause of death, resolution of data issues, quality assurance through internal consistency checks, and formal sign-off.[61-63] The systems and standards used, and the changes in methods over time (such as the replacement of ICD-9 by ICD-10) is also described, along with the impact of such changes to improve the comparability of mortality statistics across Europe and internationally.[61, 62, 64]

The ONS formal death registration database is 64% complete after 5 days, 91% by Day 14, and 99.5% complete after one year.[61] Any delay in death registration may occur due to coroner referral or inquest, which can defer the issue of the death certificate. The relatively small proportion of deaths that remain legally uncertified are often due to anomalies in certification such as: the certifying doctor not fulfilling legal requirements because they did not see the body or attend the deceased during their last illness, or the deaths of military personnel serving abroad being certified by a medical practitioner not registered with the General Medical Council in the UK.[66]

However, due to the statutory process of death registration, the ONS database is contemporaneous and is shown to be highly complete and accurate.[61] Consequently, the CRD dataset held by NHSD in the DPS platform is of similar standard since it is updated weekly from the ONS database, with an approximate delay of 3-14 days (**Appendix 6**).[61] Therefore, both ONS and NHSD death registration datasets are considered of sufficient data integrity to be used as trial data for mortality outcomes.

5.2.3 Stage 3: Linkage and extraction

The methods of linkage and extraction of CRD is similar to those of HES APC as described in **Section 5.1.3.**

5.3 Other considerations for HES APC and CRD

5.3.1 Data governance

Trialists requiring access to NHSD data must apply for access through the NHS DigiTrials Service or directly with the DARS team, using an online structured application that clearly describes the purpose and legal basis for data access (in accordance with UK GDPR). NHSD provides guidance to applicants on the information expected in data applications; these are known as “Standards” and cover important items such as the objective for processing, security assurance, and consent.[67]

The Independent Group Advising (NHSD) on the Release of Data (IGARD) is an advisory body to the NHSD Board and has two purposes: one is to make general recommendations or observations to NHSD about processes, policies, and procedures relating to data disseminations from NHSD, including transparency measures such as the data uses register; two, to independently scrutinise and advise NHSD on the appropriateness of requests for dissemination of confidential information as defined in Section 263(2) of the Health and Social Care Act 2012, including

- personal data as defined in the UK General Data Protection Regulation and Data Protection Act 2018,
- personal Information as defined in the Statistics and Registration Services Act 2007,
- data which is pseudonymised, anonymous in context or which is de-identified for limited access,
- data which is aggregated but which does not have small numbers suppressed.[20, 68]

The role of IGARD ensures that the use of patient data within research, academia, the public and private sector is in controlled environments, where any risks of disclosure are minimised.

The Information Assurance and Cyber Security Committee at NHSD is responsible for providing strategic direction and effectiveness of cyber security, information assurance and information governance risks and operations. It has specific terms of reference, reporting to the NHS Digital Board, and meeting at least four times per year.[69] This robust approach allows NHSD to detect and address data integrity weaknesses and potential failures, as well as security concerns.

5.3.2 Relevant data for the trial (i.e. is it fit-for-purpose)

Relevancy of the dataset is a trial-specific consideration.[42] The outcomes of death can be identified from CRD in a straightforward manner. However, some outcomes are more difficult to capture from healthcare data either because of lack of granularity in coding, the recording by healthcare providers may be unclear or inconsistent (such as hospital attendance at A&E for self-harm[70, 71][8]) or the information is not routinely recorded in patient records, such as fact and time of cancer recurrence. It is outside the scope of this paper to discuss the relevancy of datasets for specific trial outcomes, but further work in this arena is occurring nationally through the Trial Methodology and Research Partnership Health Informatics Working Group. The Group is undertaking a series of studies-within-a-trial to compare trial-collected and healthcare-data outcomes for diagnostic value. This follows on from existing studies that have assessed data fitness, accuracy and completeness for trial-specific outcome measures.[17, 70, 72, 73]

5.3.3 HDR UK data utility and Collibra

Health Data Research UK defined a data utility framework to assess the usefulness of a health dataset for a specific purpose and to objectively evaluate a dataset, with the overall aim to improve the quality, use, and access to health data. It covers data documentation, technical quality, coverage, access and provision, and value and interest,[74] and many of these categories are addressed for HES APC and CRD in this paper. This framework has been integrated within NHSD's Collibra Metastore so that information about each accessible dataset will be available to researchers and the public through HDRUK's Innovation Gateway.

5.3.4 Other healthcare purposes

HES APC is widely used by health and social care organisations, local authorities, and commissioners for supporting policy and planning, and improving health and care, and is a valuable resource. It is used to help commissioners and local authorities to understand the

[8] The quality of coding in HES A&E is less reliable and consistent than that of HES APC. The A&E dataset was succeeded by the ECDS in October 2017 to improve the validity of diagnoses and reduce the use of vague terms.

needs of local populations, and NHS organisations use it to monitor their performance and the quality of care given, as well as better plan their services. A clinical review describing the impact of data released through DARS gives detailed examples of the usefulness of HES data to impact service improvements and measure population health outcomes. This shows the value of HES APC and other NHSD data assets beyond their use in clinical trials.[6]

5.4 Conclusions on the integrity of HES APC and CRD

The evidence collated on the management of HES APC and CRD by NHSD demonstrate that these are integral datasets, handled robustly with appropriate controls and automation to assure their quality for secondary uses, including clinical trials. Recent technological developments in the form of the DPS Platform and associated services have improved timeliness and accuracy, so that sponsors and trial teams who choose to use the datasets can be assured of their integrity and reliability as transcribed copies of the original source data. It is the responsibility of sponsors to review the integrity of the routine datasets that are used in their trials including the key lifecycle stages, and clearly document their decisions on the acceptability of the datasets (caveats and exceptions) within the Trial Master File.

6 CASE STUDIES OF NHSD DATA ASSETS USED IN CLINICAL TRIALS

6.1 UKCTOCS

UKCTOCS is a multicentre randomised non-CTIMP trial of over 200,000 women aged 50-74 years randomly allocated between April 2001 and September 2005 to two annual screening groups for ovarian cancer (multimodal or ultrasound) or a no screening group.[16] The trial's aim was to gather evidence on whether screening reduced deaths from ovarian cancer.

The sponsor's trial team received consent from participants to be followed up through multiple sources including national registries and healthcare systems data such as cancer and death registrations (CRD) from NHSD, HES, National Cancer Intelligence Network dataset, direct communication from participants, and local trial teams using CRFs. Most participants (202,632 of 202,638; >99.999%) were electronically flagged using their NHS number for cancer and death registrations within national registries of England, Wales, and Northern Ireland (no sites in Scotland participated). To ascertain the feasibility and reliability of healthcare data for follow-up, two studies were performed comparing trial-collected outcomes on colorectal cancer, ovarian cancer and death with corresponding information from healthcare data and registries.[17, 72]

Both studies used HES (APC and Outpatients), cancer registration and CRD datasets to obtain diagnoses of malignant neoplasms of the colon or rectum, ovarian cancer, or death caused by either condition recorded after randomisation. The colorectal cancer study used cancer registration and CRD to May 2011 and HES data to July 2010, and the ovarian cancer study used later data extracts to March 2015 and had an independent outcomes review committee to confirm ovarian cancer diagnosis and death. The trialists report the sensitivity of the cancer registry for colorectal cancers registered within one year was 92% (453/491; 95% CI 90-94%) and was 99% (485/491; 95% CI 97-100) within six years. The sensitivity of CRD for colorectal cancer deaths was 97% (98/102; 95% CI 92-99) and specificity was reported as 97% (6968/7183; 95% CI 97-97%). In HES, 82% (327/397; 95% CI 78-86%) of confirmed colorectal cancers diagnosed before July 2010 were recorded, of which 11 were unique. The use of HES in addition to the registration datasets provided more comprehensive information on diagnoses of colorectal cancers and deaths (sensitivity 98% [388/396; 95% CI 96-99%]). Hence a composite approach of obtaining colorectal cancer and mortality outcomes from cancer registration, CRD and HES was recommended by the trial team. Where differences between data sources were identified, an adjudication committee reviewed the data.[72]

Kalsi *et al.* found sensitivity and specificity of ovarian cancer registration were 85.0% (1125/1324; 95% CI 83.7-86.2%) and 94.0% (1679/1786; 95% CI 93.2-94.8%) respectively.[17] For ovarian cancer death registration, sensitivity was 88.8% (605/681; 95% CI 86.4-91.2%) and specificity was 96.7% (1482/1533; 95% CI 95.8-97.6%). HES provided further information on additional cases and deaths due to ovarian cancer, so when multiple healthcare datasets were considered, sensitivity for cancer site increased to 91.1% (1206/1324; 95% CI 89.4-92.5%) and for cause of death 94.0% (640/681; 95% CI 91.9-95.5%). Furthermore, the level of agreement between CRD and the confirmed outcomes review of the cause of death was substantial (86.3%; 1763/2041; kappa 0.78; 95%CI 0.76-0.80). These studies demonstrate the accuracy and reliability of HES, cancer and death registration datasets to obtain cancer and mortality outcomes for trials.[17]

6.2 RECOVERY

RECOVERY is the UK's national clinical trial of treatments for patients hospitalised with COVID-19, recruiting over 40,000 patients across 177 hospitals. By June 2021, the trial had demonstrated that three treatments save lives in patients hospitalised with COVID-19 (dexamethasone,[13] tocilizumab,[75] REGEN-COV[76]) and shown that six treatments were not effective (hydroxychloroquine,[14] lopinavir-ritonavir,[77] azithromycin,[78] convalescent plasma,[79] colchicine,[80] and aspirin[81]) avoiding unnecessary patient exposure and allowing healthcare systems to direct resources to treatments which are effective.

Effective use of UK healthcare data has been a central component of the trial's success. Data from a one-page form completed by the local study staff at randomisation and 28 days later (with one further form required for some treatments) is integrated with data from over 25 routine datasets across the UK, including primary care, secondary care, prescribing data, critical care data, mortality data, COVID-19 testing data and others. This enables:

- Complete follow-up of participants, irrespective of where they were transferred for care
- A reduced burden on front-line staff, by avoiding lengthy questionnaires
- Additional scientific questions to be answered, including the effects of the treatment on long-term health outcomes such as 'long-COVID'

Collaborations have been established with national data custodians, including, Public Health Scotland, National Records of Scotland, NHS Digital, the SAIL Databank at the University of Swansea, the Intensive Care National Audit and Research Centre (ICNARC), the UK Renal Registry and others. NHSD was able to support the trial by pivoting their resources to accelerate the availability of datasets including SUS+ APC, HES APC, and PDS for participants in England, and CRD for participants in England and Wales.[4, 82]

Algorithms were developed to extract additional baseline data (for example, participant ethnicity) along with the primary, secondary and some of the subsidiary outcomes for all participants in England, Wales, and Scotland. A detailed description of the data sources and algorithms used is provided in the 'Definition and Derivation of Baseline Characteristics and Outcomes'.[82]

In addition to facilitating long-term follow-up and other analyses, benefits from this process included:

- Confirmation of primary outcomes, avoiding source data verification (for example, site queries about the primary outcome, all-cause mortality by 28-days, are targeted to those reports from the investigating site which are *not* later confirmed by linked healthcare data).
- Improved completeness. Of 8,000 deaths within 28 days of randomisation recorded by May 2021, 500 (1 in 16) deaths were identified solely based on the linked healthcare data and would otherwise have been missed or delayed.
- Rapid ascertainment of outcomes for the data monitoring committee analyses with notification of deaths in almost real-time from central NHS systems.

6.3 ASCEND

ASCEND (A Study of Cardiovascular Events iN Diabetes) is a double-blind randomised controlled trial assessing the effects of aspirin and, separately, in a two-by-two factorial design, of supplementation with omega-3 fatty acids, on serious vascular events in people with diabetes but who did not have prior atherosclerotic cardiovascular disease at randomisation.[18, 19] The study used novel mail-based methods with collection of study outcomes from participant, or, if necessary, General Practitioner, questionnaires, with central collection of supporting documentation and adjudication of study outcomes blind to treatment allocation.[18, 19, 83] The primary outcome is serious vascular events (a composite of cardiovascular death, non-fatal stroke, non-fatal myocardial infarction or transient ischaemic attack, excluding intra-cranial haemorrhage), with a secondary outcome of serious vascular event or arterial revascularisation. The key safety outcome was major bleed, including sight threatening eye bleeding, intracranial haemorrhage or any other bleed which was fatal or required hospitalisation. Over 90% of the primary, secondary and main safety outcomes included in the analyses were confirmed by the adjudication process.[18]

Between 2005 and 2011, 15,480 participants were randomised and then followed for a mean of 7.4 years. The main follow-up phase completed in 2017 with publication of the primary results in 2018, [18, 19] and long-term follow-up of the cohort is ongoing. Linkage with HES APC

and CRD (or equivalent datasets in Wales and Scotland) was available for over 99% of the study cohort. The occurrence of a serious vascular event or revascularisation was ascertained on the basis of ICD-10 or OPCS codes in the linked hospitalisation and mortality data, and the accuracy of these healthcare data outcomes were compared with the reference method of ascertainment from trial follow-up and adjudication systems.

Restricting to the 1099 serious vascular events which were fatal or led to hospitalisation, outcomes obtained from healthcare data performed well compared to the adjudicated follow-up methods (sensitivity 86.4% [95% CI 84.4%-88.5%]; specificity 98.8% [98.6%-98.9%]; kappa statistic 0.84 [0.83-0.86]). Furthermore, for arterial vascularisation procedures ascertainment from the healthcare data was almost identical to the adjudicated follow-up methods (sensitivity 94.6% [93.0%-96.3%]; specificity 99.7% [99.6%-99.7%]; kappa statistic 0.94 [0.92-0.95]). (manuscript under review). Additional work to assess the utility of these datasets to ascertain major bleed and other outcomes in ASCEND, along with data utility projects across several other completed clinical trials, are underway.

7 RECOMMENDATIONS

7.1 For HES APC and CRD datasets

The data lifecycles of HES APC and CRD were carefully documented, and their integrity confirmed. Therefore, these datasets should be considered as accurate, transcribed copies of the original source data and may be suitable for use as clinical trial data subject to relevance. This is a key step in de-duplicating some of the unnecessary effort currently required in conducting clinical trials.

Trial teams should understand that these datasets are effectively the data from the originating healthcare provider and the data that are being used to manage patient care, hence no source data verification should be required. Instead, sponsors may prioritise processes to assess the validity and relevance of routine data. Finally, sponsors should document in their protocol and Trial Master File why they have chosen to use these datasets, describing any validation studies and other evidence of suitability, and refer to this paper for their data integrity.

7.2 For other healthcare datasets

The structured descriptions of the data lifecycle provide a template for other healthcare data providers to follow. These should be used by NHSD and other data collators to characterise their other data assets. The assessment findings for other datasets should be made publicly available. Data collators that cannot demonstrate this quality measurement should consider steps to rectify this.

8 REFERENCES

1. Sydes MR, Barbachano Y, Bowman L*, et al.* Realising the full potential of data-enabled trials in the UK: a call for action. *BMJ Open* 2021;11(6):e043906.

2. Lensen S, Macnair A, Love SB*, et al.* Access to routinely collected health data for clinical trials - review of successful data requests to UK registries. *Trials* 2020;21(1):398.

3. McKay AJ, Jones AP, Gamble CL, Farmer AJ, Williamson PR. Use of routinely collected data in a UK cohort of publicly funded randomised clinical trials. *F1000Res* 2020;9:323.

4. Pinches H. NHS DigiTrials: how the search for COVID-19 treatments is revolutionising clinical trials. The Pharmaceutical Journal [Internet]. 2021 12 April 2021; 306(7948). Available from: https://pharmaceutical-journal.com/article/opinion/nhs-digitrials-how-the-search-for-covid-19-treatments-is-revolutionising-clinical-trials.

5. Harron K, Gamble C, Gilbert R. E-health data to support and enhance randomised controlled trials in the United Kingdom. *Clin Trials* 2015;12(2):180-182.

6. Foley TL, M. Clinical Review: The impact of data released through the Data Access Request Service. NHS Digital; 01 November 2019. Available from: https://digital.nhs.uk/data-and-information/data-insights-and-statistics/dars-clinical-review [accessed 24 July 2020].

7. Coughlan CH, Ruzangi J, Neale FK*, et al.* Social and ethnic group differences in healthcare use by children aged 0-14 years: a population-based cohort study in England from 2007 to 2017. *Arch Dis Child* 2021;10.1136/archdischild-2020-321045.

8. Muzambi R, Bhaskaran K, Smeeth L, Brayne C, Chaturvedi N, Warren-Gash C. Assessment of common infections and incident dementia using UK primary and secondary care data: a historical cohort study. *Lancet Healthy Longev* 2021;2(7):e426-e435.

9. Medicines and Healthcare products Regulatory Agency (MHRA). Consultation document: MHRA draft guidance on randomised controlled trials generating real-world evidence to support regulatory decisions. London: MHRA; 30 October 2020. Available from: https://www.gov.uk/government/consultations/mhra-draft-guidance-on-randomised-controlled-trials-generating-real-world-evidence-to-support-regulatory-

decisions/consultation-document-mhra-draft-guidance-on-randomised-controlled-trials-generating-real-world-evidence-to-support-regulatory-decisions [accessed 08 December 2020].

10. Bakerly ND, Woodcock A, New JP*, et al.* The Salford Lung Study protocol: a pragmatic, randomised phase III real-world effectiveness trial in chronic obstructive pulmonary disease. *Respir Res* 2015;16:101.

11. Bakerly ND, Woodcock A, Collier S*, et al.* Benefit and safety of fluticasone furoate/vilanterol in the Salford Lung Study in chronic obstructive pulmonary disease (SLS COPD) according to baseline patient characteristics and treatment subgroups. *Respir Med* 2019;147:58-65.

12. Yu L-M, Bafadhel M, Dorward J*, et al.* Inhaled budesonide for COVID-19 in people at higher risk of adverse outcomes in the community: interim analyses from the PRINCIPLE trial. *medRxiv* 2021;10.1101/2021.04.10.21254672:2021.2004.2010.21254672.

13. Recovery Collaborative Group, Horby P, Lim WS*, et al.* Dexamethasone in Hospitalized Patients with Covid-19. *N Engl J Med* 2021;384(8):693-704.

14. Recovery Collaborative Group, Horby P, Mafham M*, et al.* Effect of Hydroxychloroquine in Hospitalized Patients with Covid-19. *N Engl J Med* 2020;383(21):2030-2040.

15. Donnelly T. NHS DigiTrials: Already saving lives. NHS Digital; 10 May 2021. Available from: https://digital.nhs.uk/features/nhs-digitrials-already-saving-lives [accessed 04 June 2021].

16. Menon U, Gentry-Maharaj A, Burnell M*, et al.* Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *Lancet* 2021;10.1016/S0140-6736(21)00731-5.

17. Kalsi JK, Ryan A, Gentry-Maharaj A*, et al.* Completeness and accuracy of national cancer and death registration for outcome ascertainment in trials-an ovarian cancer exemplar. *Trials* 2021;22(1):88.

18. The ASCEND Study Collaborative Group. Effects of Aspirin for Primary Prevention in Persons with Diabetes Mellitus. *New England Journal of Medicine* 2018;379(16):1529-1539.

19. The ASCEND Study Collaborative Group. Effects of n−3 Fatty Acid Supplements in Diabetes Mellitus. *New England Journal of Medicine* 2018;379(16):1540-1550.

20. Health and Social Care Act 2012, c.7. 27 March 2012. Available from: https://www.legislation.gov.uk/ukpga/2012/7/contents/enacted [accessed 13 July 2021].

21. NHS Digital. Available from: https://digital.nhs.uk/ [accessed 04 June 2021].

22. NHS Digital. NHS Data Model and Dictionary. NHS Digital; 29 March 2021. Available from: https://www.datadictionary.nhs.uk/index.html [accessed 10 June 2021].

23. NHS Digital. HES Data Dictionary: Admitted Patient Care. Leeds, UK: Health and Social Care Information Centre; 17 November 2020. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hes-data-dictionary/dd-apc_v12.pdf [accessed 26 November 2020].

24. NHS Digital. SUS+ essentials: Secondary Uses Service user guide. NHS Digital; 04 January 2021. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/services/sus/sus-guidance/sus-essentials-v13.1.pdf [accessed 10 March 2021].

25. Department of Health (DH) Payment by Results team. A simple guide to Payment by Results. Leeds, UK: DH; November 2012. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/213150/PbR-Simple-Guide-FINAL.pdf [accessed 05 May 2021].

26. NHS Digital. Clinical Classifications. NHS Digital; 22 June 2021. Available from: https://digital.nhs.uk/services/terminology-and-classifications/clinical-classifications [accessed 06 July 2021].

27. Information Standards Board for Health and Social Care (ISB). ISB 0092 Amd 16/2010 Commissioning Data Sets (CDS) Version 6.2. ISB; 06 September 2012. Available from: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-

notifications/standards-and-collections/isb0092-commissioning-data-sets [accessed 26 February 2021].

28. Health and Social Care Information Centre (aka NHS Digital). Commissioning Data Set Business Rules. NHS Digital; 29 March 2021. Available from: https://www.datadictionary.nhs.uk/supporting_information/commissioning_data_set_business_rules.html [accessed 04 June 2021].

29. Bowman L, Mafham M, Stevens W*, et al.* ASCEND: A Study of Cardiovascular Events iN Diabetes: Characteristics of a randomized trial of aspirin and of omega-3 fatty acid supplementation in 15,480 people with diabetes. *Am Heart J* 2018;198:135-144.

30. Dearnaley DP, Mason MD, Parmar MK, Sanders K, Sydes MR. Adjuvant therapy with oral sodium clodronate in locally advanced and metastatic prostate cancer: long-term overall survival results from the MRC PR04 and PR05 randomised controlled trials. *Lancet Oncol* 2009;10(9):872-876.

31. The Information Centre. Medical Research Information Service [The National Archives: Archived on 27 May 2008]. The Information Centre for Health and Social Care; 2007. Available from: https://webarchive.nationalarchives.gov.uk/20080527165020/http://www.ic.nhs.uk/our-services/improving-patient-care/medical-research-information-service [accessed 15 June 2021].

32. Dahlof B, Sever PS, Poulter NR*, et al.* Prevention of cardiovascular events with an antihypertensive regimen of amlodipine adding perindopril as required versus atenolol adding bendroflumethiazide as required, in the Anglo-Scandinavian Cardiac Outcomes Trial-Blood Pressure Lowering Arm (ASCOT-BPLA): a multicentre randomised controlled trial. *Lancet* 2005;366(9489):895-906.

33. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Integrated addendum to ICH E6(R1): Guideline for good clinical practice E6(R2). ICH; 09 November 2016. Available from: https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf [accessed 30 April 2021].

34. European Medicines Agency (EMA). Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials. London: EMA; 09 June 2010. Available from:

https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/reflection-paper-expectations-electronic-source-data-data-transcribed-electronic-data-collection_en.pdf [accessed 24 September 2020].

35. Clinical Data Interchange Standards Consortium (CDISC) Electronic Source Data Interchange (eSDI) Group. Leveraging the CDISC Standards to Facilitate the use of Electronic Source Data within Clinical Trials. CDISC; 20 November 2006. Available from: https://www.cdisc.org/system/files/all/reference_material_category/application/pdf/esdi.pdf [accessed 29 September 2020].

36. Clinical Trials Transformation Initiative (CTTI). ICH E6 Guideline for Good Clinical Practice - Update on Progress. CTTI; 18 May 2021. Available from: https://www.ctti-clinicaltrials.org/briefing-room/meetings/ich-e6-guideline-good-clinical-practice-%E2%80%93-update-progress [accessed 18 May 2021].

37. Clinical Trials Transformation Initiative (CTTI). ICH E6 Guideline for Good Clinical Practice - Update on Progress: Full web conference slide deck. CTTI; 18 May 2021. Available from: https://www.ctti-clinicaltrials.org/sites/www.ctti-clinicaltrials.org/files/5.18-19.2021_ich_update_final_slide_deck.pdf [accessed 18 May 2021].

38. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH Reflection on "GCP renovation": Modernization of ICH E8 and Subsequent Renovation of ICH E6.: ICH; January 2017. Available from: https://admin.ich.org/sites/default/files/2021-05/ICH_ReflectionPaper_GCPRenovation_2021_0519.pdf [accessed 25 June 2021].

39. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E6(R3) Expert Working Group. Good clinical practice draft principles. ICH; 19 April 2021. Available from: https://database.ich.org/sites/default/files/ICH_E6-R3_GCP-Principles_Draft_2021_0419.pdf [accessed 05 May 2021].

40. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E6(R3) Expert Working Group (EWG). ICH E6(R3) EWG Work Plan. ICH; 04 January 2021. Available from: https://database.ich.org/sites/default/files/E6%28R3%29_WorkPlan_2021_0104.pdf [accessed 25 June 2021].

41. Medicines and Healthcare products Regulatory Agency (MHRA). ‘GXP’ Data Integrity Guidance and Definitions. London: MHRA; March 2018. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/687246/MHRA_GxP_data_integrity_guide_March_edited_Final.pdf [accessed 10 February 2021].

42. Food and Drug Administration. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products - Draft Guidance for Industry. Food and Drug Administration; September 2021. Available from: https://www.fda.gov/media/152503/download [accessed 25 November 2021].

43. NHS Digital. Data Quality Maturity Index. NHS Digital; 27 May 2021. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/data-and-information/data-tools-and-services/data-services/data-quality/dqmi-may-2021/dqmi-methodology-v1---current.pdf [accessed 13 July 2021].

44. NHS Digital. Secondary Uses Service (SUS) Guidance. NHS Digital; 19 April 2021. Available from: https://digital.nhs.uk/services/secondary-uses-service-sus/secondary-uses-services-sus-guidance [accessed 16 June 2021].

45. NHS Digital. Payment by Results Guidance. NHS Digital; 25 May 2021. Available from: https://digital.nhs.uk/services/secondary-uses-service-sus/payment-by-results-guidance [accessed 24 June 2021].

46. NHS Digital. Data Provision Notice: Commissioning Data Sets (CDS). NHS Digital; 13 March 2020. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/corporate-information/directions-and-data-provision-notices/data-provision-notices/commissioningdatasetsdataprovisionnoticev1.0.pdf [accessed 08 December 2021].

47. Capita. Payment by Results Data Assurance Framework: Key findings from the 2012/13 programme. Kent, UK: Capita Health and Wellbeing Limited; 21 November 2013. Available from: https://www.chks.co.uk/userfiles/files/PbR_Key_Findings_Report_2013.pdf [accessed 04 May 2021].

48. Capita. The quality of clinical coding in the NHS: Payment by Results data assurance framework. Kent, UK: Capita Health and Wellbeing Limited; September 2014. Available from: https://www.chks.co.uk/userfiles/files/The_quality_of_clinical_coding_in_the_NHS.pdf [accessed 04 May 2021].

49. NHS Digital. Data Quality Assurance: Framework for Providers - Part 1. NHS Digital; January 2020. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/data-and-information/data-tools-and-services/data-services/data-quality/nhs-digital-data-quality-assurance-framework-for-providers-part-1-v1.pdf [accessed 13 July 2021].

50. NHSX. Records Management Code of Practice 2020: a guide to the management of health and care records. NHSX; October 2020. Available from: https://www.nhsx.nhs.uk/documents/48/NHSX_Records_Management_Code_of_Practice_2020_3.pdf [accessed 28 January 2021].

51. NHS Digital. Data Quality Assurance: Framework for Providers - Part 2. NHS Digital; January 2020. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/data-and-information/data-tools-and-services/data-services/data-quality/nhs-digital-data-quality-assurance-framework-for-providers-part-2-v1.pdf [accessed 13 July 2021].

52. NHS England, NHS Improvement. Commissioning for Quality and Innovation (CQUIN): Guidance for 2020 - 2021. NHS England; February 2020. Available from: https://www.england.nhs.uk/wp-content/uploads/2020/01/FINAL-CQUIN-20-21-Core-Guidance-190220.pdf [accessed 24 October 2021].

53. Care Act 2014, c.23. 14 May 2014. Available from: https://www.legislation.gov.uk/ukpga/2014/23/contents/enacted [accessed 13 July 2021].

54. NHS Digital. NHS Digital Data Quality Strategy 2018-2020. NHS Digital; 15 August 2018. Available from: https://digital.nhs.uk/binaries/content/assets/legacy/pdf/m/c/dqa-strategy-on-a-page.pdf [accessed 22 October 2021].

55. HES Data Quality Team. The HES processing cycle and HES data quality. Health and Social Care Information Centre; 26 September 2016. Available from: https://nhs-

prod.global.ssl.fastly.net/binaries/content/assets/legacy/pdf/d/5/the_hes_processing_cycle_and_data_quality.pdf [accessed 26 November 2020].

56. HES Data Quality Team. Data quality checks performed on SUS and HES data. Health and Social Care Information Centre; 26 September 2016. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/legacy/pdf/d/r/data_quality_checks_performed_on_sus_and_hes_data.pdf [accessed 26 November 2020].

57. HES Data Quality Team. Methodology for identifying and removing duplicate records from the HES dataset. Health and Social Care Information Centre; 26 September 2016. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/legacy/pdf/m/o/hes_duplicate_identification_and_removal_methodology.pdf [accessed 26 November 2020].

58. NHS Digital. Provisional Monthly Hospital Episode Statistics for Admitted Patient Care, Outpatient and Accident and Emergency data April 2020 - March 2021 (M13). NHS Digital; 15 June 2021. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-episode-statistics-for-admitted-patient-care-outpatient-and-accident-and-emergency-data/april-2020---march-2021-m13 [accessed 22 October 2021].

59. NHS Digital. The processing cycle and HES data quality. NHS Digital; 13 July 2021. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/the-processing-cycle-and-hes-data-quality [accessed 13 July 2021].

60. NHS Digital. Announcement of methodological change: Impact of changes to Hospital Episode Statistics (HES) processing from April 2021. NHS Digital; 07 December 2020. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/data-and-information/find-data-and-publications/statement-of-administrative-sources/announcement-of-methodological-change-to-hes-v1.0.pdf [accessed 17 June 2021].

61. NHS Digital. Mortality Data Review. NHS Digital; June 2020. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/coronavirus/mortality-data-review/nhs-digital-mortality-data-review.pdf [accessed 11 June 2021].

62. Coronavirus Act 2020, c.7. 25 March 2020. Available from: https://www.legislation.gov.uk/ukpga/2020/7/introduction/enacted [accessed 13 July 2021].

63. Office for National Statistics (ONS). User guide to mortality statistics. ONS; 28 September 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017/pdf [accessed 11 June 2021].

64. Office for National Statistics (ONS). Mortality statistics in England and Wales QMI: Quality and methodology information for mortality statistics in England and Wales. ONS; 01 July 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/mortalitystatisticsinenglandandwalesqmi/pdf [accessed 14 June 2021].

65. Office for National Statistics (ONS). Annex K: User guide to mortality statistics. ONS; 28 September 2020. Available from: https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017/annexk.pdf [accessed 11 June 2021].

66. Gastrell J, Griffiths C, Devis T. An analysis of legally uncertified deaths in England and Wales, 1979-2002. *Health Stat Q* 2004, https://www.ncbi.nlm.nih.gov/pubmed/15704807(24):7-13.

67. NHS Digital. Data Access Request Service (DARS) guidance: Standards of information expected in a data access application. NHS Digital; 24 May 2021. Available from: https://digital.nhs.uk/services/data-access-request-service-dars/dars-guidance#standards-of-information-expected-in-a-data-access-application [accessed 09 December 2021].

68. NHS Digital. Independent Group Advising on the Release of Data (IGARD). NHS Digital; 30 June 2021. Available from: https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/independent-group-advising-on-the-release-of-data [accessed 07 July 2021].

69. NHS Digital. Information Assurance & Cyber Security Committee Terms of Reference. Health and Social Care Information Centre; November 2019.

70. Wright-Hughes A, Graham E, Cottrell D, Farrin A. Routine hospital data - is it good enough for trials? An example using England's Hospital Episode Statistics in the SHIFT trial of Family Therapy vs. Treatment as Usual in adolescents following self-harm. *Clin Trials* 2018;15(2):197-206.

71. NHS Digital. Emergency Care Data Set (ECDS). NHS Digital; 02 March 2021. Available from: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/emergency-care-data-set-ecds [accessed 01 July 2021].

72. Thomas DS, Gentry-Maharaj A, Ryan A*, et al.* Colorectal cancer ascertainment through cancer registries, hospital episode statistics, and self-reporting compared to confirmation by clinician: A cohort study nested within the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). *Cancer Epidemiol* 2019;58:167-174.

73. Powell GA, Bonnett LJ, Smith CT, Hughes DA, Williamson PR, Marson AG. Using routinely recorded data in a UK RCT: a comparison to standard prospective data collection methods. *Trials* 2021;22(1):429.

74. Gordon B, Barrett J, Fennessy C*, et al.* Development of a data utility framework to support effective health data curation. *BMJ Health & Care Informatics* 2021;28(1):e100303.

75. Recovery Collaborative Group. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *Lancet* 2021;397(10285):1637-1645.

76. Horby PW, Mafham M, Peto L*, et al.* Casirivimab and imdevimab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *medRxiv* 2021;10.1101/2021.06.15.21258542:2021.2006.2015.21258542.

77. Group RC. Lopinavir-ritonavir in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *Lancet* 2020;10.1016/S0140-6736(20)32013-4.

78. Recovery Collaborative Group. Azithromycin in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *Lancet* 2021;397(10274):605-612.

79. Recovery Collaborative Group. Convalescent plasma in patients admitted to hospital with COVID-19 (RECOVERY): a randomised controlled, open-label, platform trial. *Lancet* 2021;397(10289):2049-2059.

80. Horby PW, Campbell M, Spata E*, et al.* Colchicine in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *medRxiv* 2021;10.1101/2021.05.18.21257267:2021.2005.2018.21257267.

81. Horby PW, Pessoa-Amorim G, Staplin N*, et al.* Aspirin in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *medRxiv* 2021;10.1101/2021.06.08.21258132:2021.2006.2008.21258132.

82. Recovery Collaborative Group. RECOVERY: Definition and Derivation of Baseline Characteristics and Outcomes v3.0. Nuffield Department of Population Health, University of Oxford; 06 January 2021. Available from: https://www.recoverytrial.net/files/recovery-outcomes-definitions-v3-0.pdf [accessed 01 July 2021].

83. Aung T, Haynes R, Barton J*, et al.* Cost-effective recruitment methods for a large randomised trial in people with diabetes: A Study of Cardiovascular Events iN Diabetes (ASCEND). *Trials* 2016;17(1):286.

84. NHS Improvement. Transition from SUS to SUS+. NHS Improvement; May 2017. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/legacy/pdf/7/t/transition_from_sus_to_sus__v1.0.pdf [accessed 13 July 2021].

85. Zhao A, Sato L, Liu L, Panesar J. HES APC lineage examples for MHRA primary source paper [Microsoft Powerpoint presentation]. NHS Digital: Integration Data Architecture team; 25 June 2021.

GLOSSARY OF TERMS

Term	Description
AHAS	Acute Hospital Analytical System, a live contemporaneous database.
ALCOA	Accurate, Legible, Contemporaneous, Original, Attributable, Complete (ICH E6(R2) section 4.9.0)
ALCOA+	Accurate, Legible, Contemporaneous, Original, Attributable, Complete, Consistent, Enduring and Available when needed
APC	Admitted Patient Care, a dataset of hospital admissions held by NHS Digital
ASCEND	A Study of Cardiovascular Events iN Diabetes (https://clinicaltrials.gov/ct2/show/NCT00135226)
Business rules	Extract specifications
CDISC	Clinical Data Interchange Standards Consortium (https://www.cdisc.org/)
CDS	Commissioning Data Sets
CRD	Civil Registration of Deaths, a dataset held by NHS Digital
CTTI	Clinical Trials Transformation Initiative
CTU	Clinical Trials Unit
DAE	Data Access Environment
DARS	Data Access Request Service (of NHS Digital)
DPS	Data Processing Service
DQ	Data Quality
DQMI	Data Quality Maturity Index, a monthly publication provided to data submitters with timely and transparent information about data quality. It can be access via the interactive reporting tool, Power BI, or as a .csv or .xlsx file from the Data Quality section of the NHSD website.
DSPT	Data Security and Protection Toolkit
EMA	European Medicines Agency
EWG	Expert Working Group
FDA	Food and Drug Administration
GCP	Good Clinical Practice (usually attributed specifically, e.g. ICH E6)
GRO	General Register Office, register of births, deaths, and marriages.
HDR UK	Health Data Research UK
Healthcare data, Healthcare systems data	Data from healthcare systems, collected for the purpose of providing clinical care. Often described as routinely collected data rather than data directly collected for research purposes (e.g., for a clinical trial) or patient generated data (such as from wearable digital technologies).
HES	Hospital Episodes Statistics
HESID	HES patient identifier
HSCIC	Health and Social Care Information Centre, now known as NHS Digital
ICH	The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICD-9, ICD-10	9th and 10th versions of International Classification of Diseases and related health problems: classification system for diagnoses and symptoms including diseases, injuries, and causes of death. Developed by the World Health Organization.

Term	Description
IG	Information Governance
IGARD	Independent Group Advising on the Release of Data
ISB	Information Standards Board for Health and Social Care
LEC	Life Events Continuity database, an ONS database with real-time feed from RON
LEDR	Life Events Data Repository, an ONS database that holds Registration of Death and other data.
MCCD	Medical Certificate of Cause of Death
MHRA	Medicines and Healthcare products Regulatory Agency
MPSID	Master Person Service identifier (HESID transitioned to MPSID)
MPS	Master Person Service
NHSD	NHS Digital
NHSX	Joint unit of NHS England and the Department of Health and Social Care
ONS	Office for National Statistics
OPCS-4	The Office of Population Censuses and Surveys Classification of Interventions and Procedures, a statistical classification of interventions and procedures undertaken in the NHS reflecting current clinical practice.
PAS	Patient Administration System, such as Lorenzo, EMIS or Cerner systems.
PbR	Payment by Results
PCMD	Primary Care Mortality Database, updated monthly from ONS. Used by public health analysts in local authorities and analysts in NHS organisations who require death data for statistical purposes.
PDS	Personal Demographics Service, holds demographic details of users of health and care services in England.
PDS NBO	PDS National Back Office, responsible for management of NHS numbers and PDS records.
PRINCIPLE	Platform Randomised trial of Interventions against COVID-19 in older people (https://clinicaltrials.gov/ct2/show/NCT02443155)
RECOVERY	Randomised Evaluation of COVID-19 Therapy trial (https://clinicaltrials.gov/ct2/show/NCT04381936)
RON	Registration Online system used to record births, marriages, deaths, stillbirths, and civil partnerships by local registrars
RWD, RWE	Real-World Data, Real-World Evidence
SUS, SUS+	Secondary Uses Service repository. SUS+ replaced SUS in April 2017. They use different methods to construct hospital spells: SUS+ uses a ‘natural’ method based on specific data fields whereas SUS uses an algorithm to deduce which episode records are combined into a spell.[84]
SUS+ APC	SUS+ version of the Admitted Patient Care dataset
TRE	Trusted Research Environment. A secure computing environment to enable storage and remote access to sensitive data for analysis, and sometimes referred to as a “data safe haven”. It usually provides analytical and statistical tools for analysis, and individual level data cannot be exported from it.
UKCTOCS	UK Collaborative Trial of Ovarian Cancer Screening (https://www.clinicaltrials.gov/show/NCT00058032)
UK GDPR	UK General Data Protection Regulation. The Data Protection Act 2018 is the UK’s implementation of GDPR.

Term	Description
XML	Extensible Mark-up Language, which is a text-based language for encoding structured information. NHSD use an XML schema based on NHS Data Dictionary definitions which consistently error checks and includes validation of formats and values.

APPENDIX 1 ALCOA+ DIMENSIONS[9] AND CDISC REQUIREMENTS FOR ESOURCE[35]

ALCOA+ Dimension	Definition(s) including CDISC eSource standard requirements (#)
Accurate	The data captured shall be accurate and the reporting of such data should be accurate (2a). The source document shall allow for accurate copies to be made (8; ICH E6[R2] 1.51). When source data are copied, the process used shall ensure that the copy is an exact copy preserving all of the data and metadata of the original (12).
Legible	Data must be held such that, when retrieved, it can be read and understood. This includes not only storing the data such that it can be retrieved, but also storing any metadata such that the meaning of the data is clear (2b). Readable at the input and output stage in a form meaningful to an independent reviewer i.e. a human being should be able to read it, not encrypted, coded or in programmed language (EMA, 2010).
Contemporaneous	Data are recorded as soon as possible after the event to which it refers (2c).
Original	The data should be the original data and not falsified (2d). The investigator shall maintain the original source document or a certified copy (5; ICH E6[R2] 2.11, 5.15.1). When source data are copied, the process used shall ensure that the copy is an exact copy preserving all of the data and metadata of the original (12). The sponsor shall not have exclusive control of a source document (10; ICH E6[R2] 8.3.13).
Attributable	Data should be attributable to the individual, both to the subject being reported on, and those who have modified that data (2e). An audit trail shall be maintained as part of the source documents for the original creation and subsequent modification of all source data (3). Source data shall only be modified with the knowledge or approval of the investigator (6).
Complete	The data must be whole, an entire set (2f).
Consistent	The data must be self-consistent and free from self-contradiction (2g).
Enduring	Source documents and data shall be protected from destruction (7; ICH E6[R2] 4.9.3, 4.9.4, chapter 8).
Available when needed	The storage of source documents shall provide for their ready retrieval (4). The location of source documents and the associated source data shall be clearly identified at all points within the capture process (11). Source documents shall be protected against unauthorized access (9)

[9] ALCOA+: accurate, legible, contemporaneous, original, attributable, complete, consistent, enduring and available when needed

- The CDISC standard developed in line with the FDA 21 CFR Part 11 regulations on electronic records specifies 12 requirements[35] (also described in the EMA reflection paper[34]):
    1. An instrument used to capture source data shall ensure that the data are captured as specified within the protocol (ICH GCP 2.6 and 6.4.9).
    2. Source data shall be Accurate, Legible, Contemporaneous, Original, Attributable, Complete and Consistent (ICH GCP 1.51, 1.52, 4.9.0, 4.9.1, 6.4.9).
    3. An audit trail shall be maintained as part of the source documents for the original creation and subsequent modification of all source data (ICH GCP 4.9.0, 4.9.3, 5.5.4).
    4. The storage of source documents shall provide for their ready retrieval (ICH GCP 2.11, 5.15.1).
    5. The investigator shall maintain the original source document or a certified copy (ICH GCP 2.11, 5.15.1).
    6. Source data shall only be modified with the knowledge or approval of the investigator (ICH GCP 4.9.3, 4.9.4, chapter 8).
    7. Source documents and data shall be protected from destruction (ICH GCP 4.9.3, 4.9.4, chapter 8).
    8. The source document shall allow for accurate copies to be made (ICH GCP 1.51).
    9. Source documents shall be protected against unauthorized access (ICH GCP 2.11, 5.15.1).
    10. The sponsor shall not have exclusive control of a source document (ICH GCP 8.3.13).
    11. The location of source documents and the associated source data shall be clearly identified at all points within the capture process (ICH GCP 6.4.9).
    12. When source data are copied, the process used shall ensure that the copy is an exact copy preserving all of the data and metadata of the original.

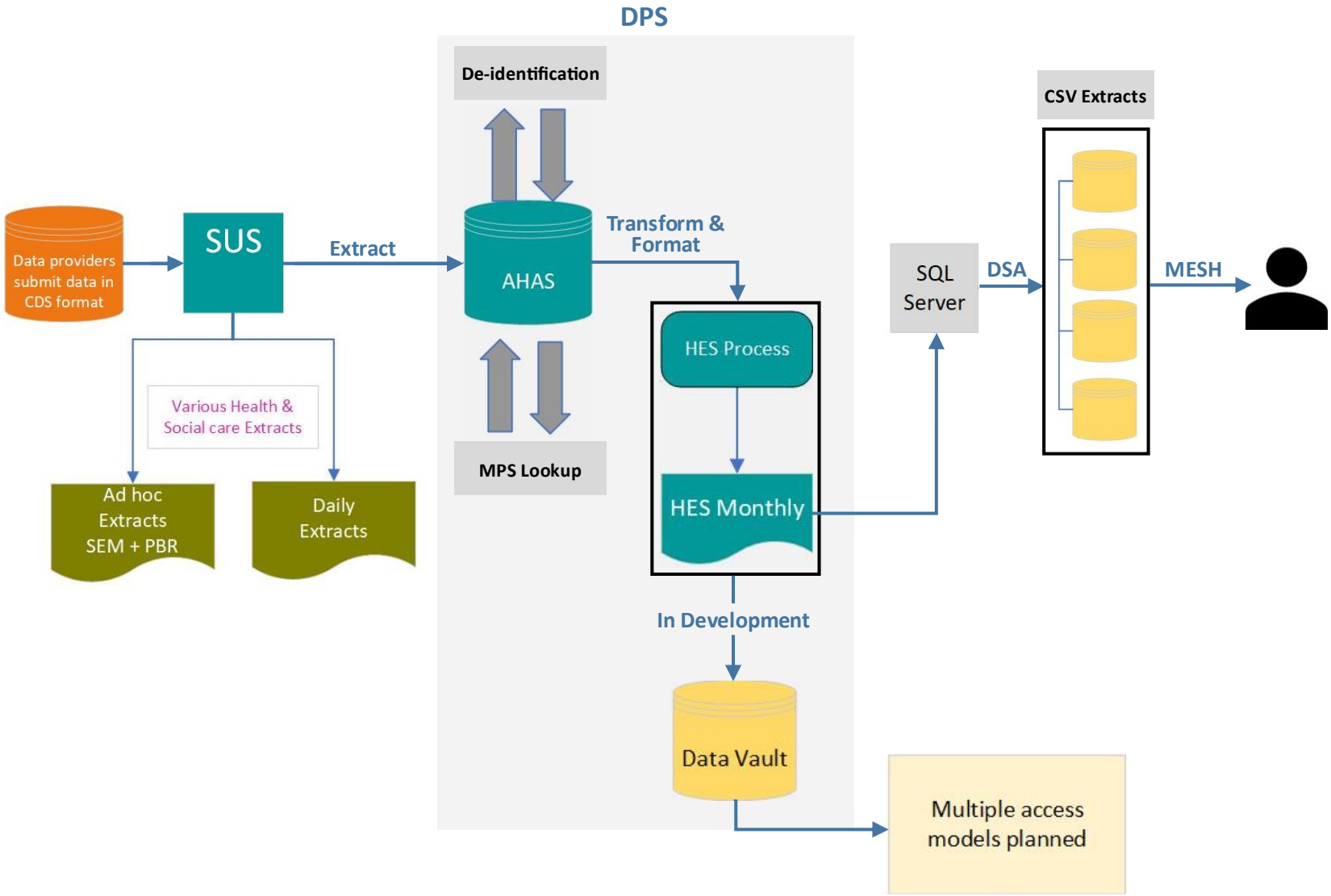
Suitability of application to healthcare data such as HES

The strict definitions of source data are difficult to apply to healthcare data such as NHSD datasets for several reasons, including:

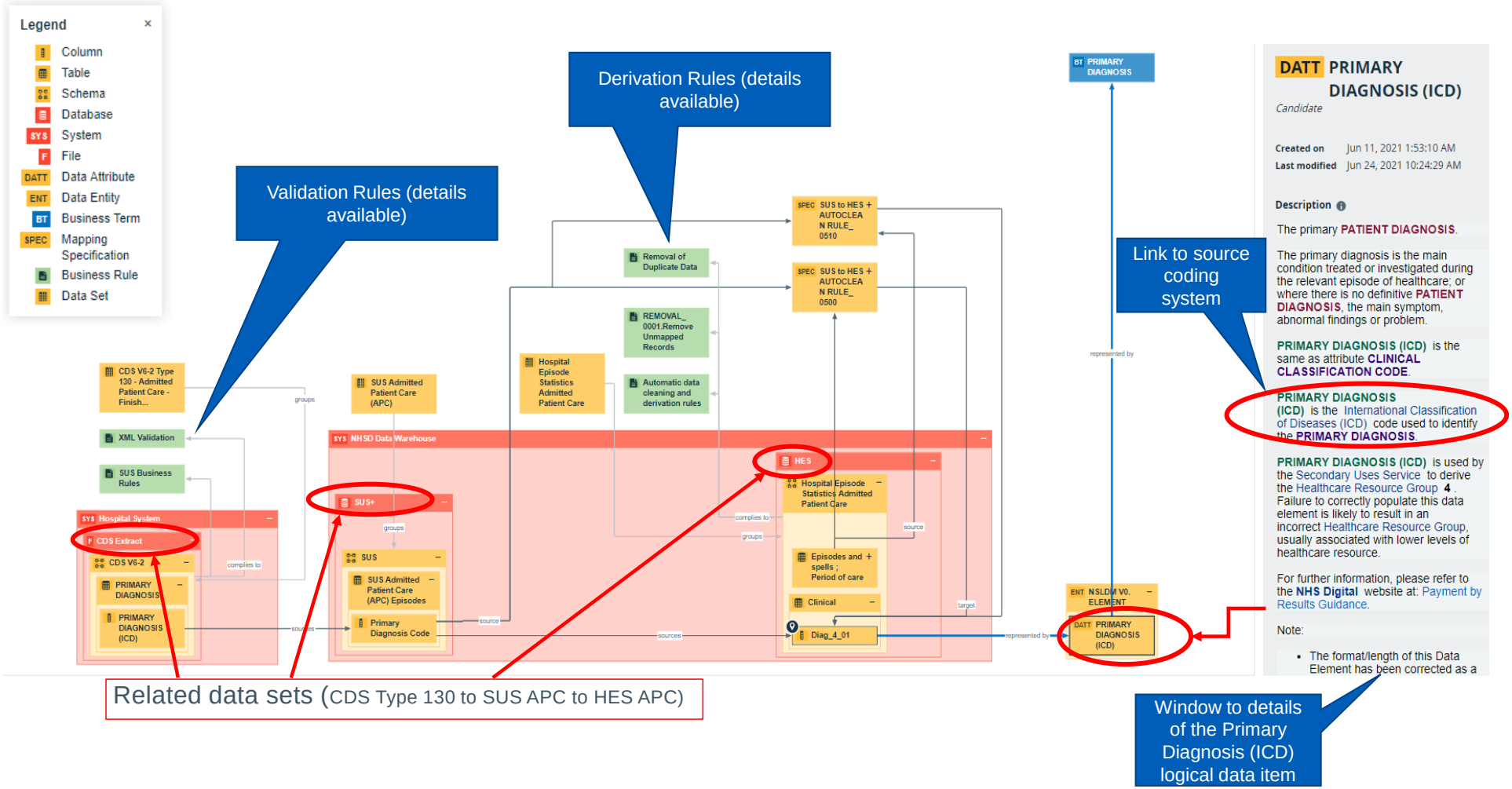
1. The Information Standard and Business rules used to abstract CDS (forming the basis of HES APC) leaves the creation/modification audit trail within the PAS.
2. Only specific coded data as defined by the Information Standard is abstracted making the consistency and completeness of the HES APC dataset difficult to demonstrate.
3. Centralised processing of HES APC by NHSD is away from the investigator's control.

4. CDS is coded as ICD-10 and OPCS, so cannot be read uncoded (although it can be translated with reference to the relevant dictionary).
5. Although CDS can be transferred from hospital trusts in a timely manner, the HES APC dataset is updated monthly and then becomes a fixed dataset on an annual basis, so it may not be suitably contemporaneous for monitoring serious safety events (as the sole source of this information).

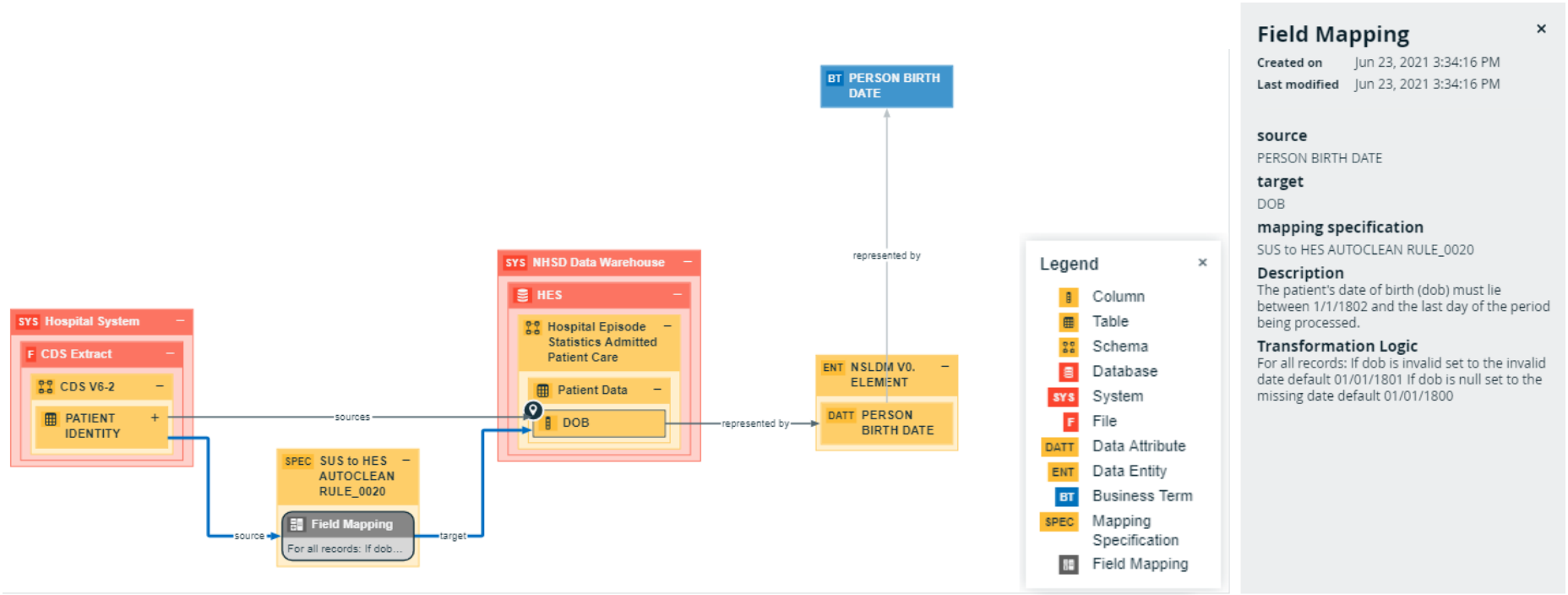
APPENDIX 2 DATA ARCHITECTURE DIAGRAM AT NHS DIGITAL



APPENDIX 3 EXAMPLE OF HES APC DATA LINEAGE FROM COLLIBRA[85]



APPENDIX 4 EXAMPLE OF HES AUTOCLEAN VALIDATION FROM COLLIBRA[85]



- Validation rules and transformation Logic: AUTOCLEAN 0020 applied to Patient.DOB in HES APC dataset
- Business/Data/Technical Lineage and Linkage for **PERSON BIRTH DATE**

APPENDIX 5 DATA FIELDS OF THE CIVIL REGISTRATION OF DEATHS DATASET

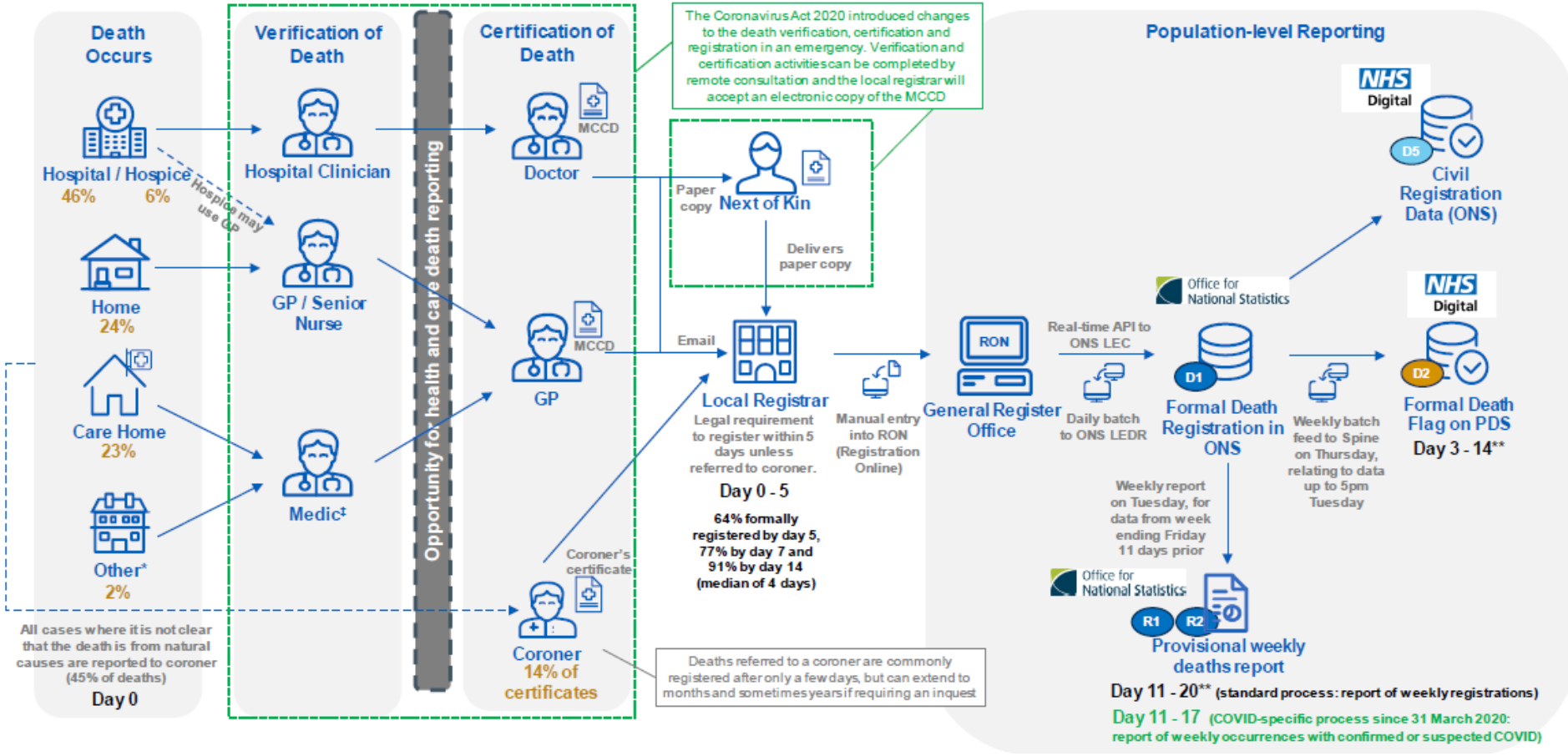
Database Field Name	Field Description	Identifiable[j]
DEC_NHS_NUMBER	NHS Number	Yes
DEC_1ST_FORENAME to DEC_4TH_FORENAME	1st to 4th Forename of deceased	Yes
DEC_EXTRA_FORENAME	Deceased's extra forenames	Yes
DEC_SURNAME	Surname of deceased	Yes
DEC_ALIASD1 to DEC_ALIASD2	Alias surnames of deceased	Yes
DEC_MAIDEN_NAME	Maiden surname	Yes
DEC_DATE_OF_BIRTH	Date of birth	Yes
DEC_SEX	Sex	Yes
DEC_MARITAL_STATUS	Marital status	No
PAR_DATE_OF_BIRTH	Spouse date of birth	Yes
PAR_RELATIONSHIP_TO_DEC	Spouse relationship to deceased	No
REG_NAMEHF	Name of spouse	Yes
DEC_AGEC	Age at death	No
DEC_AGECUNIT	Age unit code	No
REG_DATE_OF_DEATH	Date of death	No
REG_DATE	Date of Registration	No
MED_DOCTOR_NAME	Doctor/coroner certification (non-inquest)	Yes
COR_POST_MORTEM_TEXT to COR_POST_MORTEM_TEXT3	Doctor/coroner certification (non-inquest)	Yes
REG_REGISTRAR_SIGNATURE	Registrar signature	Yes
REG_REGISTRAR_DESIGNATION	Designation of registrar	No
INF_SIGNATURE	Informant's signature	Yes
INF_FORENAMES	Informant's forename	Yes
INF_SURNAME	Informant's surname	Yes
INF_QUAL_CODE	Informant's Qualification	No
INF_RELATIONSHIP_TO_DEC	Informant's relationship to deceased	No
INFAD_FLAT_NUMBER	Informant's flat number	Yes
INFAD_BUILDING_OR_HOUSE	Informant's usual Building or House number	Yes
INFAD_BUILDING_NAM	Informant's usual Building name	Yes
INFAD_LINE_1 to INFAD_LINE_4	Informant's usual Street Lines 1 to 4	Yes
INFAD_TOWN	Informant's address Town	Yes
INFAD_COUNTY	Informant's address County	Yes

[j] In accordance with UK GDPR, a natural person can be directly identified from the information.

Database Field Name	Field Description	Identifiable[j]
INFAD_POSTCODE	Informant's address postcode	Yes
REG_DISTRICT_CODE	Reg district	No
REG_SUB_DISTRICT_CODE	Reg sub-district	No
REG_NUMBER	Register number	No
REG_ENTRY_NUMBER	Entry number	No
REG_TYPE	Re-registration indicator	No
COR_DESIGNATION	Coroner designation (Inquest)	No
COR_NAME	Coroner name (Inquest)	Yes
COR_AREA_NAME	Coroner's area (Inquest)	No
COR_INQ_DATES	Date of Inquest text	No
DEC_OCCUPATION	Deceased's occupation	No
DEC_OCC_TYPE	Occupation type code	No
DEC_OCCUPATION_FREE to DEC_OCCUPATION_FREE4	Free format occupation text	No
REN_OCCUPATION	Spouse/parent occupation	No
MOT_OCCUPATION	Occupation of mother of deceased juvenile	No
OUT_SECCATDM	National Statistics Socio-economic Classification operational category for deceased or mother of deceased juvenile	No
OUT_SECCATHF	National Statistics Socio-economic Classification operational category for husband or father of deceased juvenile	No
SOC2KDM	2010 Standard Occupation Classification of deceased or mother of deceased juvenile	No
SOC2KHF	2010 Standard Occupation Classification of husband or father of deceased juvenile	No
DEC_RETIRED_IND	Retired indicator for deceased or mother	No
PAR_RETIRED_IND	Retired indicator for husband or father	No
POD_CONCAT	Place of death	No
DECAD_FLAT_NUMBER	Usual Address Flat No	Yes
DECAD_BUILDING_OR_HOUSE	Usual Address Building or House No	Yes
DECAD_BUILDING_NAME	Usual Address Building Name	Yes
DECAD_LINE_1 to DECAD_LINE_4	Usual Address Line 1 to 4	Yes
DECAD_TOWN	Usual Address Town	Yes
DECAD_COUNTY	Usual Address County	Yes
DECAD_STAT_POSTCODE	Postcode	Yes
HROR	SHA of residence [PCMD]	No
HAUTR	Primacy care org of residence [PCMD]	No
DEC_BIRTHPLACE	Place of birth	Yes

Database Field Name	Field Description	Identifiable[j]
DEC_BIRTH_COUNTRY_CODE	Country code of place of birth of deceased	No
GORR	Gov office region of residence [PCMD]	No
UTLA9R	Upper tier LA of residence [PCMD]	No
CTYR	County code of residence [PCMD]	No
CTYDR	County district code of residence [PCMD]	No
CCGR	CCG code of residence [PCMD]	No
LSOAR	Lower super output area of residence [PCMD]	No
WARDR	Ward code of residence [PCMD]	No
POD_CODE	Place of death code	No
POD_STAT_POSTCODE	Postcode of place of death	Yes
CCGPOD	CCG code place of death [PCMD]	No
UTLA9POD	Upper tier LA of place of death [PCMD]	No
HAUTPOD	Primary Care Organisation code for place of death [PCMD]	No
POD_NHS_ESTABLISHMENT	NHS Establishment indicator	No
POD_ESTABLISHMENT_TYPE	Establishment type where death occurred	No
MED_C_OF_D_LINE_1 to MED_C_OF_D_LINE_5	Cause of death lines	Yes
S_UNDERLYING_COD_ICD10	ICD10 Original Underlying code	No
S_COD_CODE_1 to S_COD_CODE_15	ICD10 Original mention codes	No
S_INJURY_EXTERNAL	Nature of injury code where the underlying cause of death (ICD10U) is an external cause	No
S_COD_LINE_1 to S_COD_LINE_15	Cause of death row positions 1 to 15	No
NEO_NATE_FLAG	Indicates whether cause of death is neonatal (0 or 1)	No
S_UNDERLYING_COD_ICD9	Cause of death underlying ICD9 coding	No
ICD9_ORIG_MENTION_1 to ICD9_ORIG_MENTION_15	ICD9 original mention codes	No
MED_C_OF_D_FREE_FORMAT	Free format cause of death	Yes
DEC_MARITAL_STATUS_TEXT	Marital status	No
DECAD_ADDR_CONCAT	Usual address concatenated	Yes
INFAD_ADDR_CONCAT	Concatenated informant address	Yes
REG_DISTRICT_NAME	Reg district name	No
REG_SUB_DISTRICT_NAME	Reg sub-district name	No
GP_PRACTICE_CODE	GP practice code [PCMD]	No

APPENDIX 6 STATUTORY REPORTING PROCESS THROUGH CIVIL REGISTRATION (REPRODUCED FROM NHS DIGITAL'S MORTALITY DATA REVIEW)[59k]



[k] API: Application Programming Interface; LEC: Life Events Continuity database; LEDR: Life Events Data Repository; MCCD: Medical Certificate of Cause of Death; PDS: Personal Demographics Service;

* Other: all other establishments/locations, e.g. military, prison, hotel, school.

‡ Medic: usually a GP or qualified medical professional, depending on circumstance and location of the death.

** Dates based on civil registration within 5 days.