Title page

Working title

Critical Care Health Informatics Collaborative (CCHIC): data, tools and methods for reproducible research: a multi-centre UK intensive care database

Author list

- Steve Harris[a,h] *
- Sinan Shi[b] *
- Spiros Denaxas[g] *
- David Brealey[a,h]
- Niall S MacCallum[a,h]
- David Perez-Suarez[b]
- Ari Ercole[c]
- Peter Watkinson[d]
- Andrew Jones[e]
- Simon Ashworth[f]
- Richard Beale[e,i]
- Duncan Young[d]
- Stephen Brett[f]
- Mervyn Singer[a]

*Joint first authorship (based on contribution to research and manuscript preparation)

Affiliations

a. Bloomsbury Institute of Intensive Care Medicine, University College Hospital, London, UK

b. Research Software Engineering, University College London, London, United Kingdom

c. Division of Anaesthesia, Department of Medicine, Cambridge University, UK

d. Critical Care Research Group (Kadoorie Centre), Nuffield Department of Clinical Neurosciences, Medical Sciences Division, Oxford University

e. Critical Care, Guy's and St. Thomas' NHS Foundation Trust, London, UK

f. Critical Care, St. Mary's Hospital, Imperial College Healthcare NHS Trust, London, UK 8 Critical Care, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK

g. Institute of Health Informatics, University College London, Gower Street, London, WC1E 6BT, United Kingdom

h. Critical Care, University College London Hospitals NHS Foundation Trust, London, UK

i. Division of Asthma, Allergy and Lung Biology, King's College, London, UK

Acknowledgements and support

This research was funded by the National Institute for Health Research Health Informatics Collaborative and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

Abstract

Objective

To build and curate a linkable multi-centre database of high resolution longitudinal electronic health records (EHR) from adult Intensive Care Units.

To develop a set of open-source tools to make these data ‘research ready’ while protecting patient’s privacy with a particular focus on anonymization.

Materials and Methods

We developed a scalable EHR processing pipeline for extracting, linking, normalising and curating and anonymising EHR data. Patient and public involvement was sought from the outset, and approval to hold these data was granted by the NHS Health Research Authority’s Confidentiality Advisory Group (CAG). The data are held in a certified Data Safe Haven. We followed sustainable software development principles throughout, and defined and populated a common data model that links to other clinical areas.

Results

From January 2014 to January 2017, this amounted to 21,930 and admissions (18,074 unique patients). Typical admissions have 70 data-items pertaining to admission and discharge, and a median of 1030 (IQR 481–2335) time-varying measures. Training datasets were made available through virtual machine images emulating the data processing environment. An open source R package, cleanEHR, was developed and released that transforms the data into a square table readily analysable by most statistical packages. A simple language agnostic configuration file will allow the user to select and clean variables, and impute missing data. An audit trail makes clear the provenance of the data at all times.

Discussion

Discussion: Making health care data available for research is problematic. CCHIC is a unique multi-centre longitudinal and linkable resource that prioritises patient privacy through the highest standards of data security, but also provides tools to clean, organise, and anonymise the data. We believe the development of such tools are essential if we are to meet the twin requirements of respecting patient privacy and working for patient benefit

Conclusion

The CCHIC database is now in use by health care researchers from academia and industry. The ‘research ready' suite of data preparation tools have facilitated access, and linkage to national databases of secondary care is underway.

Keywords

Electronic health records; database; clinical decision support; critical care; reproducibility

Introduction

Empirical observation, or measurement, was the foundation of the Scientific Revolution, but was historically expensive [1]. Digitalisation and the computer age have changed this, and the electronic health record (EHR) is health care's version of 'big data'. Critical care will inevitably be at the forefront of the big data revolution because there is no other environment where patients are monitored more closely, or with such a broad range of measures.

However, making such data available for research is problematic for three reasons. Firstly, health data is sensitive, and the protection of patient privacy must trump all other issues. Secondly, such data is frequently unusable in its raw format. The pace of research must not be mired by the need to repeatedly prepare and clean the data. Thirdly, the data should not exist in isolation. A critical care admission is just one part of an illness pathway. There are antecedents and consequences, and those consequences will impact the patient, their family, and the health service.

Underlying these issues, there is also the thornier problem of data ownership. If the default position is that organisations are temporary guardians of personal data, then there is an expectation that the data should be used in the best interests of patients.

In response to this we have developed the Critical Care Health Informatics Collaborative (CCHIC), a partnership between the UK's National Institute of Health Research (NIHR) and five leading NHS hospital trusts. CCHIC attempts to deliver critical care 'big data' to researchers thereby facilitating research for patient benefit. Demographics, diagnostic, physiological and treatment data are abstracted from critical care admission to discharge creating a high-resolution, longitudinal EHR of unprecedented depth and breadth.

Uniquely, the resource is designed to be explicitly linkable. This means that other clinical specialties can understand the disease process in their most vulnerable and unwell patients. It means that we can begin to share with patients and families a true picture of survivorship following critical care. We can report on long term outcomes, subsequent disease profiles, and use of health resources. We can in theory understand whether people return to work, and the impact of the illness on the wider family.

CCHIC has a specific focus on open-access, reproducible research that is done with patient and public involvement from the outset. Making the data research ready yet robustly anonymised for as wide a community of academic and clinical collaborators as possible fulfils our ethical responsibility to the patients who provide these data. In this paper we describe the database, the pipeline (extracting, cleaning, curating, and distributing), and the tools built to enable reproducible research.

Objectives

The objectives of our research were threefold:

1. To build and curate a linkable multi-centre database of high-resolution, longitudinal and multi-modal EHR data from adult Intensive Care units (ICU)

2. To create a scalable pipeline ('Extract Transform Load', ETL) for extracting, linking, cleaning, encoding and anonymising ICU data across multiple secondary healthcare providers

3. To develop a set of open source tools and methods for undertaking reproducible research using the database

Materials and methods

In 2014, CCHIC started to recruit consecutive admissions to the general adult medical and surgical critical care units at the five founding National Institute of Health Research (NIHR) BRCs at Cambridge, Guy's, Kings' and St Thomas', Imperial, Oxford and University College London (UCL). The current dataset (version 1.0) includes 264 fields comprising 108 hospital, unit, patient and episode descriptors (recorded once per admission), and 154 time-varying physiology and therapeutic fields (recorded hourly, daily etc.). Data are currently exported on a quarterly basis with the ambition to move to near realtime collection.

Table 1: Participating hospitals and critical care units (ICU: Intensive Care Unit, HDU: High Dependency Unit, OIR: Overnight Intensive Recovery).

Biomedical Research Centre	Hospital	Unit
Cambridge	Addenbrooke's Hospital	ICU/HDU
Cambridge	Addenbrooke's Hospital	Neuro
GSTT	Guy's Hospital	ICU
GSTT	St Thomas' Hospital	ICU/HDU
GSTT	St Thomas' Hospital	OIR
GSTT	St Thomas' Hosptial	HDU
Imperial	Hammersmith Hospital	ICU/HDU
Imperial	St Mary's Hospital London	ICU
Oxford	John Radcliffe	ICU
UCLH	University College Hospital	ICU/HDU
UCLH	Westmoreland Street	ICU/HDU

Regulatory Approval

To be of benefit to researchers the database must allow access to data that is reflective of the entire critical care cohort for their full critical illness. A direct consent model would face two challenges: 1. practicability of consenting more than 10 000 patients per year initially and more as the project expands 2. temporary or permanent lack of capacity to consent amongst the critically ill either due to the severity of the illness, the use sedation during mechanical ventilation or a high (circa 15%) early mortality rate

The project therefore approached the Confidentiality Advisory Group (CAG) who provided a legal basis for data sharing for essential medical research, and granted an exemption to the common law duty of confidentiality for the project under Section 251 of the NHS Act 2006 (14/CAG/1001). A favourable opinion was provided by the National Research Ethics Service (14/LO/103). Data sharing agreements were signed between the participating NHS Trusts and UCL which hosts the Data Safe Haven (DSH) where the data are stored. The DSH is certified to the ISO/IEC 27001:2013 information security standard and conforms to the NHS Information Governance Toolkit level 2.

All patients are provided with information regarding the project and an option by which to opt out and a Standard Operating Procedure (SOP) exists to ensure this wish is carried out, and

public and patient involvement is actively sought through notifications at each participating unit, and other media.

CCHIC Design principles

The design of CCHIC has been based on the following principles:

1. to protect the privacy of the patients

2. to support research for patient benefit (specifically excluding commercial exploitation)

3. to facilitate that research by building a scalable pipeline for extracting, processing, and sharing the data

Principle 1: patient privacy

Being able to protect patient's privacy with confidence is the first and foremost consideration for this data resource. Extensive patient and public engagement work has been performed to ensure that this resource is seen as a public good by a broad cross-section of constituents. The particular problem with critical care research is that the patients themselves are either temporarily or permanently incapacitated and therefore unable to offer explicit permission. In the UK, this triggers the need for an application to the Secretary of State for Health to hold these data without consent (as per Section 251 of the NHS Act 2006). Permission is only granted when the physical security of the data can be guaranteed, and when the justification for holding the data is in the public interest (hence principle 2).

The data itself is encrypted before leaving each hospital, and then moved to the data safe haven at University College London. Access to the identifiable data is strictly controlled, but an anonymisation step in the data pipeline makes an extract of the data ready for the end-researcher (principle 3).

Principle 2: research for patient benefit

Even after privacy is protected, there is a widely reported distinction in the public perception of rights to use data. Recent furore over the partnership between the Royal Free NHS Foundation Trust and Google DeepMind in 2016 was driven by suspicion of the motives of commercial organisations especially those with the pervasive reach of Google [3].

In the DeepMind case, the purported use of the data was to simply develop an alerting system for patients with acute kidney injury. However calculating the AKI class from a laboratory creatinine is so simple that it is hard to believe this was Google's end game. In fact the Information Sharing Agreement that was signed in 2015 placed no restrictions on the data to be analysed, or the technologies that might be used [4].

For CCHIC, in contrast, the data cannot be used for profit, the research question must be explicitly for patient benefit, and even anonymised data releases must be proportional to the researcher's need.

Principle 3: research ready

Principle (1) protects the patient, and Principle (2) justifies the risks, however small, of making health care data available. Principle (3) enables the researcher to deliver on the promise of their research. Most data analysis requires a huge amount of preparation. We therefore developed an automated data processing pipeline to process, curate, and make available the data (Fig.1).

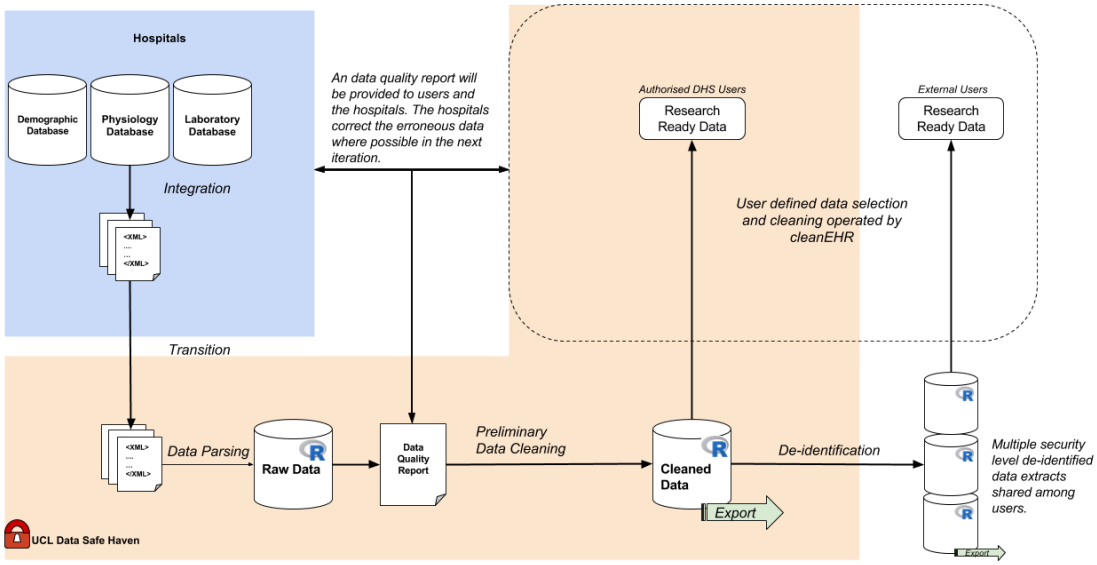
Data specification We developed an XML-based format for individual ICUs to store and transmit the extracted EHR data. The common data model was developed in collaboration with clinicians, clinical information systems architects and researchers. A description of the XML data model used is provided via the NIHR's Health Data Finder [5].[4]

We extract EHR data from each ICU using a combination of manual, semi-automatic or entirely automatic methods adapted to local ICU clinical information systems. Currently, this includes systems from Phillips Healthcare and Epic Systems, but there is no barrier to extraction from other EHR providers. Data items are extracted as frequently as they were reported (typically hourly) from ICU admission to discharge.[5]

This includes bedside physiology, near patient testing, laboratory testing, and drug administration. In addition, diagnostic coding, patient co-morbidities, admission and discharge pathways, demographics and other information typically used for risk adjustment are extracted on a per admission basis.

Uniquely, patient identifiers (NHS number, name, and date of birth) are retained with the record to enable linkage to other health and social care resources. This includes but is not limited to data curated by NHS digital (e.g. Hospital Episode Statistics, and mortality data from the Office of National Statistics), primary care, and clinical trial data sets.

below: (a) data extraction, (b) quality assessment and (c) anonymisation.



Data quality

Our approach to data quality was based on the philosophy of reproducing accurately the local EHR rather than curating a data set for audit, benchmarking or quality control. For example, aberrant invasive blood pressure readings of 300mmHg occur when the transducer system is flushed, and exposed to the attached pressure bag instead of the patient. For benchmarking, it is important to identify and exclude these values before using them to adjust for patient outcomes. However, it is exactly this sort of artefact that must be handled by the designer of a clinical monitoring system. Such use cases are very much part of the justification for CCHIC. Similarly, missing data might need imputation to allow between patient comparison for audit, but the pattern of missingness has turned out to be an important input for computer scientists trying to predict patient outcomes.
Data extracts therefore had only to contain a minimum set of fields (critical care unit, episode identifier and data item timestamps). With these data, an audit trail could be constructed and the quality could be reported. Modifications to data items were only requested where reporting did not match the schema standards (e.g. reporting PaO_2 in mmHg rather than kPa), or where entire fields were missing because of a problem with local exporting. A data quality report summarised the completeness of each time-invariant field, and the sampling frequency

of the time-varying fields. It would also summarise the characteristics of each field, and compare data extracts between submitting institutions, and with previous extracts from the same institution. Again, the purpose of the report was to identify when local extraction procedures were not accurately capturing the local EHR.

Data anonymisation

Patient confidentiality can be protected either by security measures (physical security, access control, and appropriate governance), or by anonymisation. We argue that relying solely on the former will impede our 'research ready' philosophy partly because of the necessary administrative overhead, but also because the data *storage* environment also becomes the *development* environment. The pace of change of modern machine learning, statistical, and software tools would mean that the development environment needs continuous updating. This is a burden, and a security risk as each update requires an external ingest of code. As the number of researchers grows then so will the number of tools, and risk of external exposure. Making available a proportionate anonymised extract of the data both solves this problem, and allows researchers to develop work with their own tools. However, the importance of getting the anonymisation process right cannot be understated. The process must balance the following demands: 1. the likelihood of identification being successful 2. minimising incentives for re-identification 3. the quality of the data post-anonymization. Here we have followed the guidance1 provided by the Information Commissioner's Office (ICO) which states that > (there is) clear legal authority for the view that where an organisation converts personal data into an anonymised form and discloses it, this will not amount to a disclosure of personal data.

Likelihood of identification being successful

There is a trade off between information loss and disclosure risk so that as the risk of disclosure decreases then so does utility of the data. To define this we need to measure the information content, and quantify the disclosure risk. Ignoring *direct identifiers* (e.g. NHS numbers which have a uniquely identify an individual) which must be removed, then it is still possible to re-identify individuals by the intersection of *key variables*. K-anonymity counts the number of individuals identified by the intersection of these*key variables*. We would wish this to be above so lower limit (say 10 or more).2 We have therefore implemented an anonymization algorithm which iteratively perturbs and aggregates these key variables until a convergences criterion based on the k-anonymity metric is reached.

Minimising incentives for re-identification

---

1 see Information Commissioner's Office. *Anonymisation: managing data protection risk code of practice*. 2014. The legal basis for this guidance comes from the Data Protection Act (DPA) 1988, and Recital 26 of the European Data Protection Directive (95/46/EC) which in turn is based on the following principles: (1) Personal data has to be about a living person, meaning that the DPA does not apply to mortality or other records about the deceased; (2) information or a combination of information, that does not relate to and identify an individual, is not personal data

2 For example, if we release individual data describing 'species', and 'favourite sandwich filling', then the intersection of 'bears' and 'marmalade' would uniquely identify Paddington Bear. If we generalise 'favourite sandwich filling' to 'prefers sweet sandwiches' then because Pooh Bear likes honey as well as Paddington liking marmalade, the k-anonymity would rise to two.

While a cliche, there is anonymity in obscurity. For this reason, records of publicly prominent individuals3 are removed prior to a data release (just as inviduals opt-outs are removed prior to data storage). However, because of the sensitivity of medical data, this risk remains to others. We therefore prospectively identify *sensitive data items* such as those recording cirrhosis, or HIV status. These are only released where sufficient heterogeneity is present such that the status of a single patient cannot be determined.4

Quality of the data post-anonymization

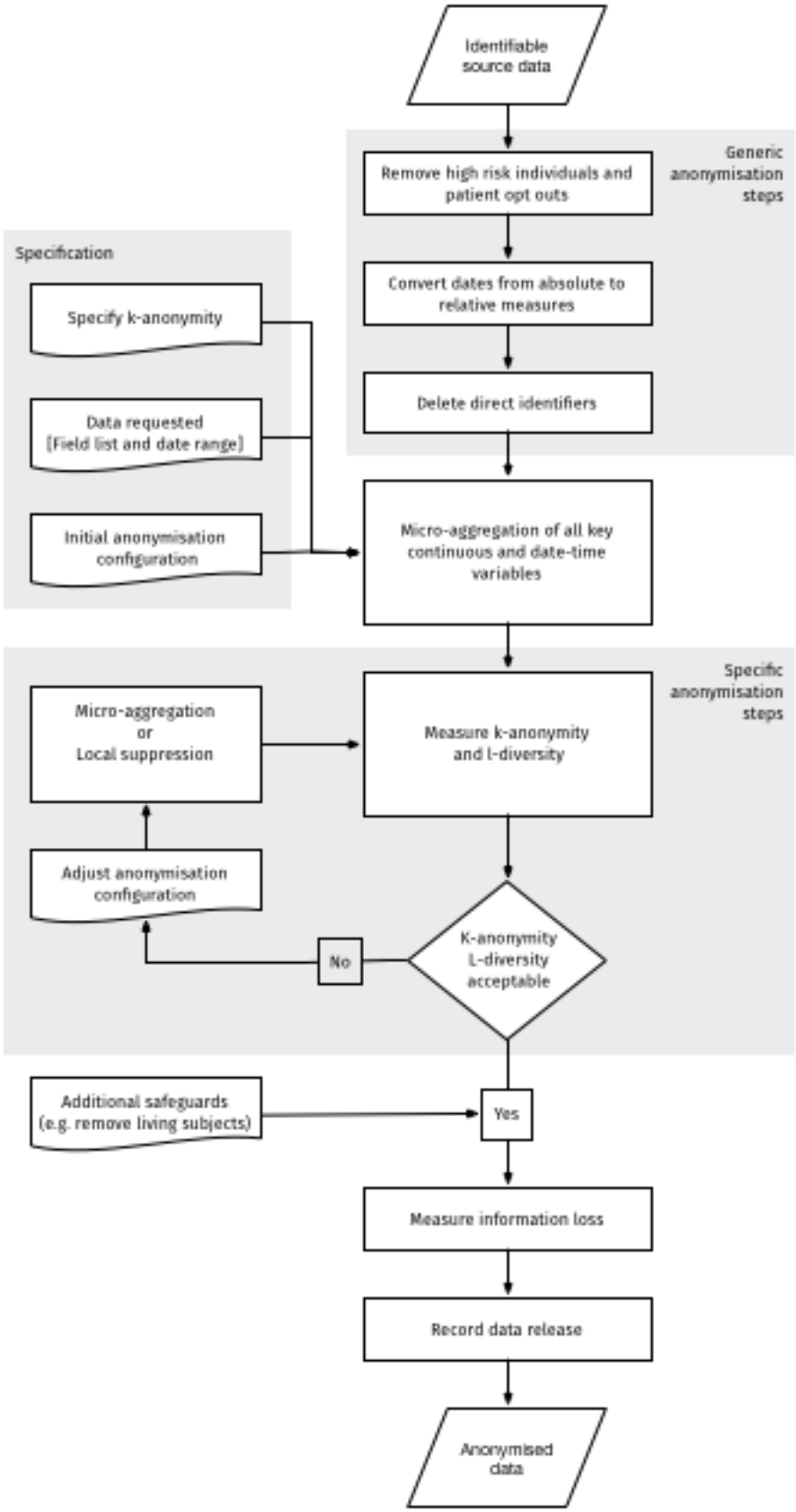
For non-identifying variables, (e.g. heart rate), there is no information loss. For key variables and sensitive fields, a balance must be reached. For example, a project examining the weekend effect on critical care outcomes might have to sacrifice granularity in other key variables (e.g. age) in order to extract the data. Such a compromise is not normally an impediment. Where information loss is not acceptable, then the research team will have to go through a vetting process to work with the original data, and be prepared to work with the more limited set of tools available in the IDHS.

Data anonymization: algorithm implementation

A summary of the anonymisation process applied before any data release. 1. Removal of direct identifiers: All unique identifiers including NHS number and hospital number will be removed from the data before release. 2. Remove high risk individuals and specific opt-outs 3. Date and time metadata: All timestamps are converted to data and time differences from the instant of critical care admission. 4. Aggregate continuous and date-time key variables: Because we cannot group patients by a continuous measure, the concept of k-anonymity only applies to categorical variables (e.g. we can group patients by gender, but not by hair length). Where a key variable is continuous then we will run an initial conversion to a categorical version by aggregating. The unit of aggregation will be the natural unit of the measurement (e.g. years for age, or kilograms for weight), and the initial aggregation will be some multiple of that unit (e.g. 2 years, and 5 kg respectively). This multiples will be initially small in order to minimise information loss, but will be increased during the iterative specific anonymisation step until the necessary k-anonymity and l-diversity is reached. 5. Remove living subjects (where possible): The Data Protection Act only applies to living individuals so where possible data will only be released for non-survivors.

3 The team managing the MIMIC-III database at MIT report that there were several attempts at identifying the victims of the Boston marathon bombing in 2013. Although their database is open source, they have removed this individuals from the publicly released version.

4 This is known as *l-diversity* and guarantees that even if an individual can be identified as belonging to a small group (cell) there is sufficient variability of these sensitive items within that group that uncertainty remains as to an specific individual's status.



In practice, we used a heuristic algorithm within the *sdcMicro* R package developed by the International Household Survey Network5 to suppress (remove) quasi-identifiers from the dataset until the target k-anonymity is reached. Quasi-identifiers are aggregated to increase the granularity

5 http://www.ihsn.org/software/disclosure-control-toolbox

before the k-anonymity suppression. Additionally, for the public release, the remaining quasi-identifiers are perturbed with noise.

Data anonymization: tiered data access model

The algorithm above provides a mechanistic level of security that is supplemented by additional administrative safe guards. For example, in contrast to a member of the general public, a medically qualified researcher is expected to follow a code of professional ethics with associated sanctions for breach of this code. Releases to the general public are more strictly anonymised than releases to medical researchers. We have two standard tiers of data release based on the likelihood of re-identification being attempted: general public, or quasi-public. The general public extract is a small subset of the original dataset, where direct identifiers are removed, and quasi-identifiable variables are heavily aggregated and perturbed. It thus has the lowest disclosure risk but also the lowest data usability. Although the physiology fields are unaltered, the analysis results cannot be directly used for publication. The purpose of this dataset is for users to familiarise themselves with the data structure and to develop hypotheses that could be tested on the full data. To gain access to this dataset, researchers still must sign data sharing agreement, identifying themselves and their institution, confirming that they will be only be using the data for clinical research (in line with our research ethics permissions), and undertaking to be respectful of the data (specifically not to pass it on, nor to attempt to re-identify individuals). A quasi-public data extract is distributed to researchers who have submitted a data request that has been vetted by the CCHIC governance structure. Researchers are recommended to request the minimum set of fields necessary for their planned analysis. Where this balance cannot be achieved with a public release, then the analysis may initially proceed using the anonymised data. The analysis script is then tested on a virtual machine that simulates the development inside the data safe haven. Finally, the tested script is deployed within the safe haven, and the outputs are released to the investigator after inspection to ensure that these too pose no re-identification risk.

Research ready: the cleanEHR toolkit

As described above, the data that is released is a 'warts and all' version of the electronic health care record integrated across the sites. Although being faithful to the original record is a design principle, it leaves researchers without a specific focus on artefacts, missingness and errors with the huge task of cleaning the data. We therefore provide alongside the data an open source R package *cleanEHR* that presents a standardised set of methods for transforming the data into research-ready datasets for statistical analysis. This includes the most common data pre-processing and post-processing operations. The most important of these is a function that converts the various asynchronous lists of time-dependent measurements into a table of measurements with a customisable cadence. For example, if the researcher wishes to analyse? the data every hour then a skeleton table is built with one row per critical care admission per hour from the time of admission to the time of discharge. For time-invariant data, the data items are repeated across all rows. For time-varying items, a value is inserted if a value has been recorded in that hour.6 The end result is a data frame that is ready for analysis in applications from Microsoft Excel to SPSS, from R to Python. Additional functionality includes the ability to relabel the data fields at will, to perform range and consistency checks, and to either impute missing values or to remove episodes with excess missingness. All of this is performed by providing a simple text file7 with the configuration requests so that even users not familiar with the R programming language can configure the data processing and cleaning pipeline to match their requirements. The entire package is provided with tutorials and documentation.

Database characterization

We characterize the CCHIC database in terms of data fields, clinical data, cause of admission – would need to be aligned with other papers out there from clinical perspective

This requires applying to become a collaborator of CCHIC, and receiving training to work with the UCL Identifiable Data Handling Solution.

6 Where more than one item is available in that time period, the most recent measurement is used by default although other selection algorithms are possible.

7 The text file is specified using the human readable and writeable version of XML called YAML. Learning the formatting rules for this should take no more than ten minutes.