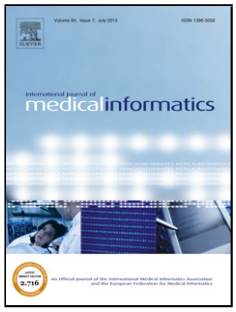
Accepted Manuscript



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

Authors: Steve Harris, Sinan Shib, David Brealey, Niall S. MacCallum, Spiros Denaxas, David Perez-Suarez, Ari Ercole, Peter Watkinson, Andrew Jones, Simon Ashworth, Richard Beale, Duncan Young, Stephen Brett, Mervyn Singer

PII: S1386-5056(18)30007-8
DOI: https://doi.org/10.1016/j.ijmedinf.2018.01.006
Reference: IJB 3637

To appear in: *International Journal of Medical Informatics*

Received date: 2-10-2017
Revised date: 6-12-2017
Accepted date: 8-1-2018

Please cite this article as: Steve Harris, Sinan Shib, David Brealey, Niall S.MacCallum, Spiros Denaxas, David Perez-Suarez, Ari Ercole, Peter Watkinson, Andrew Jones, Simon Ashworth, Richard Beale, Duncan Young, Stephen Brett, Mervyn Singer, Critical Care Health Informatics Collaborative (CCHIC): data, tools and methods for reproducible research: a multi-centre UK intensive care database, International Journal of Medical Informatics https://doi.org/10.1016/j.ijmedinf.2018.01.006

1 Title page

1.1 Working title

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

1.2 Author list

- Steve Harris[a,h] *
- Sinan Shi[b] *
- David Brealey[a,h]
- Niall S MacCallum[a,h]
- Spiros Denaxas[g]
- 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)

1.3 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

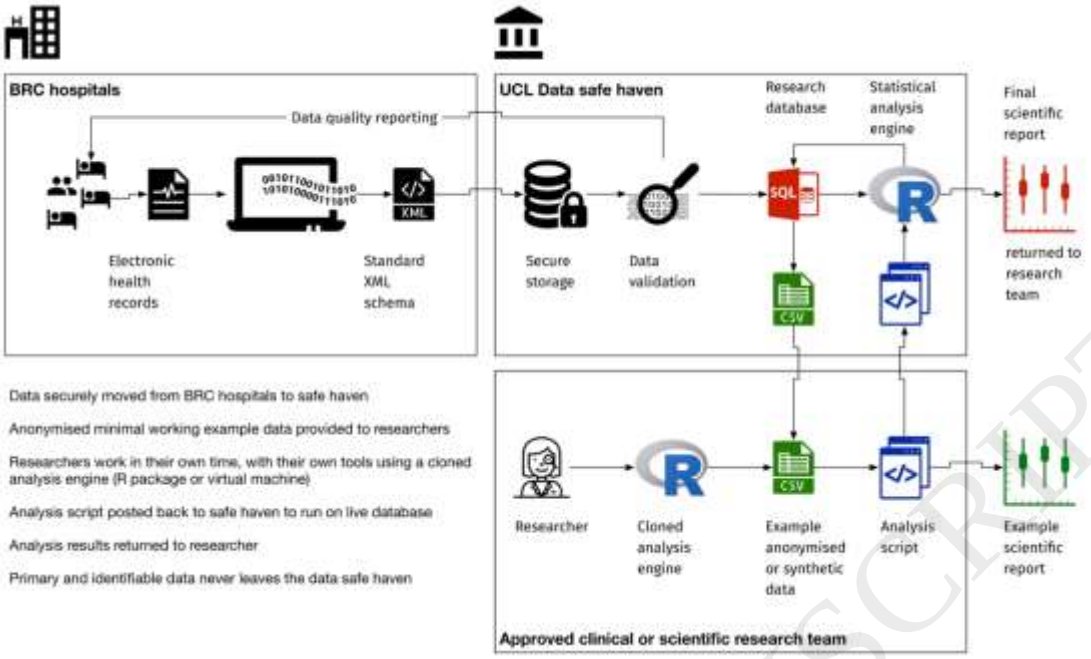
f. Critical Care, St. Mary's Hospital, Imperial College Healthcare NHS Trust, London, UK 8 Critical Care, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK

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

Graphical abstract



 Version 2018-01-10t13:23:28

2 Structured abstract

Objective

1. To build and curate a *linkable* multi-centre database of high resolution longitudinal electronic health records (EHR) from adult Intensive Care Units (ICU)
2. To develop a set of open-source tools to make these data 'research ready' while protecting patient's privacy with a particular focus on anonymisation

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

Longitudinal EHR data were loaded into the CCHIC database from eleven adult ICUs at 5 UK teaching hospitals. 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 to 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

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

3 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.

3.1 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

4 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.

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

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

4.1 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. The practicability of consenting thousands of patients per year, and, more importantly, the lack of capacity to consent for many critically ill patients. This is either due to the severity of the illness, the use of sedation during mechanical ventilation, or a high (circa 15%) early mortality rate. A consent based model would under-represent the most unwell patients.

The project therefore approached the NHS Health Research Authority's 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 Digital's Information Governance Toolkit. [2]

All patients are provided with information regarding the project and an option by which to opt out. Public and patient involvement is actively sought through notifications at each participating unit, and other media.†

* The data set is available via the http://www.hdf.nihr.ac.uk/catalogue/#/catalogue/dataModel/13

† Videos explaining the programme are available on the internet (https://www.youtube.com/watch?v=NjE9VQo-nP4&t=11s, and https://www.youtube.com/watch?v=aQJmV6i58H4)

4.2 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

4.2.1 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).

4.2.2 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.

4.2.3 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.

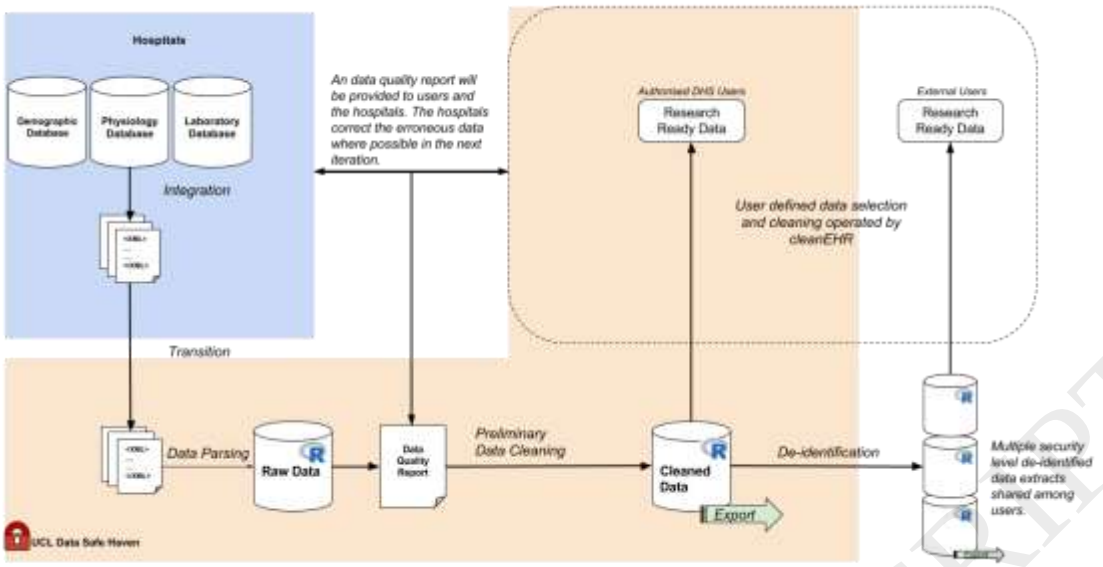


Figure 1: Data processing pipeline: Data moves from the hospital EHR to the UCL data safe haven as an XML file, and is validated before appending to the central database. A data quality report is then returned to the submitting site. Preliminary cleaning removes out of range and invalid entries. The database can then be queried in its identifiable form by authorised users within the safe haven, or a separate anonymiser can produce extracts for external collaborators.

4.2.3.1 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]*

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.† 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.

4.2.4 Data quality

Our approach to data quality is based on the philosophy of reproducing accurately the local EHR rather than curating data 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

* Of note, this XML schema is common to other clinical schemes under the umbrella Health Informatics Collaborative programme including acute coronary syndromes, ovarian cancer, renal transplant anf viral hepatitis.

† Some data items such as waveform data are often recorded at microsecond intervals, but are only reported to the local EHR solution at hourly or similar intervals.

a clinical monitoring system. Such use cases are very much part of the justification for CCHIC. Similarly, some projects will automatically impute missing data or discard incomplete records whereas others use the pattern of missingness for clinical diagnostics. [6]
Hence data extracts were accepted if the provenance (submitting unit, file name and timestamp) and the indexing information (critical care unit, episode identifier, data item label and timestamp) were complete. A data quality report summarised the completeness of each time-invariant field, and the sampling frequency of the time-varying fields.
Field level characteristics of new data ingests were compared to existing data within and across institutions in order to identify failure of local extraction procedures to accurately capture the local EHR. Fresh extracts were requested where reporting did not meet 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.

4.2.5 Data anonymisation

Researchers may apply to work with the primary identifiable data where necessary. However, limiting this access is clearly desirable with respect to data security. Moreover, working directly within the data safe haven (DSH) means 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. Each update requires an external ingest of code, and as the number of researchers grows then so will the number of tools, and the risk of external exposure.
We therefore minimise this risk by undertaking to make available anonymised data extracts to approved researchers. Here we follow guidance from the Information Commissioner's Office (ICO) [7] which is in turn based on the UK Data Protection Act (DPA) 1988 and Recital 26 of the European Data Protection Directive (95/46/EC)* The key principle is that "information or a combination of information, that does not relate to and identify an individual, is not personal data". [7] Moreover,

(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.

The anonymisation focussed on three areas:

1. Minimising the likelihood of re-identification
2. Minimising incentives for re-identification
3. Maximising the quality of data post-anonymisation

4.2.5.1 Minimising the likelihood of re-identification

We first delete all *direct identifiers* (e.g. NHS numbers which have a uniquely identify an individual). However, other *key variables* can be combined by a motivated intruder, particularly one with access to external data sources, to re-identify individuals by the intersection of specific rare values.
K-anonymity counts the number of individuals identified at this intersection, and we set k so that this smallest group still provides anonymity for its members.† In practice, we use a heuristic algorithm within the *sdcMicro* R package [8] developed by the International Household Survey Network to suppress quasi-identifiers from the dataset until the target k-

* On 25 May 2018, this will be superseded by the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679).

† 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.

anonymity is reached.[9] Quasi-identifiers are aggregated to increase the granularity before the k-anonymity suppression.* Additionally, for the public release, the remaining quasi-identifiers are perturbed with noise.

4.2.5.2 Minimising incentives for re-identification

While a cliche, there is anonymity in obscurity. For this reason, records of publicly prominent individuals† are removed prior to a data release (just as invidual opt-outs are removed prior to data storage). However, because of the sensitivity of medical data, this risk remains to others. In addition, we prospectively identify *sensitive data items* such as those recording (alcoholic) cirrhosis, or HIV status. These are either suppressed if homogeneous, or released if heterogenous. In this way the disease status of the members of even the smallest (k) group remains uncertain.‡

4.2.5.3 Maximising the quality of data post-anonymisation

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.

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

* For example, if the two most elderly patients were 101 and 109 years old, there is a risk of re-identification. These extreme values might be replaced (perhaps with the local median of 105 years). K-anonymity could then be (re)evaluated, and is likely to increase.

† 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.

‡ 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.

the more limited set of tools available in the Data Safe Haven.

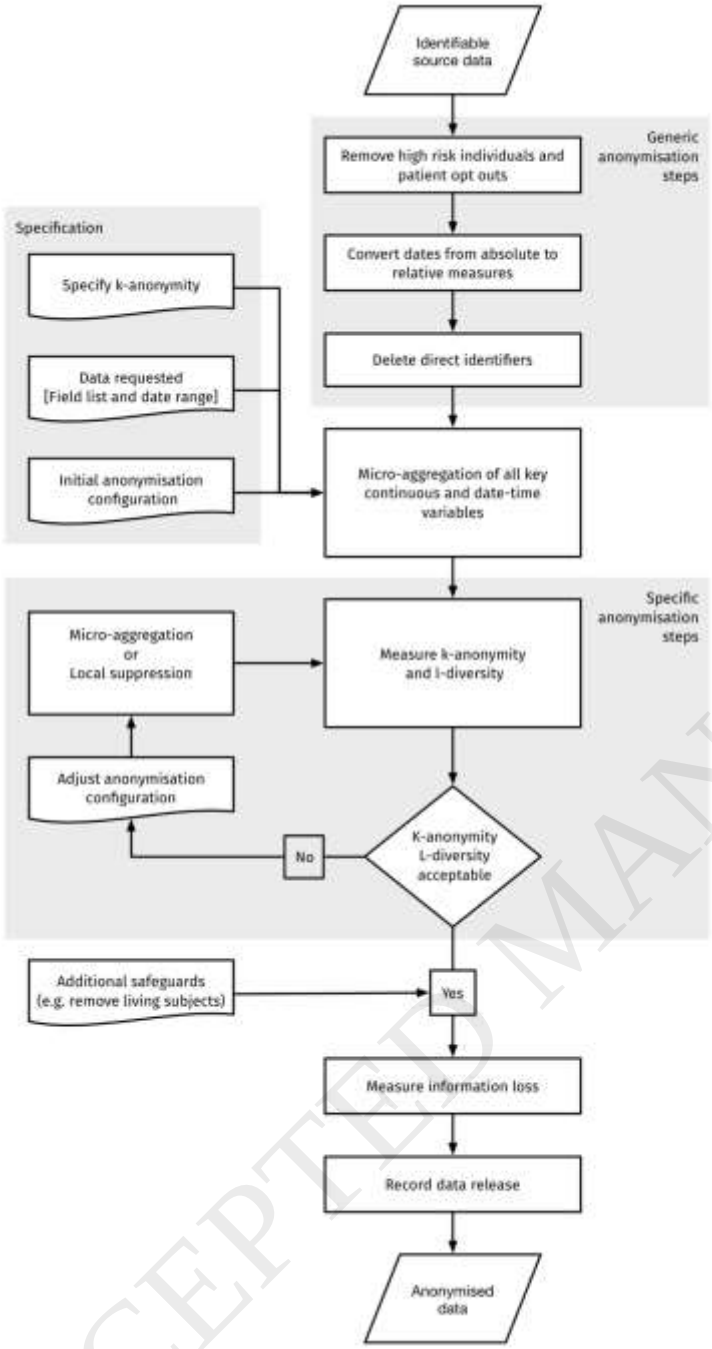


Figure 2: Data anonymisation: algorithm implementation: A summary of the anonymisation process applied before any data release.

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 eye colour, but not 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). These 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.

4.2.5.4 Data anonymisation: 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 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. The data may be suitable for a complete analysis but this will depend on the balance between the fields requested, and resolution required. 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.

4.2.5.5 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 most researchers with the huge task of cleaning the data. We therefore provide alongside the data a set of tools covering the most common data pre-processing and post-processing operations. These are provided as an open source package cleanEHR for the R statistical programming language.

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 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.* The end result is a data frame that is ready for analysis in applications from Microsoft Excel to SPSS, from R to Python.

A second function is used to stitch together separate but sequential critical care admissions into a unified illness spell. Regardless of whether care for that spell of illness is provided in a single facility, or across multiple facilities, the longitudinal data is appropriately concatenated. This is a particular problem in the UK where similar patients may step down from an ICU to an High Dependency Unit (HDU) in one institution, but may have all their care delivered in a single critical care unit in another institution.

Additional functionality includes the ability to relabel the data fields at will, to perform

* 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.

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 file 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. The *cleanEHR* toolkit is freely available from the Comprehensive R Archive Network (CRAN) and GitHub.

* 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. [10]

5 Current data

The initial data set specification (version 1.0) was released to contributing sites in 2013. Data collection started in 3 ICUs from 3 hospitals in February 2014, and expanded to 11 ICUs from 5 hospitals by July 2017 with regular quarterly updates by which time, the database contained 21930 critical care admissions.

The data set contains 258 variables describing each admission plus additional unit and hospital level metadata. 165 variables are time-dependent (e.g. drugs, physiology etc.), and the remaining 93 are captured on admission or discharge to the ICU, or discharge from the hospital. We used the ICNARC coding method to capture admission diagnosis as per the UK's national audit.[11] The data specification permits multiple levels of metadata to be associated with each measurement (i.e. site and units of measurement, route and method of drug administration etc.). We hope to expand the data set to include additional structured data items, narrative text, and waveform data in the near future.

A typical admission would have 70 time-invariant measures, and a median of 1030 (IQR 481 to 2335) time-varying measures. The database therefore contained more than 60 million data items plus associated meta data.

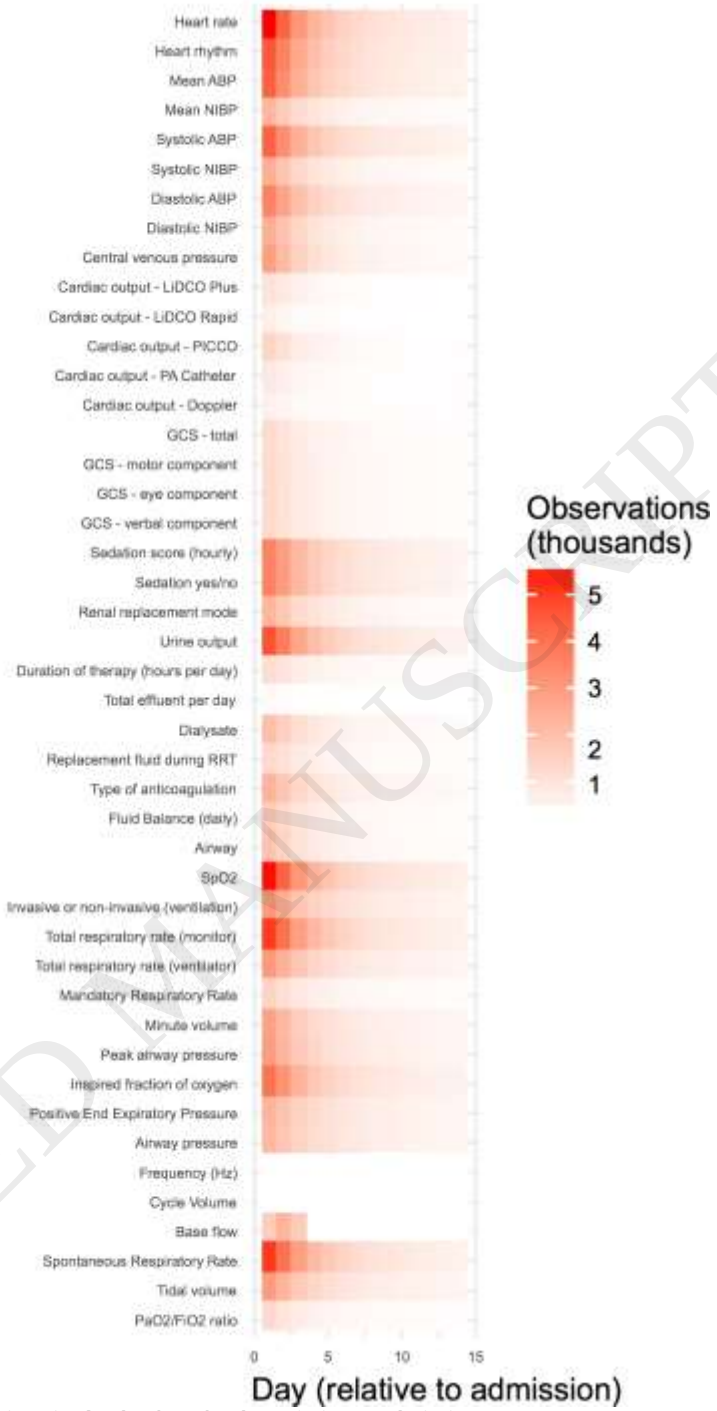


Figure 3: Number of physiology observations in the database by day relative to admission

A user may therefore recreate, in detail, the longitudinal profile of an individual patient, or examine the distribution of variables across all patients.

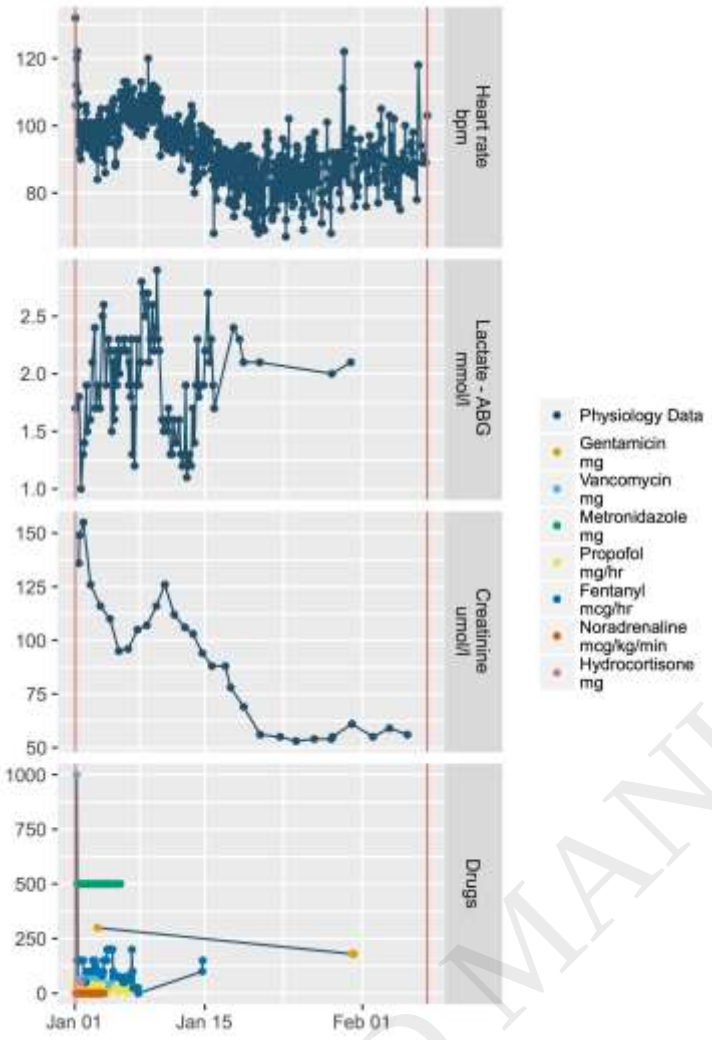


Figure 4: Selected physiology measures and drug administration from an admission with Inhalation pneumonitis

CCHIC population across each individual ICU unit and in total with descriptive statistics

		Demographics	Drugs	Laboratory	Other	Physiology	All
On admission							
	Dates	1					1
	Admission descriptors	6					6
	Patient characteristics	12					12
	Pre-admission descriptors	10					10
	Prognostic scoring information	24					24
On discharge							
	Dates	2					2
	End of life	12					12
	Episode descriptors	6					6
	Organ dysfunction summary	1					1
	Post-admission descriptors	1					1
	Prognostic scoring information	4					4
Late follow-up							
	Dates	3					3
	Episode descriptors	2					2
	Post-admission descriptors	9					9
Daily							
	Organ dysfunction	9				9	
	Fluid balance					1	1
Within 30 minutes of input							
	Anti-microbials		45				45
	Cardiovascular					14	14
	Chemistry			15			15
	CNS		9				9
	CVSvasoactive		15				15
	Dates	2					2
	Haematology			5			5
	Microbiology			3			3
	Neurology					6	6
	Position				1		1
	Renal					8	8
	Respiratory			4		17	21
	Temperature				2		2
		104	69	26	3	55	258

Table 2: Count of data fields (variables) classified by type and time dependence

6 Discussion

The widespread adoption of EHR platforms coupled with technical advancements in clinical information systems and biomedical information standards has enabled the collection and re-use of clinical data for research. Historically however, researchers typically only get to see the tip of the iceberg: coded administrative data relating to healthcare claims with mainly record billable diagnoses and procedures. The rich data generated across the clinical pathway remain submerged and inaccessible. It is to this challenge that CCHIC is responding.

6.1 Comparison with other databases

6.1.1 MIMIC

Notable resources already exist in the United States, such as the Medical Information Mart for Intensive Care III (MIMIC-III) database, but these are single centre initiatives.[12] MIMIC has nonetheless set the precedent for open access health data, and has been enormously successful in this regard. The full MIMIC database is available to researchers who complete a human research ethics training programme, and sign a data use agreement. In contrast, CCHIC makes available a restricted fully anonymised exemplar data set for exploration, and code development. Access to the full data set currently requires approval by the CCHIC data advisory group. Because the source data is fully identifiable, and until we have tested our anonymisation process more widely, we feel this is an appropriate balance. The only external data that is routinely linked to MIMIC is mortality via social security records. CCHIC, in contrast, was designed from the outset to link regularly to a wide range of health and social care databases. The aim is to eventually collate a cradle to grave perspective of health for patients who experience critical illness. Permissions are already in place to link to hospital episode data thereby defining secondary care use following discharge, and comorbidities prior to admission. Permissions will next be sought for long term survival, and primary care episodes.

6.1.2 ICNARC

The other major UK critical care database belongs to the Intensive Care National Audit and Research Centre's Case Mix Programme (ICNARC CMP). This is now, with the exception of Scotland, a national audit programme with more than twenty years of data. However, the CMP is designed for benchmarking not research.[13] As such it only contains selected data during the first 24 hours of admissions to critical care with summarised outcome measures. Linkage is possible but not explicitly part of the remit of the design.
We see ICNARC and CCHIC as two synergistic programmes: one with a wide-angled historical view, and one with a detailed, longitudinal view enriched with secondary sources.

6.2 Limitations

The future of CCHIC depends on our meeting the obligations to patient privacy, and research for patient benefit. The initial technical hurdle has been in transforming the database into a research ready resource. In this we believe we have made significant progress.
The next major technical challenge is to extend the data set, and expand the group of participating hospitals beyond the founding academic centres. Currently, both of these endeavours would require individual sites to write further local ETL (extract, transform, load) schemes. This is a significant burden that is multiplied by each data request and each participating site. Our experience is that even where sites share similar EHR systems, each has been so extensively modified that the ETL scripts are not transferrable.
One solution is to shift the burden of data *transformation* centrally. [14] Each site is then only required to write a smaller data *extraction* routine. This routine identifies data items associated with critical care admissions (e.g. by filtering HL7 messages), and then transfers

them centrally. Since all messages are archived, the data set can be expanded variable by variable as transformation and loading routines are developed. Moreover, these could be applied retrospectively to the existing data archive. The barrier to new sites joining would also be much lower.

6.3 Conclusion

Making health care data available to researchers is a huge challenge. The data is both sensitive, and the research needs are many. Resources such as MIMIC and ICNARC already have their own answers to this, but CCHIC brings several advantages. It is an explicitly linkable, multi-centre collaboration with a focus on making the data *research ready*. This is more than the technical challenge of protecting personal information, and effectively anonymising data. It is also about creating a culture that promotes collaboration and the best quality reproducible science, and we therefore look forward to meeting our future collaborators.

7 Authors' contributions

Manuscript preparation: SH, SS, SD
Concept and design of database and model catalogue: NM & DB
Concept and design of data pipeline: SH, SS, DPS, NM
Design, data sharing and critical review: AE, PW, AJ, SA, RB, DY, SB, MS

8 Acknowledgements

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.

9 Statement on conflicts of interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed.. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

10 Summary table

- Electronic health record (EHR) research is being led by critical care (for example, the MIMIC database at MIT) however all these projects face a common set of conflicting challenges: patient privacy, and usability for the researcher
- CCHIC is a new multi-centre critical care database from the UK that holds data from five hospitals, eleven ICUs and more than 20,000 admissions
- Uniquely CCHIC is an explicitly linkable with patient identifiers retained to allow mapping of health from the cradle to the grave
- CCHIC is provided with a set of *research ready* open source software tools in order to facilitate the final part of the contract with patients in using their data: that we can show patient benefit. These tools include:
    - Anonymisation
    - Data cleaning
    - Data extraction in a language agnostic manner

11 References

[1] S.M. Stigler, The History of Statistics, Harvard University Press, 1986.

[2] Standard ISB 0086: Information Governance Toolkit, (2017). http://webarchive.nationalarchives.gov.uk/+/http://www.isb.nhs.uk/library/standard/151 (accessed September 28, 2017).

[3] H. Shah, The DeepMind debacle demands dialogue on data., Nature. 547 (2017) 259–259. doi:10.1038/547259a.

[4] J. Powles, H. Hodson, Google DeepMind and healthcare in an age of algorithms, Health Technol. 29 (2017) 1–17. doi:10.1007/s12553-017-0179-1.

[5] NIHR HIC Locality: Critical Care, (2015). http://www.hdf.nihr.ac.uk/catalogue/#/catalogue/dataModel/13 (accessed September 28, 2017).

[6] [1611.05146] A Semi-Markov Switching Linear Gaussian Model for Censored Physiological Data, (n.d.). https://arxiv.org/abs/1611.05146 (accessed September 28, 2017).

[7] Information Commissioner's Office, Anonymisation: managing data protection risk code of practice, 2014.

[8] M. Templ, A. Kowarik, B. Meindl, Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro, Journal of Statistical Software. 67 (2015). doi:10.18637/jss.v067.i04.

[9] M. Templ, B. Meindl, A. Kowarik, S. Chen, Introduction to Statistical Disclosure Control (SDC), 2014.

[10] YAML Ain't Markup Language (YAML™) Version 1.2, (n.d.). http://www.yaml.org/spec/1.2/spec.html (accessed September 29, 2017).

[11] J.D. Young, C. Goldfrad, K. Rowan, Development and testing of a hierarchical method to code the reason for admission to intensive care units: the ICNARC Coding Method, Br J Anaesth. 87 (2001) 543–548. doi:10.1093/bja/87.4.543.

[12] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data, Published Online: 15 March 2016; | Doi:10.1038/Sdata.2016.18. 3 (2016) 160035–10. doi:10.1038/sdata.2016.35.

[13] D.A. Harrison, A.R. Brady, K. Rowan, Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit & Research Centre Case Mix Programme Database., Crit Care. 8 (2004) R99–111. doi:10.1186/cc2834.

[14] C.B. Turley, Leveraging a Statewide Clinical Data Warehouse to Expand Boundaries of the Learning Health System, eGEMs (Generating Evidence & Methods to Improve Patient Outcomes). 4 (2016). doi:10.13063/2327-9214.1245.

12 Appendices

12.1 Data specification