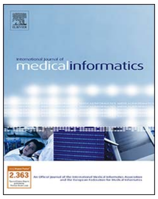
International Journal of Medical Informatics 112 (2018) 82–89

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



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

Steve Harris[a,i,*,1], Sinan Shi[b,1], David Brealey[a,i], Niall S. MacCallum[a,i], Spiros Denaxas[h], David Perez-Suarez[b], Ari Ercole[c], Peter Watkinson[d], Andrew Jones[e], Simon Ashworth[f,g], Richard Beale[e,j], Duncan Young[d], Stephen Brett[f,g], Mervyn Singer[a]

[a] *Bloomsbury Institute of Intensive Care Medicine, University College Hospital, London, UK*
[b] *Research Software Engineering, University College London, London, UK*
[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, UK*
[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*
[g] *Critical Care, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK*
[h] *Institute of Health Informatics, University College London, Gower Street, London, WC1E 6BT, UK*
[i] *Critical Care, University College London Hospitals NHS Foundation Trust, London, UK*
[j] *Division of Asthma, Allergy and Lung Biology, King's College, London, UK*

ARTICLE INFO

Keywords:
Electronic health records
Database
Clinical decision support
Critical care
Reproducibility

ABSTRACT

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

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 481–2335) 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.

---

* Corresponding author at: Steve Harris Critical Care Department, University College Hospital London, 235 Euston Road, London, NW1 2BU, UK.
E-mail address: s.harris8@nhs.net (S. Harris).

[1] Joint first authorship (based on contribution to research and manuscript preparation).

https://doi.org/10.1016/j.ijmedinf.2018.01.006
Received 2 October 2017; Received in revised form 6 December 2017; Accepted 8 January 2018

1. 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.

1.1. 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

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

2. 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.).[2] Data are currently exported on a quarterly basis with the ambition to move to near realtime collection (Table 1).

2.1. 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.[3]

2.2. 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

---

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

[3] 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)