MIMIC code: a repository for deriving clinical concepts

Alistair E. W. Johnson [a], Tom J. Pollard [a]

[a] *Massachusetts institute of Technology, Cambridge*

Abstract

Secondary analysis of electronic health records is an important method for gaining insight into clinical care. Retrospective studies frequently require similar clinical concepts, so there is benefit in providing open, standardized tools for deriving these concepts to ensure the consistency and efficiency of future studies. We present the MIMIC Code Repository, a collection of open source code for deriving clinical concepts using the MIMIC Critical Care Database. Concepts include severity of illness scores, organ failure indices, and duration of treatments such as ventilation and dialysis. The MIMIC Code Repository is in active development on GitHub. All code is made available under a permissive MIT license unless otherwise indicated.

Key words: critical care; open data; data mining; secondary use of electronic health records.

1 Introduction

There is substantial heterogeneity in intensive care populations, particularly in aspects such as patient physiology, presence of disease, and intervention types. This heterogeneity presents a significant challenge to understanding the relationships between provision of care and patient outcomes, and as a result the benefit of many widely-practiced treatments and interventions remains inconclusive [REFS: Ioannadis].

Vast quantities of data are routinely collected by modern hospital monitoring systems and even more so in intensive care units where patients frequently suffer organ failure and require close observation. There is optimism that increasing availability of large scale clinical databases will offer opportunities to

Email addresses: aewj@mit.edu (Alistair E. W. Johnson), tpollard@mit.edu (Tom J. Pollard).

overcome many of the challenges associated with heterogeneity and offer new insights into critical care medicine [REF: BIG DATA etc].

One such database is the Medical Information Mart for Intensive Care (MIMIC-III), collected from patients admitted to intensive care units in the Beth Israel Deaconess Medical Center, Boston, MA, USA [**?**]. The latest version of MIMIC-III, v1.3, houses data spanning 11 years between 2001 and 2012, and is made freely available to researchers upon signing of a data use agreement and proof of a human studies training course.

MIMIC-III is an unmatched research resource in the area of critical care informatics that promotes crowdsourcing of knowledge generation and sharing of fully reproducible studies. Analysis of critical care data often requires definition of clinical concepts, such as severity of illness scores, organ failure indices, and duration of treatments including ventilation and dialysis. Historically, this code has been produced by independent researchers; a process which is time consuming, inefficient, and error-prone. Furthermore, many decisions are made when extracting clinical concepts, and these decisions may have a large impact on the resultant analysis. One example is code written to extract the Glasgow Coma Scale (GCS). Severity of illness scores stipulate a value of 15 ("normal") should be assigned when the GCS cannot be obtained (e.g. due to patient sedation). However, it is common for clinical staff to record values of 3 for the GCS of sedated patients. This results in a systematic bias for sedated patients unless appropriate measures are taken to correct the recorded values. Even if correctly performed, these details would likely be omitted from publications, making study methods difficult to reproduce.

There is a great opportunity for unifying studies on MIMIC with the creation of a centralized repository for data extraction code. Here we describe such a centralized repository, with the hope that is fulfills this need and enhances the reproducibility of research on the MIMIC-III database.

2 Results

A prerequisite for using much of the code in the MIMIC Code Repository is access to the MIMIC-III Database, so we provide scripts to enable researchers to build local copies of the MIMIC database in a variety of database systems including PostgreSQL, MySQL, Oracle, and MonetDB. The set of core clinical concepts which have been extracted using structured query language (SQL) [1] are as follows:

[1] All queries have been developed and tested using PostgreSQL 9.5.1.

The Angus criteria for defining sepsis: Sepsis is a serious illness caused by infection and is a major focus of clinical research. Angus criteria utilize billing codes to classify a hospital admission as being related to sepsis [REF TO ANGUS], and the criteria have been recently validated [RECENT ANGUS VALIDATION].

Severity of illness scores: measures of illness are a useful tool for adjusting for patient severity when comparing patient populations. Scores implemented include the Simplified Acute Physiology Score (SAPS), Simplified Acute Physiology Score II (SAPS II), the Acute Physiology Score III (APS III), and the Oxford Acute Severity of Illness Score (OASIS) [REFS].

Organ failure scores: Multi-organ failure is a hallmark of acute illness and quantify the morbidity for a given patient. The Sequential Organ Failure Assessment (SOFA) score [REF] and Logistic Organ Dysfunction System (LODS) [REF] both assess six organ systems for failure. Single organ failure scores implemented include MELD [REF], commonly used to determine suitability for a liver transplant, RIFLE [REF], which quantifies acute kidney injury, and KDIGO [REF], also used for acute kidney injury.

Comorbidity scores: Critically ill patients frequently have many comorbidities which influence both their overall health and their individual stay. The Elixhauser Comorbidity Index summarizes the level of comorbidities in a patient using billing codes collected at hospital discharge [REFS].

Treatment durations: Studying the effect of treatments on patient health is of great interest, though deriving the timing of these treatments from a database can be non-trivial. We provide views with (i) vasopressor use start and stop times for all vasopressors and individual medications and (ii) mechanical ventilation start and stop times. The duration of these interventions is also a useful measure of treatment intensity.

More detail on the extraction of these concepts is provided in the supplemental material. An example of the importance of the variance caused by a non-centralized code base is shown in 1, where the performances of two different implementations of the SOFA score in discriminating hospital mortality are shown. Both of these implementations have been used in previous publications.

3 Discussion

Numerous research studies have been carried out on the MIMIC database in the past, but the code used in analysis has largely been developed independently and often not shared. Lack of code sharing is typical of scientific soft-

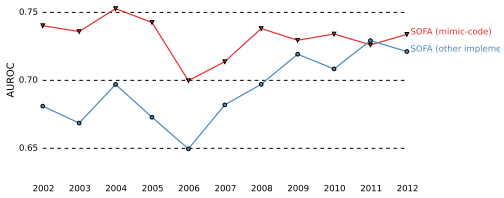

Fig. 1. Discrimination of two implementations of SOFA across fiscal years as measured by the area under the receiver operator characteristic curve (AUROC).

ware development and is a widely recognized issue [Ref: http://www.nature.com/news/2010/101013/
By creating the MIMIC Code Repository we have provided a central hub for development of clinical concepts, which we believe will help to standardize and improve future analyses. The repository is intended to be a continuously developed resource enhancing the sustainability of the code and creating a community around data analysis in MIMIC.

The code described herein follows guidelines for good practice in scientific programming, including incremental development with a distributed version control system, unit tests, and a public issue tracker [Ref: G Wilson paper].

The MIMIC Code Repository is an important resource for researchers working with the MIMIC critical care database. The repository provides code for deriving a variety of clinical concepts and will continue to incorporate new concepts as they are calculated, allowing for rapid prototyping of clinical questions in a large retrospective database. Fully reproducible analytical workflow are encouraged by the openly available nature of both the data and code. Finally, the code written is modular and generalisable, and may support research in other similarly structured clinical databases.

4 Materials and methods

Acknowledgements

The authors would like to thank Professor Roger G. Mark, the MIT Laboratory for Computational Physiology, Philips Healthcare and the Beth Israel Deaconess Medical Center for the creation of the MIMIC-III database.

Funding

This work has been supported by grants NIH-R01-EB017205, NIH-R01-EB001659, and NIH-R01-GW104987 from the National Institutes of Health.