The MIMIC Code Repository - Towards Reproducibility in Secondary Analysis of Health Records

Alistair E. W. Johnson[1*] David J. Stone[2]
Leo Anthony Celi[1,3] Tom J. Pollard[1]

April 01, 2017

Abstract

Lack of reproducibility in medical studies is a barrier to the generation of a robust knowledge base to support clinical decision-making. In this paper we outline the MIMIC Code Repository, a framework for generating reproducible studies on an openly available critical care dataset. By providing open source code alongside the freely accessible MIMIC-III database, we enable end-to-end reproducible analysis of electronic health records.

1 Introduction

Concerns about the reproducibility of results in science are becoming increasingly prominent in both scientific and mainstream literature (Baker 2016). Some commentators have gone so far as to call the current state a 'crisis', citing causes such as pressure to publish positive results, the cost of replicating studies such as double blind randomized controlled clinical trials, and the lack of emphasis on reproducibility as a requirement for sound science [Leo Anthony Celi "Beyond Open Data"]. In medicine, the heterogeneity of treatment effect across subpopulations presents an additional challenge in

the evaluation of interventions [reference]. As a result, the risk-benefit profile of many widely-practiced treatments and tests remains unknown [reference].

In parallel, health care has been undergoing a digital revolution in recent years. The Health Information Technology for Economic and Clinical Health Act has catalyzed the transition of hospitals and care institutions from paper based systems to electronic ones [?]. Vast quantities of digital data are now routinely collected by modern hospital monitoring systems, even more so in intensive care units (ICUs) where patients require close observation. There is optimism that increasing availability of large scale clinical databases will offer opportunities to overcome many of the challenges associated with lack of evidence in medical practice [REF: BIG DATA etc].

The Medical Information Mart for Intensive Care (MIMIC-III) database is an example of such a data repository (Johnson et al. 2016). The database comprises detailed clinical information regarding over 60,000 stays in intensive care units (ICU) at the Beth Israel Deaconess Medical Center in Boston, MA, USA, collected as part of routine clinical care. Uniquely, the MIMIC-III dataset is freely available to researchers around the world, for use in research and education. Derivation of key clinical concepts on an EHR database is a resource-intensive task, however, and is a significant barrier to those unfamiliar with the clinical environment. Moreover, if concepts are not defined collaboratively with those who are familiar with the workflows, including how the data is captured, the validity of any findings becomes suspect.

In this paper, we describe the MIMIC code repository, a large body of work which derives concepts that are relevant to critical care research. Detailed descriptions on how the concepts are defined and extracted from the database are provided, including the assumptions that are made and the conditions for which a code or query is valid. The code is open source, follows good documentation practices, and is contributed to by members of the research community using MIMIC-III.

The repository provides a framework for collaboration around research. While the case for open data has been already been strongly made [references], we believe *open code* is equally important. We would make the argument that the use of an openly available code repository will improve secondary analysis of health data by accelerating the understanding of datasets by researchers, and improving the consistency and validity of future studies.

2 The MIMIC Code Repository

The MIMIC code repository is available online [?] and is open source. Code is available as standardised scripts in languages including SQL, Python, and R. Scripts are modified to allow an individual who has been granted access to the MIMIC-III database to generate a number of "views" of the data, with each view being an extraction from the raw data. Each script is associated with an automatically generated unique commit hash that acts as an identifier for the code. Publications that use the code repository can further cite the commit hash, allowing other researchers to download a copy of the code used regardless of any modifications since. All code follows the principles of good scientific programming as outlined by Wilson et al [Ref: G Wilson paper], including incremental development with a distributed version control system, unit tests, and a public issue tracker. The repository was tested on MIMIC-III v1.4 at the time of this publication.

There are three components to the repository that facilitate navigation of the data for research purposes. These components are:

1. Executable documents: notebooks that allow text and analytical code to be seamlessly combined into a single executable document, allowing studies and tutorials to be reproduced.
2. Concepts: code to extract important concepts from the health records. For example, a module on acute kidney injury (AKI) uses the criteria as specified by the Kidney Disease Improving Global Outcomes (KDIGO) and provides the code to identify patients with AKI in MIMIC.
3. Community: public discussions to facilitate contributions from members of the MIMIC research community

2.1 Executable documents

When both data and code is freely available to researchers - as is now the case for MIMIC-III - this provides a framework that allows a study to be entirely reproduced. This is especially powerful when toolkits such as R Markdown and Jupyter Notebook are employed, allowing documentation and code to be seamlessly combined to create executable documents. Figure 1 shows an example of a Jupyter Notebook that extracts patient demographics and displays the results for the user to view. Jupyter Notebooks are language

agnostic, supporting code written in Python, R, MATLAB, SAS, and others (Pérez & Granger 2007, Kluyver et al. (2016)).

If we are interested in the length of stay for the ICU patients, we can query the `intime` and `outtime` columns, adding in some SQL specific syntax for calculating the difference between two dates.

In [3]:
```
query = query_schema + """
SELECT subject_id, hadm_id, icustay_id
, outtime - intime as icu_length_of_stay_interval
, EXTRACT(EPOCH FROM outtime - intime) as icu_length_of_stay
FROM icustays
LIMIT 10
"""
df = pd.read_sql_query(query, con)
df.head()
```

Out[3]:

	subject_id	hadm_id	icustay_id	icu_length_of_stay_interval	icu_length_of_stay
0	268	110404	280836	3 days 05:58:33	280713.0
1	269	106296	206613	3 days 06:41:28	283288.0
2	270	188028	220345	2 days 21:27:09	250029.0
3	271	173727	249196	2 days 01:26:22	177982.0
4	272	164716	210407	1 days 14:53:09	139989.0

Figure 1: Example of a notebook providing a tutorial with MIMIC-III data.

We have found executable documents particularly valuable for research in cross-disciplinary fields such as healthcare, because they facilitate collaboration between data analysts and domain experts. Notebooks primarily serve three purposes: (i) they allow documentation of the logic behind the code in an organized and easy to read manner; (ii) they aid rapid writing of the code particularly during group discussions; and (iii) they provide a means of sharing details of a published study that captures the learning that takes place during the evolution of a research project.

Executable documents are also an platform well-suited to tutorials. Harmonisation of text and code allows for explanations of the subject matter, while the interactive nature of the document allows for experimentation and facilitates learning. A number of tutorials have been made available to explain key concepts important for working with MIMIC. For example, the transformation of recorded clinical parameters, such as hemofiltration settings, into desired clinical concepts, such as length of continuous renal replacement therapy (CRRT), is non-trivial and requires both domain and database expertise. A *CRRT* tutorial overviews the process of exploring MIMIC-III, assessing the data stored within and producing a measure of the clinical concept of interest (duration of CRRT). The tutorial provides a starting point

for all researchers who work on the secondary analysis of electronic health records. Additional tutorials include, for example, an introduction to Structured Query Language; a step by step guide to select a study cohort, and an outline of the data capture process for commonly recorded parameters in the database.

2.2 Concepts

Code to extract concepts that that are broadly applicable to research questions in critical care are provided in the repository. For example, severity of illness scores are frequently required to adjust for confounding factors in a study, but are complex to derive, and so scripts are provided for reuse. These and other concepts are coded in a modular fashion to reduce redundancy in code and allow for extension. An example of the modular nature of the code is shown in Figure 2.

In the figure, a set of severity of illness scores is shown alongside a set of concepts that consitute separate components of the scores. Each component can easily be isolated and employed on its own, which could occur if, for example, the researcher is interested in determining which patients are mechanically ventilated on the first day (by using `ventfirstday`). The following sections describe various concepts currently available in the repository.

2.2.1 Severity of illness scores

Severity of illness scores have been developed over recent decades to provide an assessment of the patient's acuity, particularly but not exclusively, at the time of admission to the ICU (Knaus 2002). The principal aim of these scores is for risk-adjusting patient populations for benchmarking and research purposes such as comparison of cohorts in clinical trials and observational studies. In the context of performing research using MIMIC-III, the use of severity of illness scores for risk-adjustment is almost always required to address confounding.

While severity of illness scores are integral to risk adjustment, their calculation, if done retrospectively, presents challenges. Most severity scores were developed with well-curated datasets, put together through prospective data collection or manual data abstraction by dedicated trained personnel. As a

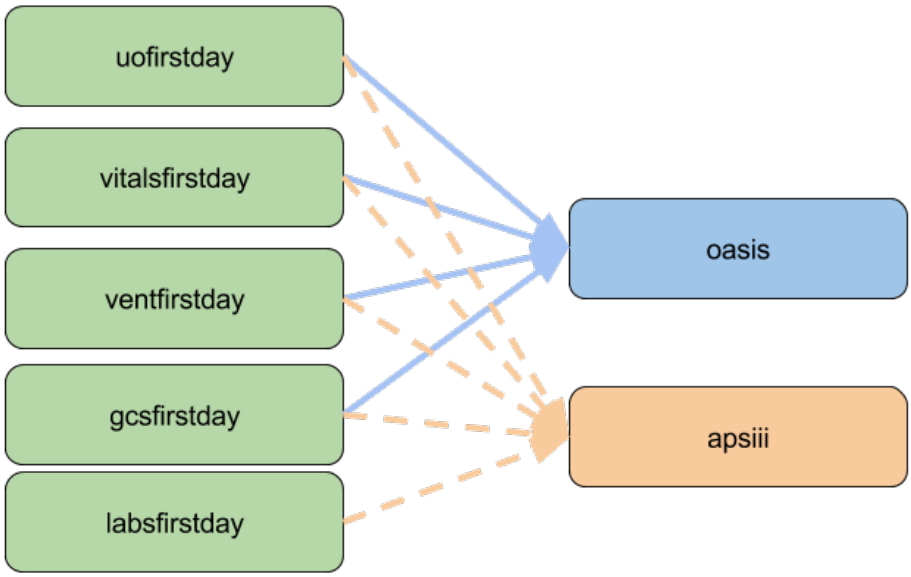

Figure 2: Block diagram demonstrating modular components of severity scores. These components can be used individually by researchers to quickly extract data of interest.

result, the data tends to be cleaner and often has, perhaps more importantly, a distribution that is markedly different from routinely collected data such as that present in an electronic health record.

Secondly, routinely collected data often lacks data elements required to compute the score. For example, the comorbidity ''biopsy proven cirrhosis" is required for the Acute Physiology Score and Chronic Health Evaluation (APACHE) system, but this concept is not documented in a structured manner during routine care. Finally, the data definitions for the same concept can vary between the original dataset used to define the severity score and the electronic health records being analyzed. To illustrate this potential disparity, the Glasgow Coma Scale (GCS), a common marker of neurological dysfunction which ranges from 3 (worst) to 15 (best), is usually assumed to be 15 for patients who are unable to be assessed due to sedation or ventilation, but otherwise appear to be neurologically intact. In an electronic health record however, this definition is not strictly adhered to as there is no defined protocol, and as a result, sedated patients may be assigned a score of 15 by some care providers, and a score of 3 by others.

Working with local nurses and doctors has helped us to address these kinds of issues that potentially impact the code, helping to ensure the derived scores accurately reflect the true severity of patient illness. There are five severity of illness scores currently implemented in the MIMIC Code Repository: APS-III (Knaus et al. 1991), SAPS (Le Gall et al. 1993), SAPS-II (J. R. Le Gall et al. 1996a), and OASIS (Johnson et al. 2013). A more detailed comparison of the severity scores is provided in the supplementary material, along with discussion of the assumptions made in calculating the scores. Organ dysfunction scores are also available and detailed later.

Each score is comprised of at least ten independent components. The APS III, SAPS II, SOFA, LODS, and OASIS scores are generally calculated using data from the first 24 hours of the patient's stay. SIRS and qSOFA are screening tools with scores calculated on admission to the ICU which is concretely defined as up to 2 hours after the admission time. Details of score derivation are available in the supplemental material (Appendix A). The distribution of these scores is shown in in Figure 3, and calibration curves are shown in Figure 4.

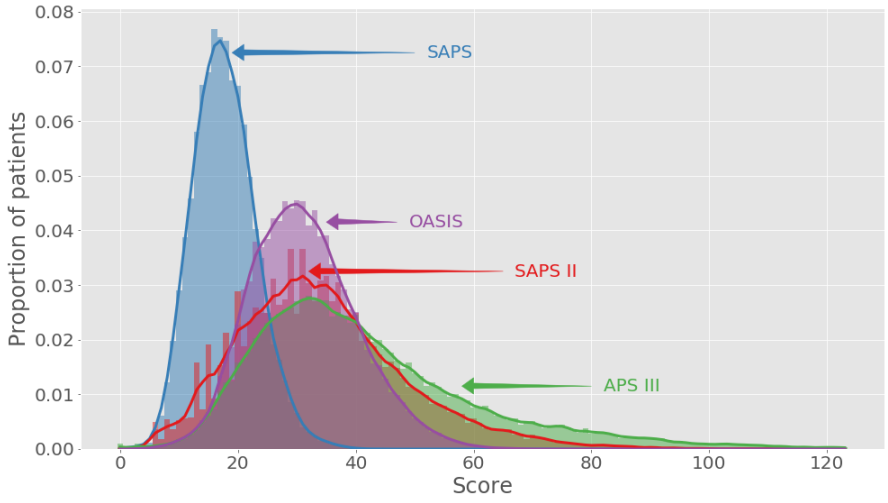

Figure 3: Comparison of severity of illness score distributions.

2.2.2 Organ dysfunction scores

Organ failure is a hallmark of acute illness and is quantified in numerous scores. Some scores assess multiple organ systems: the Sequential Organ Failure Assessment (SOFA) score (Vincent et al. 1996) and Logistic Organ Dysfunction System (LODS) (J. R. Le Gall et al. 1996b) both assess six organ systems for failure. Others are organ specific. Examples include the Model for End-stage Liver Disease (MELD) (Wiesner et al. 2003), the Risk/Injury/Failure/Loss/End stage renal disease (RIFLE) criteria (Bellomo et al. 2004, Kellum et al. (2002)), the Acute Kidney Injury Network (AKIN) classification (Mehta et al. 2007), and the Kidney Disease Improving Global Outcomes (KDIGO) criteria (CKD-MBD Work Group & others 2009). The latter three scores assess the degree of acute kidney injury in a patient. A variety of lab, diagnostic, and therapeutic data are needed to calculate these scores.

To highlight the discrepancies that can arise from the way a concept is defined, we contrast two versions of the SOFA score: one derived by prior researchers, and one available in the MIMIC code repository. Figure 5 shows the area under the receiver operator characteristic curve (AUROC) for hospital mor-

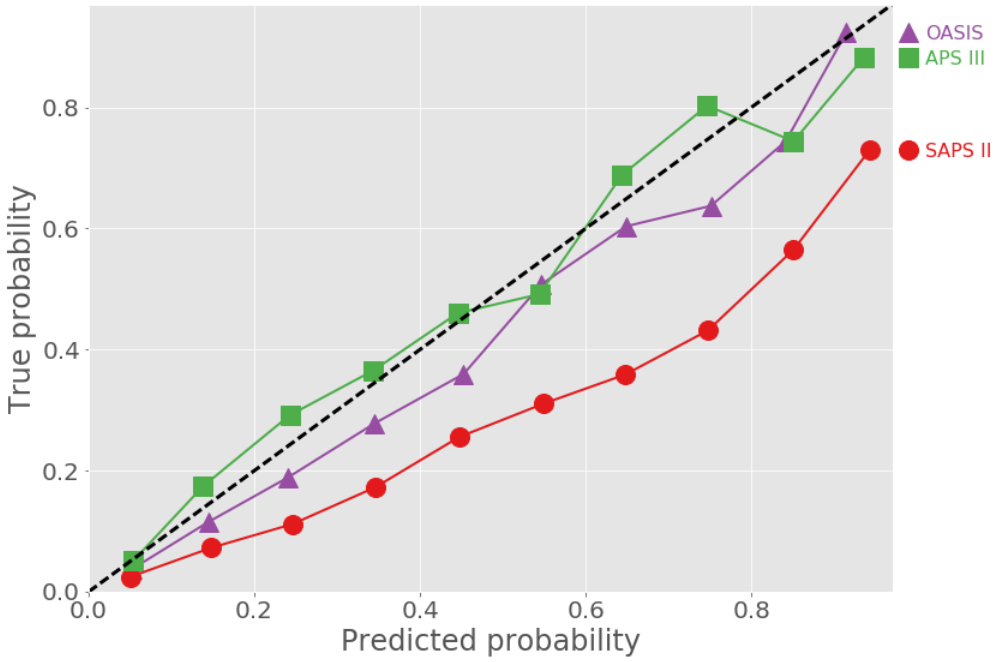

Figure 4: Calibration curves for three severity of illness scores with published equations for calculating the probability of mortality.

tality for patients admitted in the MIMIC-III database between 2001-2008 using two versions of SOFA, grouped by the year of admission.

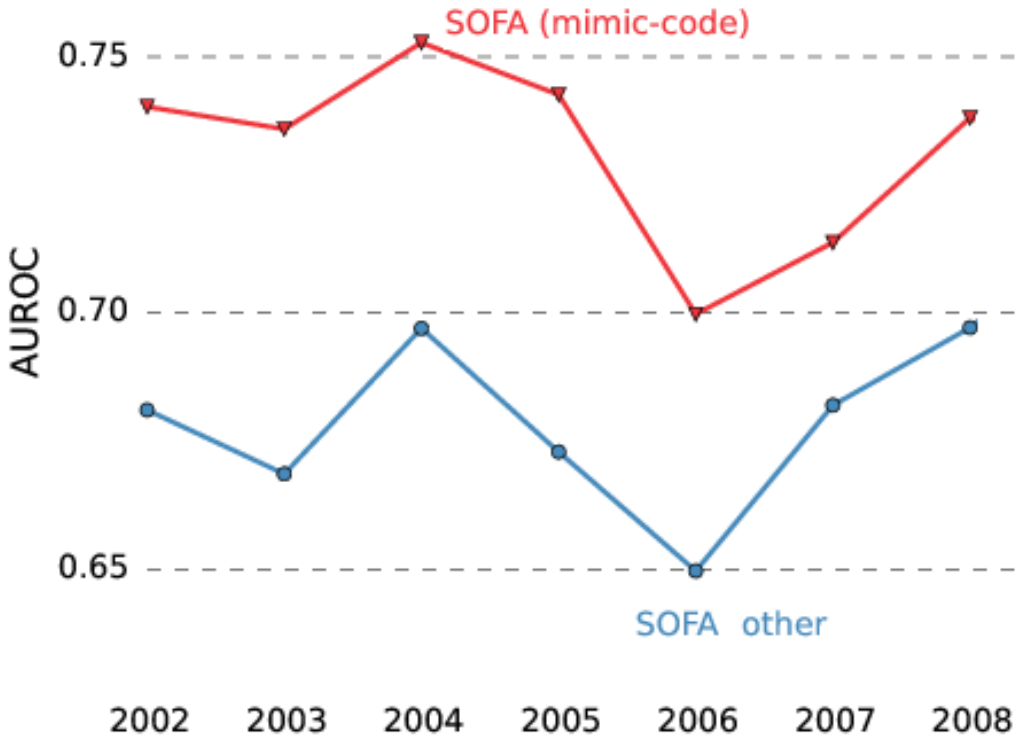

Figure 5: Comparison of AUROCs for SOFA scores calculated from mimic-code and a prior research report.

The disagreement between the two modalities is multifactorial, but a major contributing factor relates to an important variable: the Glasgow Coma Scale (GCS). In the original paper describing the SOFA score, clinicians were instructed to set GCS to its maximum value (15) if they were unable to assess the patient fully (for example, as a result of sedation to facilitate mechanical ventilation). In contrast, the documentation of GCS for these patients in the MIMIC-III database is usually a value of 3, the minimum value, with a note

that they are unable to assess the patient. Naive use of GCS values results in a dramatic differences in the capability of the score to discriminate severely ill patients and highlights the need to understand variables and how they are captured or derived. In the MIMIC Code Repository, special extraction steps are used to detect a GCS value of 3 due to sedation, and these values are corrected to 15 in the calculation of scores.

2.2.3 Timing of treatment

The timing and duration of treatment is an important concept for researchers who are interested in understanding issues that relate to the intensity of the administered intervention. Duration may serve as an indirect metric of severity and has been used in the development of decision support tools (Ghassemi et al. 2017).

Due to the method of data capture, the timing and durations of many medications and treatments are not explicitly available and as such must be derived. This derivation may involve identification of surrogate data documented by clinical staff contemporaneous to the treatment and done with a high level of compliance. Figure 6 shows a schema for the derivation of the start and stop times of mechanical ventilation. Similar rules are used to define the timing of vasopressor administration and CRRT available in the repository. Clinical expertise is invaluable in developing these rules and interpreting the fine points of the medical chart that determine them.

An example of a patient undergoing mechanical ventilation and receiving vasopressor agents is provided in Figure 7.

2.2.4 Sepsis

Sepsis is a major and costly disease in the ICU, costing over $20 billion in the US in 2011 (5.2% of all US hospital costs) (Torio CM 2013), and growing to over $23 billion in 2013 (6.2% of all US hospital costs) (Torio CM 2016). Sepsis has traditionally been defined as the concurrent presence of systemic inflammation andinfection, but a recent re-examination of the problem has suggested redefining the disease as a life-threatening organ dysfunction caused by a dysregulated host response to infection (Singer et al. 2016). The precise onset of sepsis is not typically documented in the EHR, and is, in

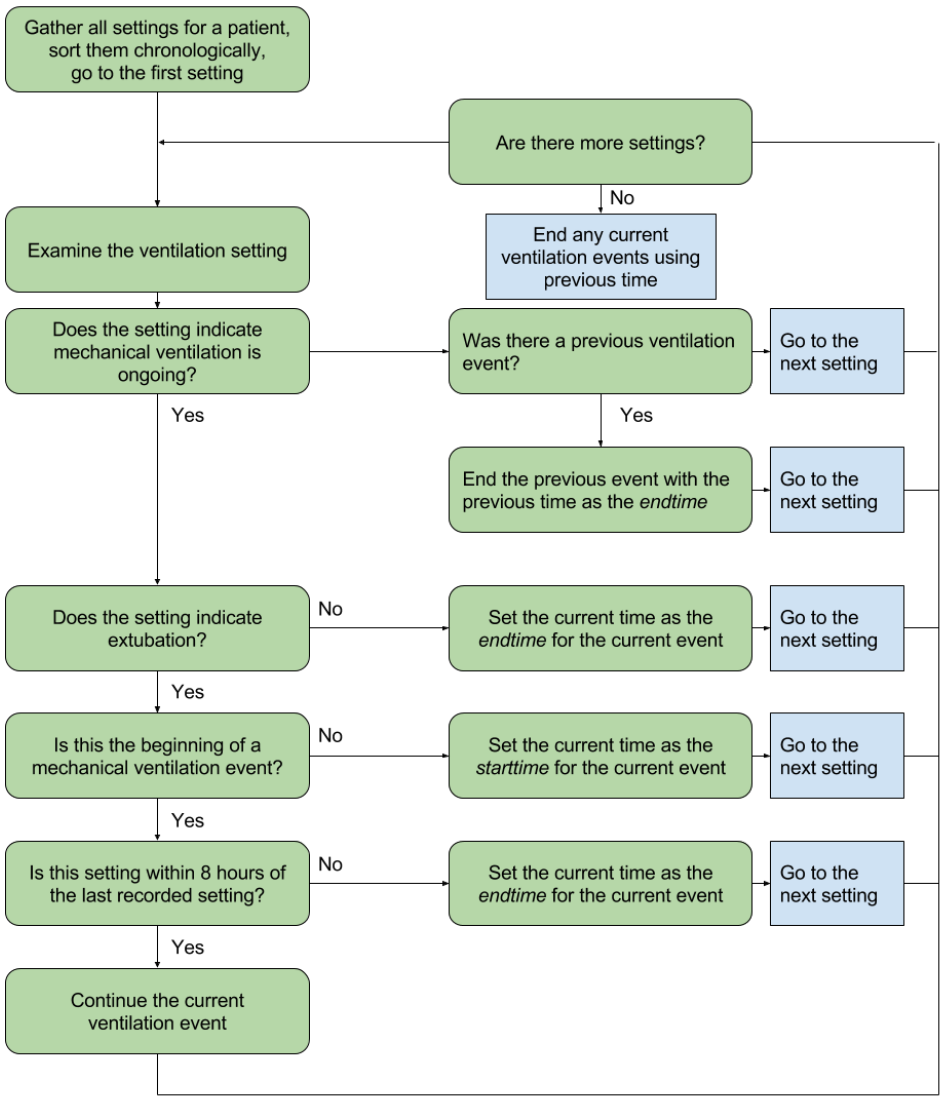

Figure 6: Logic behind the query for converting aperiodically recorded ventilator settings into durations of mechanical ventilation.

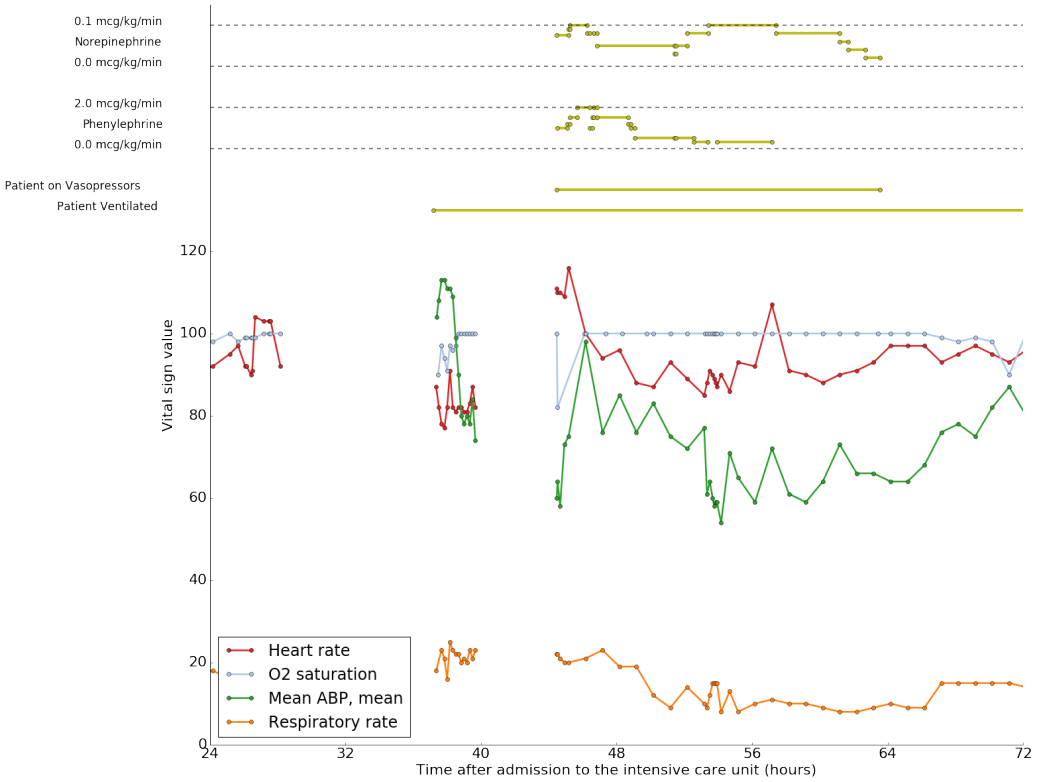

Figure 7: Example of a patient who was both mechanically ventilated and receiving vasopressors for cardiovascular support.

fact, a difficult item to capture clinically. In their quantitative evaluation of septic patients, Seymour et al. (Seymour et al. 2016) first identified patients suspected of infection by cross-referencing antibiotic use with requesting a microbiology assessment. We have implemented a similar approach, defining suspected infection as the acquisition of a microbiology culture followed by or shortly after ICU admission. Using this definition, and following the Sepsis-3 guidelines, we define sepsis as suspicion of infection associated with organ failure as quantified by an increase in SOFA >= 2. This definition is admittedly a proxy for the actual onset of sepsis, but in the absence of more precise markers, it serves as a reasonable approximation of onset time and could be used for development of decision support tools. Scripts for these concepts are available and a notebook describing the derivation is also available.

Identification of sepsis has also been done retrospectively using administrative data, and in particular billing codes acquired on hospital discharge. Angus et al. (Angus et al. 2001) and Martin et al. (Martin et al. 2003) describe algorithms for defining sepsis using a set of diagnostic and procedural ICD-9 codes. The criteria as proposed by Angus et al. (Angus et al. 2001) were validated in a later study by Iwashyna et al. (Iwashyna et al. 2014). Both criteria, those as proposed by Angus et al. (Angus et al. 2001) and those proposed by Martin et al. (Martin et al. 2003), are available in the repository. Figure 8 shows a Venn diagram for three groups of patients identified as septic.

2.2.5 Comorbidities

Many ICU patients have chronic conditions prior to their acute presentation that affect their probability of surviving critical illness. Elixhauser et al. (Elixhauser et al. 1998) codified these comorbidities into 29 categories using administrative data, specifically ICD-9 codes. The American Health and Research Quality group (AHRQ) continues to maintain these administrative codes via the Healthcare Cost and Utilization Project (HCUP), adapting them accordingly as changes are made to diagnosis and treatment coding (Steiner et al. 2001). Finally, Quan et al. (Quan et al. 2005) proposed an enhanced ICD-9 coding methodology based upon examining inconsistencies among previous definitions. Diagnosis related groups (DRG), which are used to bill for the principle diagnosis for a patient hospitalization, are used to

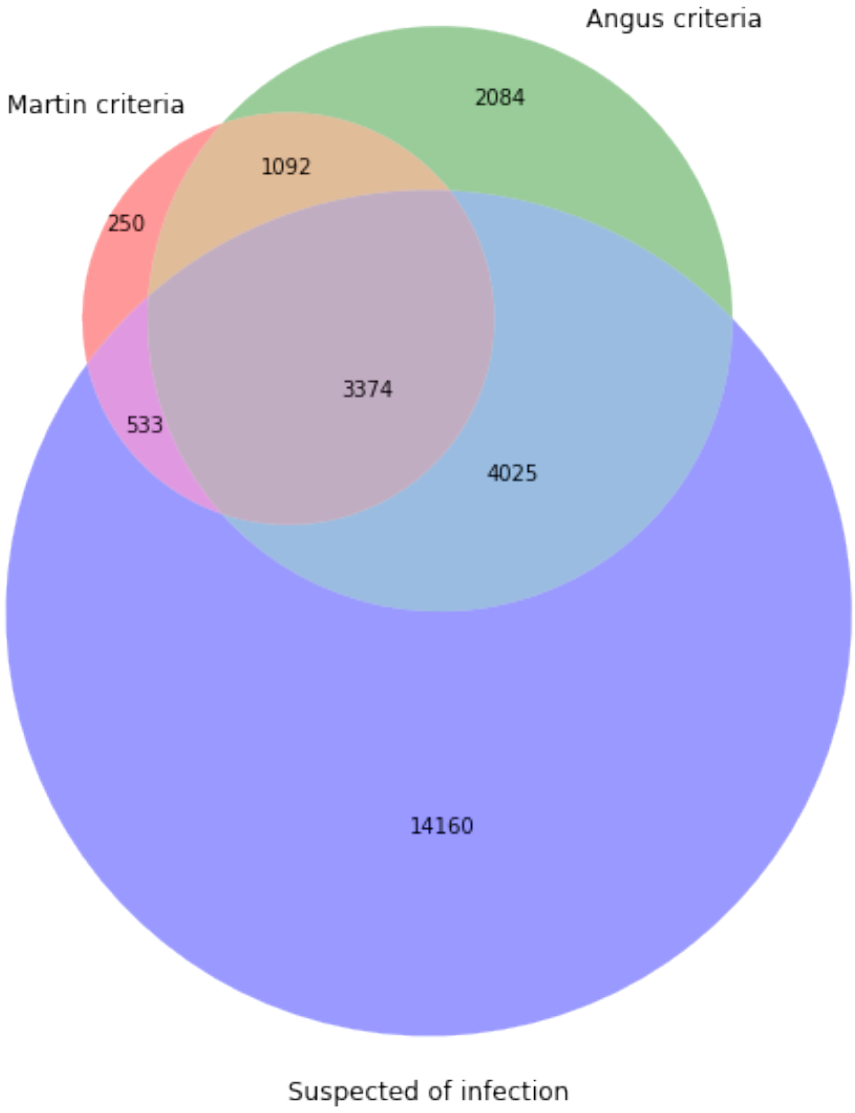

Figure 8: Venn diagram of three groups of patients who may have sepsis using: the Sepsis-3 clinical criteria, criteria proposed by Angus et al., and criteria proposed by Martin et al.

filter out those conditions that are not present prior to hospitalization. A comparison of these three methods is provided in Figure 9. These representations of comorbidities are provided in the repository, both with and without DRG filtering.

Van Walraven et al. (Walraven et al. 2009) later aggregated comorbidities codified by Elixhauser et al. (Elixhauser et al. 1998) into a single point score for in-hospital mortality prediction which is also available in the repository.

3 Conclusion

Transparent research processes can help to improve the quality of evidence that underpins health care. To achieve transparency, researchers must be able to provide both the data used for analysis *and* the code used to process it. By supplementing the MIMIC-III Database with the MIMIC Code Repository, we provide a framework to allow completely reproducible research in critical care. While cultural barriers that discourage some researchers from sharing code may pervade the research community, it is now clear that the barriers in the case of MIMIC-III are not technical. Readers and reviewers of MIMIC-III studies should be aware that a route for allowing reproducibility is available, whether or not the authors have chosen to take it.

Additionally, our efforts demonstrate how the concerns of open data raised by Longo et al in the well publicised editorial on Data Sharing might be addressed [longo2016data]. Longo argues that researchers not involved in the collection of data may lack understanding of its underlying detail, which we believe is a valid concern. Our framework connects researchers who reuse the MIMIC-III dataset with the laboratory and clinical staff who collect and produce the data, helping to provide context for downstream data analysis. Contributions to the MIMIC code repository by researchers are encouraged, progressively improving the codebase and helping to accelerate research in critical care. Source code control allows for transparency both in the authorship of the code and in the nature of modifications.

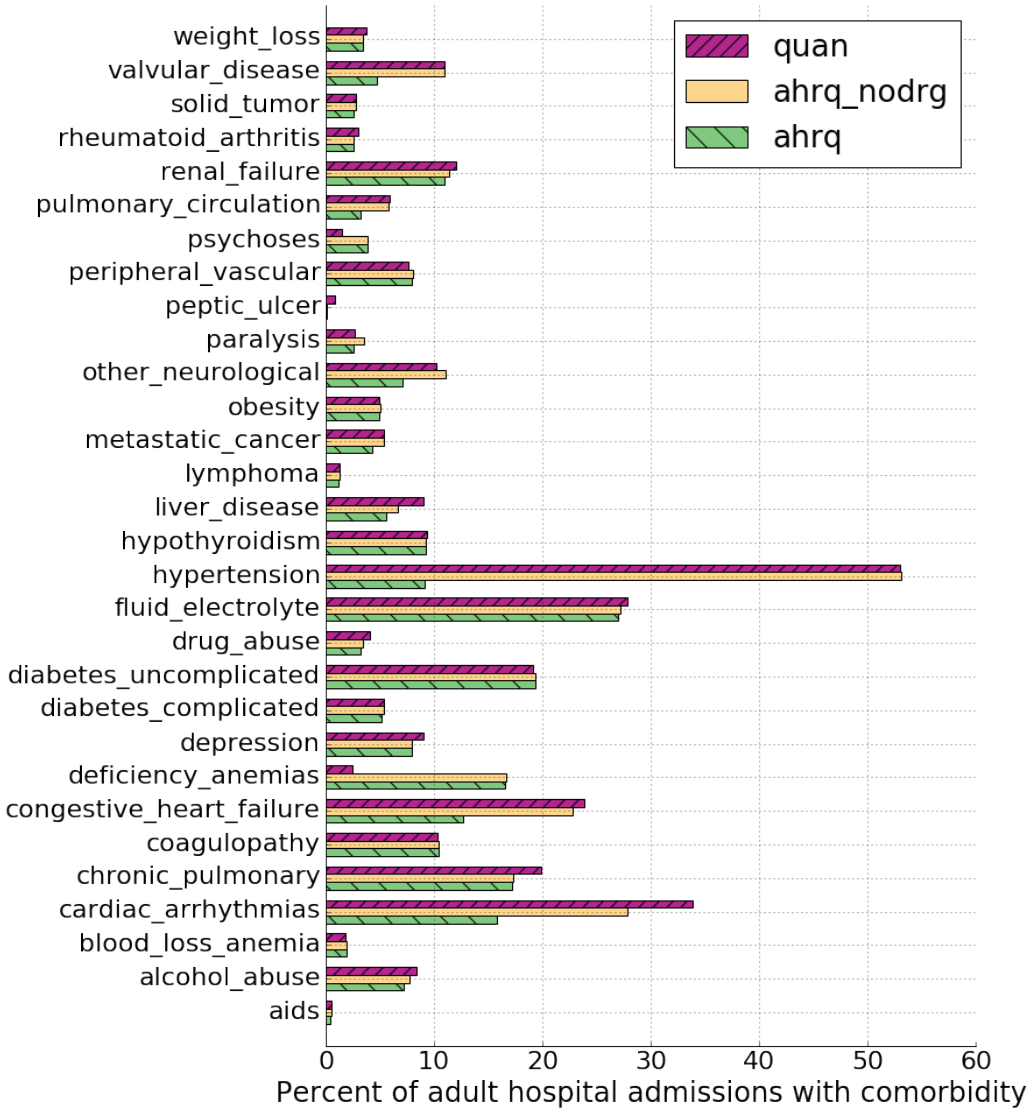

Figure 9: Comparison of three methods for calculating presence of a comorbidity for a patient using billing data: an updated coding from the AHRQ which uses DRG codes to mask non-comorbid conditions, the same coding without the DRG masking, and finally an alternative coding which does not use DRG masking proposed by Quan et al. [?].

Acknowledgments

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.

Author contributions

AEWJ and TJP collaborated to build the MIMIC code repository and write the paper.

Competing interests

The authors have no competing interests to declare.

References

Angus, D.C. et al., 2001. Epidemiology of severe sepsis in the united states: Analysis of incidence, outcome, and associated costs of care. *Critical care medicine*, 29(7), pp.1303–1310.

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(1), pp.452–454.

Bellomo, R. et al., 2004. Acute renal failure–definition, outcome measures, animal models, fluid therapy and information technology needs: The second international consensus conference of the acute dialysis quality initiative

(adqi) group. *Critical care*, 8(4), p.R204.

CKD-MBD Work Group, K.D.I.G.O. (KDIGO) & others, 2009. KDIGO clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of chronic kidney disease-mineral and bone disorder (ckd-mbd). *Kidney international. Supplement*, (113), p.S1.

Elixhauser, A. et al., 1998. Comorbidity measures for use with administrative data. *Medical care*, 36(1), pp.8–27.

Ghassemi, M. et al., 2017. Predicting intervention onset in the icu with switching state space models. *AMIA CRI*.

Iwashyna, T.J. et al., 2014. Identifying patients with severe sepsis using administrative claims: Patient-level validation of the angus implementation of the international consensus conference definition of severe sepsis. *Medical care*, 52(6), p.e39.

Johnson, A.E., Kramer, A.A. & Clifford, G.D., 2013. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Critical Care Medicine*, 41(7), pp.1711–1718.

Johnson, A.E. et al., 2016. MIMIC-iii, a freely accessible critical care database. *Scientific data*, 3.

Kellum, J.A. et al., 2002. The first international consensus conference on continuous renal replacement therapy. *Kidney international*, 62(5), pp.1855–1863.

Kluyver, T. et al., 2016. Jupyter notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, p.87.

Knaus, W.A., 2002. APACHE 1978-2001: the development of a quality assurance system based on prognosis: milestones and personal reflections. *Archives of Surgery*, 137(1), pp.37–41.

Knaus, W.A. et al., 1991. The APACHE III iii prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6), pp.1619–1636.

Le Gall, J., Lemeshow, S. & Saulnier, F., 1993. A New Simplified Acute Physiology Score (SAPS II) based on a European/North American Multicenter

Study. *JAMA*, 270(24), pp.2957–2963.

Le Gall, J.R. et al., 1996a. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. *JAMA*, 276(10), pp.802–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17166357.

Le Gall, J.R. et al., 1996b. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. *JAMA*, 276(10), pp.802–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17166357.

Martin, G.S. et al., 2003. The epidemiology of sepsis in the united states from 1979 through 2000. *New England Journal of Medicine*, 348(16), pp.1546–1554.

Mehta, R.L. et al., 2007. Acute kidney injury network: Report of an initiative to improve outcomes in acute kidney injury. *Critical care*, 11(2), p.R31.

Pérez, F. & Granger, B.E., 2007. IPython: A system for interactive scientific computing. *Computing in Science and Engineering*, 9(3), pp.21–29. Available at: http://ipython.org.

Quan, H. et al., 2005. Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Medical care*, pp.1130–1139.

Seymour, C.W. et al., 2016. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8), pp.762–774.

Singer, M. et al., 2016. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8), pp.801–810.

Steiner, C., Elixhauser, A. & Schnaier, J., 2001. The healthcare cost and utilization project: An overview. *Effective clinical practice: ECP*, 5(3), pp.143–151.

Torio CM, A.R., 2013. National inpatient hospital costs: The most expensive conditions by payer, 2011. hcup statistical brief #160. *Agency for Healthcare Research and Quality, Rockville, MD*. Available at: http://www.hcup-us.ahrq.gov/reports/statbriefs/sb160.pdf.

Torio CM, M.B., 2016. National inpatient hospital costs: The most expensive conditions by payer, 2013. hcup statistical brief #204. *Agency for Healthcare Research and Quality, Rockville, MD*. Available at: http://www.hcup-us.

ahrq.gov/reports/statbriefs/sb204-Most-Expensive-Hospital-Conditions.pdf.

Vincent, J.-L. et al., 1996. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22, pp.707–710.

Walraven, C. van et al., 2009. A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. *Medical care*, pp.626–633.

Wiesner, R. et al., 2003. Model for end-stage liver disease (meld) and allocation of donor livers. *Gastroenterology*, 124(1), pp.91–96.