Reproducibility in critical care:  
a mortality prediction case study

Alistair E. W. Johnson  
Institute of Medical Engineering & Science  
Massachusetts Institute of Technology  
Cambridge, MA, USA  
AEWJ@MIT.EDU

Tom J. Pollard  
Institute of Medical Engineering & Science  
Massachusetts Institute of Technology  
Cambridge, MA, USA  
TPOLLARD@MIT.EDU

Roger G. Mark  
Institute of Medical Engineering & Science  
Massachusetts Institute of Technology  
Cambridge, MA, USA  
RGMARK@MIT.EDU

Abstract

Mortality prediction of intensive care unit (ICU) patients facilitates hospital benchmarking and has the opportunity to provide caregivers with useful summaries of patient health at the bedside. The development of novel models for mortality prediction is a popular task in machine learning, with researchers typically seeking to maximize measures such as the area under the receiver operator characteristic curve (AUROC). The number of ‘researcher degrees of freedom’ that contribute to the performance of a model, however, presents a challenge when seeking to compare reported performance of such models.

In this study, we review publications that have reported performance of mortality prediction models based on the Medical Information Mart for Intensive Care (MIMIC) database and attempt to reproduce the cohorts used in their studies. We then compare the performance reported in the studies against gradient boosting and logistic regression models using a simple set of features extracted from MIMIC. We demonstrate the large heterogeneity in studies that purport to conduct the single task of ‘mortality prediction’, highlighting the need for improvements in the way that prediction tasks are reported to enable fairer comparison between models.

We reproduced datasets for 38 experiments corresponding to 28 published studies using MIMIC. In half of the experiments, the sample size we acquired was 25% greater or smaller than the sample size reported. The highest discrepancy was 11,767 patients. While accurate reproduction of each study cannot be guaranteed, we believe that these results highlight the need for more consistent reporting of model design and methodology to allow performance improvements to be compared. We discuss the challenges in reproducing the cohorts used in the studies, highlighting the importance of clearly reported methods (e.g. data cleansing, variable selection, cohort selection) and the need for open code and publicly available benchmarks.

©2017.
1. Introduction

Intensive care units (ICUs) provide support to the most severely ill patients in a hospital, offering radical life saving treatments. Patients are monitored closely within the ICU to assist in the early detection and correction of deterioration before it becomes fatal, an approach has been demonstrated to improve outcomes (Kane et al., 2007). Quantifying patient health and predicting future outcomes is an important area of critical care research. One of the most immediately relevant outcomes to the ICU is patient mortality, leading many studies toward development of mortality prediction models. Typically researchers seek to improve on previously published measures of performance such as sensitivity and specificity, but other goals may include improved model interpretability and novel feature extraction.

Recent advances in both machine learning and hospital networking have facilitated better prediction models using more detailed granular data. Interpreting studies that report advances in mortality prediction performance, however, is often a challenge, because like-for-like comparison is prevented by the high degree of heterogeneity amongst studies. For example, approaches may differ in areas such as exclusion criteria, data cleaning, creation of training and test sets, and so on, making it unclear where performance improvements have been gained.

In many areas of machine learning, datasets such as ImageNet (Deng et al., 2009) have facilitated benchmarking and comparison between studies. Key to these datasets is that they are publicly available to researchers, allowing code and data to be shared together to create reproducible studies. Barriers to data sharing in healthcare have limited the accessibility of highly granular clinical data and largely prevented publication of reproducible studies, but with freely-available datasets such as the Medical Information Mart for Intensive Care (MIMIC-III) end-to-end reproducible studies are attainable (Johnson et al., 2016). The use of mortality prediction models to evaluate ICUs as a whole has found great success, both for identifying useful policies and comparing patient populations. In order to focus contributions to the state of the art in mortality prediction, however, it should be clear where performance is being gained and further gains might be achieved.

In this study, we review publications that have reported performance of mortality prediction models based on the Medical Information Mart for Intensive Care (MIMIC) database and attempt to reproduce their studies. We then compare the performance reported in the studies against gradient boosting and logistic regression models using features extracted from MIMIC. The goal of this exercise is twofold: the primary hypothesis is that textual description of patient selection criteria are insufficient to reproduce studies; the secondary hypothesis is that data extraction using domain knowledge remains an often overlooked but useful tool to improve model performance.

2. Methods

2.1 Data

We use the MIMIC-III (v1.4) database (Johnson et al., 2016). MIMIC-III is a large, publicly available dataset of ICU admissions at the Beth Israel Deaconess Medical Center in Boston, MA. MIMIC-III has over 50,000 patient admissions and is the source of data for all studies.
evaluated here. MIMIC-II is a prior version of the database and a subset of MIMIC-III: patients in MIMIC-II are also contained in MIMIC-III and can be identified in the dataset.

2.2 Study selection

We reviewed all publications between January 2015 and March 2017 which referenced the papers describing MIMIC-II and MIMIC-III. Of these, we identified all studies which presented results on a mortality prediction task. We examined these studies and added any references which also presented results on a mortality prediction task. Finally, we excluded studies if (i) they incorporated waveform data, (ii) they did not report on the AUROC, or (iii) their exclusion criteria could not be reproduced without obtaining additional information from the author. As our study is not intended to be an exhaustive review of the literature, we did not attempt to include every recent study on mortality prediction.

2.3 Cohort Selection

Each study assessed here presented distinct patient inclusion criteria. Our process for replicating this was as follows: we first defined a base set of four exclusion criteria, which we deemed fundamental for all studies. First, we removed non-adults, specifically those aged at ICU admission $\geq$ 15 years old. Neonatal patients were not the focus of this study (or any of those assessed). Second, we removed invalid admissions defined as: no charted observations, no measurements of heart rate, or an incomplete administrative recording of ICU admission and discharge. Many of these stays correspond to clerical errors. Third, we removed organ donor accounts, which are often recorded as "readmissions" for administrative purposes. Lastly, we removed stays less than 4 hours. These stays correspond to situations for which an ICU mortality prediction system would be of little value (e.g. surgical preparation).

We reviewed each study and identified all respective inclusion criteria. After extracting cohorts, we compared the sample size we extracted and that reported in the original study. In some cases, we inferred that additional inclusion criteria were implied but not stated (most frequently this was a minimum amount of time in the ICU). While some studies were originally performed in MIMIC-II, all extractions were done in MIMIC-III as it is a superset of MIMIC-II.

2.4 Data Extraction

We extracted the same features for all possible windows, which varied from study to study. For example, the baseline cohort window began at ICU admission and ended up to 24 hours after ICU admission. For vital sign measurements (heart rate, blood pressure, respiratory rate, oxygen saturation), we extracted the first, last, minimum, and maximum value across the window. For laboratory measurements, we extended the window backwards by 24 hours and extracted the first and last measurement. We extended the window in order to improve

---

1. Note that MIMIC-III v1.4 does not contain pediatric patients.
2. Since laboratory values are available outside the ICU in MIMIC-III, it is possible to extend windows before ICU admission
data completion as laboratory measurements are infrequently sampled. More detail on the features can be found in the Appendix (Table 5).

2.5 Evaluation

We built mortality prediction models using gradient boosting (GB) as implemented in \texttt{xgboost} v0.6 (Chen and Guestrin, 2016) and logistic regression (LR) as implemented in \texttt{scikit-learn} v0.18 (Pedregosa et al., 2011). The target for prediction was defined by the study, and was one of the following: in-hospital mortality, 30-day post ICU admission mortality, 48-hour post ICU discharge mortality, 30-day post ICU discharge mortality, 30-day post hospital discharge mortality, 6-month post hospital discharge mortality, 1-year post hospital discharge mortality, and 2 year post hospital discharge mortality. We use 5-fold cross-validation to obtain estimates of model performance. When a patient had multiple stays in the dataset, we ensured that stays were grouped in the same fold. We did not attempt to optimize hyperparameters.

All comparisons use the area under the receiver operator characteristic curve (AUROC). We evaluate the AUROC of classifiers trained using replication datasets for each study, and compare this AUROC to that reported by each study. We also compare sample size and frequency of outcome. To ensure reproducibility of our analysis, we have made all the code openly available\(^3\) (Johnson, 2017).

3. Results

3.1 Study selection

We identified 328 studies which used the MIMIC dataset, of which 27 reported on the development of a mortality prediction model. An additional nine studies were identified from references. We excluded six studies that used waveforms or that did not report AUROC. Finally, we removed two studies which had complex exclusion criteria that could not be reproduced \(^4\). Our final selection included 28 published studies. Inclusion criteria for the studies which used in-hospital mortality as the outcome of interest are shown in Table 1. Inclusion criteria for the studies with outcomes of interest other than in-hospital mortality are shown in Table 2. Together, these studies reported on a total of 38 distinct experiments which varied the time window, outcome definition, or inclusion criteria. All studies extracted data from a fixed window centered on ICU admission unless otherwise noted. A brief description of the inclusion criteria is also noted: for further detail the reader is referred to the original publication.

3.2 Comparison to other studies

Table 3 compares the sample size, mortality rate, and the AUROCs presented by original studies with our reproduction for experiments where the outcome was in-hospital mortality. Table 4 shows the results comparison for other outcomes.

\(^3\) https://github.com/alistairewj/reproducibility-mimic

\(^4\) For example, one study required identification of patients with acute hypoxemic respiratory failure (AHRF), a diagnosis which would require free-text processing, specifically identifying chest radiograph reports for mention of bilateral infiltrates (Khemani et al., 2009; Purushotham et al., 2017).
Table 1: Inclusion criteria for each study that used in-hospital mortality as the outcome of interest.

*Window start time defined as 17 hours before ICU discharge or death. The PhysioNet 2012 Challenge dataset is a subset of MIMIC-II (Silva et al., 2012).

<table>
<thead>
<tr>
<th>Study</th>
<th>Window, $W$ (hours)</th>
<th>Inclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caballero Barajas and Akella (2015)</td>
<td>24</td>
<td>Age &gt; 18, Random fixed size subsample</td>
</tr>
<tr>
<td>Caballero Barajas and Akella (2015)</td>
<td>48</td>
<td>Age &gt; 18, Random fixed size subsample</td>
</tr>
<tr>
<td>Caballero Barajas and Akella (2015)</td>
<td>72</td>
<td>Age &gt; 18, Random fixed size subsample</td>
</tr>
<tr>
<td>Calvert et al. (2016b)</td>
<td>5*</td>
<td>Age &gt; 18, In MICU, &gt; 1 obs. for all features, LOS ≥ 17 hr, ICD-9 codes indicating alcohol withdrawal</td>
</tr>
<tr>
<td>Calvert et al. (2016a)</td>
<td>5*</td>
<td>Age &gt; 18, In MICU, &gt; 1 obs. for all features, 500 hr ≥ LOS ≥ 17 hr</td>
</tr>
<tr>
<td>Celi et al. (2012)</td>
<td>72</td>
<td>ICD-9 code 584.9</td>
</tr>
<tr>
<td>Celi et al. (2012)</td>
<td>24</td>
<td>ICD-9 code 430 or 852</td>
</tr>
<tr>
<td>Che et al. (2016) (b)</td>
<td>48</td>
<td>PhysioNet 2012 Challenge dataset</td>
</tr>
<tr>
<td>Ding et al. (2016)</td>
<td>48</td>
<td>PhysioNet 2012 Challenge dataset</td>
</tr>
<tr>
<td>Ghassemi et al. (2014)</td>
<td>12</td>
<td>Age &gt; 18, &gt; 100 words across all notes</td>
</tr>
<tr>
<td>Ghassemi et al. (2014)</td>
<td>24</td>
<td>Age &gt; 18, &gt; 100 words across all notes</td>
</tr>
<tr>
<td>Ghassemi et al. (2015)</td>
<td>24</td>
<td>Age &gt; 18, &gt; 100 words across all notes, &gt; 6 notes</td>
</tr>
<tr>
<td>Grnarova et al. (2016)</td>
<td>Entire stay</td>
<td>Age &gt; 18, stays with only one hospital admission</td>
</tr>
<tr>
<td>Harutyunyan et al. (2017)</td>
<td>48</td>
<td>Age &gt; 18, only one ICU stay during the hospital admission</td>
</tr>
<tr>
<td>Hoogendoorn et al. (2016)</td>
<td>24</td>
<td>&gt; 18, 1 obs. for BUN/Hct/GCS/HR/IV medication, LOS ≥ 24 hr</td>
</tr>
<tr>
<td>Johnson et al. (2012)</td>
<td>48</td>
<td>PhysioNet 2012 Challenge dataset</td>
</tr>
<tr>
<td>Johnson et al. (2014)</td>
<td>48</td>
<td>PhysioNet 2012 Challenge dataset</td>
</tr>
<tr>
<td>Lee and Maslove (2017)</td>
<td>24</td>
<td>Not missing data</td>
</tr>
<tr>
<td>Lehman et al. (2012)</td>
<td>24</td>
<td>Have SAPS-I, LOS ≥ 24 hr, first ICU stay only</td>
</tr>
<tr>
<td>Pirracchio et al. (2015)</td>
<td>24</td>
<td>Age &gt; 15</td>
</tr>
<tr>
<td>Ripoll et al. (2014)</td>
<td>24</td>
<td>No missing data, only septic patients</td>
</tr>
</tbody>
</table>
Table 2: Inclusion criteria for studies with outcomes of interest other than in-hospital mortality. * Window start time defined as 12 hours after ICU admission. 1–2 Post ICU discharge mortality: 1 48-hour, 2 30-day. 3–7 Post hospital discharge mortality: 3 30-day, 4 6-month, 5 1-year, 6 2-year.

<table>
<thead>
<tr>
<th>Study</th>
<th>Window, W (hours)</th>
<th>Inclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Che et al. (2016)¹ (a)</td>
<td>48</td>
<td>None described</td>
</tr>
<tr>
<td>Hug and Szolovits (2009)²</td>
<td>24</td>
<td>&gt;1 obs. for HR/GCS/Hct/BUN, Not NSICU/CSICU, first ICU stay, full code, no eventual brain death</td>
</tr>
<tr>
<td>Joshi et al. (2016)²</td>
<td>LOS ≥ 48hr</td>
<td>As in Hug and Szolovits (2009)</td>
</tr>
<tr>
<td>Luo et al. (2016)²</td>
<td>12**</td>
<td>Have a discharge summary, have SAPS-II</td>
</tr>
<tr>
<td>Luo and Rumshisky (2016)²</td>
<td>Entire stay</td>
<td>Age&gt;18, &gt;100 words across all notes</td>
</tr>
<tr>
<td>Ghassemi et al. (2014)³</td>
<td>12</td>
<td>Age&gt;18, stays with only one hospital admission</td>
</tr>
<tr>
<td>Grnarova et al. (2016)³</td>
<td>Entire stay</td>
<td>Age ≥ 65, Alive at hospital discharge</td>
</tr>
<tr>
<td>Lee et al. (2015)³</td>
<td>24</td>
<td>Only ICU stays with complete SAPS data</td>
</tr>
<tr>
<td>Lee and Maslove (2017)³</td>
<td>24</td>
<td>Only ICU stays with complete SAPS data</td>
</tr>
<tr>
<td>Lee (2017)³</td>
<td>24</td>
<td>Only ICU stays with complete SAPS data</td>
</tr>
<tr>
<td>Wojtusiak et al. (2017)³</td>
<td>Entire stay</td>
<td>Age&gt;18, &gt;100 words across all notes</td>
</tr>
<tr>
<td>Luo and Rumshisky (2016)⁴</td>
<td>Entire stay</td>
<td>Age&gt;18, &gt;100 words across all notes, &gt;6 notes</td>
</tr>
<tr>
<td>Ghassemi et al. (2014)⁵</td>
<td>12</td>
<td>Age&gt;18, stays with only one hospital admission</td>
</tr>
<tr>
<td>Ghassemi et al. (2015)⁵</td>
<td>24</td>
<td>Only ICU stays with complete SAPS data</td>
</tr>
<tr>
<td>Grnarova et al. (2016)⁶</td>
<td>Entire stay</td>
<td>Age&gt;18, stays with only one hospital admission</td>
</tr>
<tr>
<td>Lee and Maslove (2017)⁶</td>
<td>24</td>
<td>Only ICU stays with complete SAPS data</td>
</tr>
</tbody>
</table>
Table 3: Comparison of results shown from original studies ("study") and the reproduction here ("repro."). While the model used in studies varied, we grouped them as either linear (Lin) or non-linear (NonLin). We evaluated two models: a linear model (logistic regression, LR) and a non-linear model (gradient boosting, GB). The outcome was in-hospital mortality for all studies.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Sample size</th>
<th>Outcome (%)</th>
<th>AUROC GB</th>
<th>AUROC LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caballero Barajas and Akella (2015), W=24</td>
<td>11,648</td>
<td>11,648</td>
<td>-</td>
<td>13.01</td>
</tr>
<tr>
<td>Caballero Barajas and Akella (2015), W=48</td>
<td>11,648</td>
<td>11,648</td>
<td>-</td>
<td>13.01</td>
</tr>
<tr>
<td>Caballero Barajas and Akella (2015), W=72</td>
<td>11,648</td>
<td>11,648</td>
<td>-</td>
<td>13.01</td>
</tr>
<tr>
<td>Calvert et al. (2016b)</td>
<td>3,054</td>
<td>1,985</td>
<td>12.84</td>
<td>13.8</td>
</tr>
<tr>
<td>Calvert et al. (2016a)</td>
<td>9,683</td>
<td>18,396</td>
<td>10.68</td>
<td>14.71</td>
</tr>
<tr>
<td>Celi et al. (2012), AKI</td>
<td>1,400</td>
<td>4,741</td>
<td>30.7</td>
<td>23.92</td>
</tr>
<tr>
<td>Celi et al. (2012), SAH</td>
<td>223</td>
<td>350</td>
<td>25.6</td>
<td>24.86</td>
</tr>
<tr>
<td>Che et al. (2016) (b)</td>
<td>4,000</td>
<td>4,000</td>
<td>13.85</td>
<td>14.35</td>
</tr>
<tr>
<td>Ding et al. (2016)</td>
<td>4,000</td>
<td>4,000</td>
<td>13.85</td>
<td>14.35</td>
</tr>
<tr>
<td>Ghassemi et al. (2014), W=12</td>
<td>19,308</td>
<td>28,172</td>
<td>10.84</td>
<td>12.2</td>
</tr>
<tr>
<td>Ghassemi et al. (2014), W=24</td>
<td>19,308</td>
<td>23,442</td>
<td>10.80</td>
<td>12.92</td>
</tr>
<tr>
<td>Ghassemi et al. (2015)</td>
<td>10,202</td>
<td>21,969</td>
<td>-</td>
<td>13.51</td>
</tr>
<tr>
<td>Grnarova et al. (2016)</td>
<td>31,244</td>
<td>29,572</td>
<td>13.82</td>
<td>12.49</td>
</tr>
<tr>
<td>Harutyunyan et al. (2017)</td>
<td>42,276</td>
<td>45,493</td>
<td>-</td>
<td>10.54</td>
</tr>
<tr>
<td>Hoogendoorn et al. (2016)</td>
<td>13,923</td>
<td>17,543</td>
<td>-</td>
<td>14.97</td>
</tr>
<tr>
<td>Johnson et al. (2012)</td>
<td>4,000</td>
<td>4,000</td>
<td>-</td>
<td>14.35</td>
</tr>
<tr>
<td>Johnson et al. (2014)</td>
<td>4,000</td>
<td>4,000</td>
<td>-</td>
<td>14.35</td>
</tr>
<tr>
<td>Joshi and Szolovits (2012)</td>
<td>10,666</td>
<td>10,696</td>
<td>12.0</td>
<td>4.14</td>
</tr>
<tr>
<td>Lee and Maslove (2017)</td>
<td>17,490</td>
<td>20,961</td>
<td>17.73</td>
<td>17.86</td>
</tr>
<tr>
<td>Lehman et al. (2012)</td>
<td>14,739</td>
<td>21,738</td>
<td>14.6</td>
<td>12.32</td>
</tr>
<tr>
<td>Pirracchio et al. (2015)</td>
<td>24,508</td>
<td>28,795</td>
<td>12.2</td>
<td>12.72</td>
</tr>
<tr>
<td>Ripoll et al. (2014)</td>
<td>2,002</td>
<td>2,251</td>
<td>21.10</td>
<td>39.63</td>
</tr>
</tbody>
</table>
Table 4: Comparison of results shown from original studies (“study”) and the reproduction here (“repro.”). While the model used in studies varied, we grouped them as either linear (Lin) or non-linear (NonLin). We evaluated two models: linear model (logistic regression, LR) and a non-linear model (gradient boosting, GB).

- Post ICU discharge mortality: 1 48-hour, 2 30-day.
- Post hospital discharge mortality: 3 30-day, 4 6-month, 5 1-year, 6 2-year.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Sample size</th>
<th>Outcome (%)</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Study</td>
<td>Repro.</td>
<td>Study</td>
</tr>
<tr>
<td>Che et al. (2016) (a)</td>
<td>19,714</td>
<td>26,508</td>
<td>8.7</td>
</tr>
<tr>
<td>Hug and Szolovits (2009)</td>
<td>10,066</td>
<td>10,696</td>
<td>17.0</td>
</tr>
<tr>
<td>Luo et al. (2016)</td>
<td>7,863</td>
<td>8,931</td>
<td>17.0</td>
</tr>
<tr>
<td>Joshi et al. (2016)</td>
<td>17,000</td>
<td>26,508</td>
<td>-</td>
</tr>
<tr>
<td>Ghassemi et al. (2014)</td>
<td>19,308</td>
<td>28,172</td>
<td>3.23</td>
</tr>
<tr>
<td>Ghassemi et al. (2014), W=12</td>
<td>31,244</td>
<td>29,572</td>
<td>3.70</td>
</tr>
<tr>
<td>Lee et al. (2015)</td>
<td>17,490</td>
<td>20,961</td>
<td>15.1</td>
</tr>
<tr>
<td>Lee and Maslove (2017)</td>
<td>17,490</td>
<td>20,961</td>
<td>23.56</td>
</tr>
<tr>
<td>Lee (2017)</td>
<td>17,152</td>
<td>23,443</td>
<td>15.1</td>
</tr>
<tr>
<td>Luo and Rumshisky (2016)</td>
<td>18,412</td>
<td>27,747</td>
<td>3.4</td>
</tr>
<tr>
<td>Wojtusiak et al. (2017)</td>
<td>21,651</td>
<td>22,699</td>
<td>7.74</td>
</tr>
<tr>
<td>Luo and Rumshisky (2016)</td>
<td>18,412</td>
<td>27,747</td>
<td>9.5</td>
</tr>
<tr>
<td>Ghassemi et al. (2014), W=12</td>
<td>19,308</td>
<td>28,172</td>
<td>3.34</td>
</tr>
<tr>
<td>Ghassemi et al. (2015)</td>
<td>10,202</td>
<td>21,969</td>
<td>-</td>
</tr>
<tr>
<td>Ghassemi et al. (2016)</td>
<td>31,244</td>
<td>29,572</td>
<td>12.06</td>
</tr>
<tr>
<td>Lee and Maslove (2017)</td>
<td>17,152</td>
<td>23,443</td>
<td>43.82</td>
</tr>
</tbody>
</table>
4. Discussion

We attempted to reproduce the datasets for 38 experiments from 28 published studies that used MIMIC-II or MIMIC-III for mortality prediction. Due to the limited detail provided in the majority of papers, the heterogeneity in reporting style, and the lack of code sharing, this task was a challenge. As summarized in Tables 3 and 4, many of the datasets reported in the original papers differed in sample size from our reproduced datasets: our extraction usually resulted in a larger cohort. Similarly, also reported in Tables 3 and 4, we found the proportion of patients who died to vary widely between the reported and reproduced datasets. Given that we attempted to reproduce the original dataset using the same source of data, this wide variation should not occur. The exact reason for differences are difficult to establish without in-depth analysis or engaging with each of the study authors. However, we have noted a few cases where differences are clear. Studies on 1-year mortality by (Grnarova et al., 2016; Ghassemi et al., 2014; Luo and Rumshisky, 2016) all report hospital mortalities at least 10% lower than found in the reproduction datasets; likely explained by an exclusion of patients who die during their hospital admission. While this criteria was not explicitly stated to our knowledge, it may be “obvious” and would explain the mismatch in mortality rates.

Many of the studies reviewed omitted details necessary to fully reproduce the work: the minimum length of stay required for a patient to be included, which age to use for identifying adults, or whether readmissions to the ICU should be excluded. Most publications limit space in some way and, as a result, methodology is often forced to be described sparingly. One of the most faithfully reproduced cohorts was that of Hug and Szolovits (2009) (10,066 vs. 10,696 reproduced) as the cohort is described in a PhD thesis (Hug, 2009). In lieu of a thesis with full detail, studies should at a minimum describe any constraints on the population (age restrictions, length of stay restrictions), data completeness requirements, and how multiple stays for a single subject are treated. Furthermore, explicit technical description of a criteria was extremely useful in reproducing that criteria. For example, instead of stating “excluded patients missing data”, stating “included patients with at least 1 heart rate observation” was much more useful. Other examples are more subtle: while some studies stated they only included medical ICU patients, it is unclear whether this was defined using the physical location of the patient or the service the patient was admitted under. This distinction exists as a subset of patients were physically located in a unit which is not associated with the service of care they received. Examples such as this one are numerous when working with medical data. As a result, even with extremely detailed exclusion criteria, exact reproduction of a study may still be difficult. This difficulty may be further exasperated by discrepancies between the implementation of the exclusion criteria and the stated criteria due to a number of reasons such as technical error, sparse wording, or imprecise terminology. We would argue that openly available code is the simplest and most effective manner of ensuring exact reproduction of a study. It is worth noting that only 3 of the 28 papers included in this study had code openly available.

Tables 3 and 4 also display wide inter-study heterogeneity in cohort sizes, model performance, and outcome frequency. To some degree this is expected: certain studies focused on specific patient groups (Celi et al., 2012; Calvert et al., 2016b), while others required clinical notes (Ghassemi et al., 2014, 2015 Lehman et al., 2012). However, in Tables 1 and 2, it is
evident that large discrepancies in sample size exist even among studies of a similar cohort. Hoogendoorn et al. (2016) and Calvert et al. (2016a) both have similar inclusion criteria (age over 18, minimum stay of 17-24 hours), but the small differences in criteria result in a sample size difference of almost 4,000 (a difference which was smaller, but still significant, in our reproduction). This highlights a unique challenge in retrospective analysis of databases such as MIMIC-III. While controlled clinical trials require prior specification of measured parameters in great detail, research using observational data necessitates a data extraction step, and this step has marked effect on the resulting analysis and interpretation. When we compared the performance of a logistic regression with the best model performance reported for each individual study, we found logistic regression was equivalent or better in 64% of cases. Similarly, gradient boosting was equivalent or better in 82% of cases. While direct comparison is confounded by the difficulty of reproduction discussed earlier, we believe this highlights the importance of the data abstraction step which is often overshadowed by a description of modeling methodology. The establishment of benchmark datasets, such as those proposed by Silva et al. (2012) and Harutyunyan et al. (2017), or the use of a common set of open source abstractions, such as those described by Johnson et al. (2017), are key steps to addressing this issue.

Our study has limitations. First, there are slight differences in the data contained in MIMIC-II and the data we used in our study (MIMIC-III), which while minor, prohibit exact reproduction of MIMIC-II studies using MIMIC-III (see Appendix C). Second, the aim of many of the studies presented here was not mortality prediction. Many studies attempted to create patient phenotypes or summarize patient state in meaningful way, and only used mortality prediction as a "sanity check" on the model. While this does not impact our comments regarding reproducibility, it does limit the extent to which we can claim data abstraction is critical for model performance. Finally, we did not attempt to contact any authors of the publications. While certainly this would have improved our ability to reproduce their study, our aim was to demonstrate the difficulty in reproducing these studies from the publication alone.

5. Conclusion

We attempted to reproduce the patient cohorts for 28 studies that predicted mortality using the freely-available MIMIC-III database. Our results demonstrate that, in spite of best efforts, reproducing cohorts using textual descriptions of patient selection criteria is difficult. Detailed technical description of data abstraction is crucial to contextualize prior work. More than this, we believe that the public dissemination of open source code is central to facilitating iterative improvement in the field.

Acknowledgments

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


Appendix A. - Study Flow diagram

Figure 1 shows a flow diagram of the literature review.

Figure 1: Study identification and exclusion flow diagram.
Appendix B

Table 5 lists all variables and features extracted. These features were chosen to capture patient physiology and exclude explicit treatment data, though it is worth noting that some measurements will act as surrogates for treatment (e.g. the partial pressure of oxygen to fraction of inspired oxygen ratio is usually present only if the patient is treated with mechanical ventilation).

Table 5: Features extracted during the window examined. $t_{i,w}$ represents the end of the window $w$ for each patient $i$. $W$ represents the length of the window. The window is extended backward by 24 hours for laboratory and blood gas measurements. Some variables are repeated as the source of measurement differs (e.g. fingerstick glucose vs. laboratory obtained glucose). *All these features are extracted from arterial blood gases.

<table>
<thead>
<tr>
<th>Time window</th>
<th>Feature extracted</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[t_{i,w} - W, t_{i,w}]$</td>
<td>Minimum, Maximum, First, Last</td>
<td>Heart rate, Systolic/Diastolic/mean blood pressure, Respiratory rate, Temperature, Peripheral Oxygen Saturation, Glucose</td>
</tr>
<tr>
<td>$[t_{i,w} - W, t_{i,w}]$</td>
<td>Minimum</td>
<td>Glasgow coma scale</td>
</tr>
<tr>
<td>$[t_{i,w} - W, t_{i,w}]$</td>
<td>Last</td>
<td>Glasgow coma scale, Glasgow coma scale components (motor, verbal, eyes), unable to collect verbal score</td>
</tr>
<tr>
<td>$[t_{i,w} - W - 24, t_{i,w}]$</td>
<td>First, last</td>
<td>Oxygen saturation, Partial pressure of oxygen, Partial pressure of carbon dioxide, Arterial-alveolar gradient, Ratio of partial pressure of oxygen to fraction of oxygen inspired, pH, Base excess, Bicarbonate, Total carbon dioxide concentration, Hematocrit, Hemoglobin, Carboxyhemoglobin, Methemoglobin, Chloride, Calcium, Temperature, Potassium, Sodium, Lactate, Glucose</td>
</tr>
<tr>
<td>$[t_{i,w} - W - 24, t_{i,w}]$</td>
<td>First, last</td>
<td>Anion gap, Albumin, Immature band forms, Bicarbonate, Bilirubin, Creatinine, Chloride, Glucose, Hematocrit, Hemoglobin, Lactate, Platelet, Potassium, Partial thromboplastin time, International Normalized Ratio, Prothrombin time, Sodium, Blood urea nitrogen, White blood cell count</td>
</tr>
<tr>
<td>$[t_{i,w} - W - 24, t_{i,w}]$</td>
<td>Sum</td>
<td>Urine output</td>
</tr>
</tbody>
</table>
Appendix C - MIMIC-II vs. MIMIC-III

MIMIC-II was contains data for all critical care admissions between 2001-2008 at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA, USA. MIMIC-III is an extension of MIMIC-II, containing admissions from an additional four years (2008-2012). As the BIDMC changed their clinical information system across their ICUs during 2008, it is relatively straightforward to identify patients in MIMIC-III who comprise MIMIC-II by isolating to the database source "carevue". Nevertheless, there are two major differences between MIMIC-II and MIMIC-III which may cause discrepancies when comparing cohorts extracted from the two systems.

First, MIMIC-III defined ICU admissions based on a hospital administrative database, whereas MIMIC-II utilized the ICU database. The hospital administrative database tracks patients hospital wide, and as a result the use of this database expands the scope of patient tracking from ICU specific to all floors in the hospital. However, this hospital wide database is not linked to the ICU clinical information system, and as such patients are tracked independently in the two systems. For the most part, this means that patient admission and discharge times slightly differ between MIMIC-II and MIMIC-III by no more than a few hours, but larger differences do occur. These differences can manifest via exclusion criteria which utilize length of stay.

Second, severity of illness scores (such as SAPS-I) were derived by the laboratory releasing the data and distributed with MIMIC-II. These scores were not similarly distributed in MIMIC-III. While code for deriving these scores is publicly available Johnson et al. (2017), this code was written separately to that written for MIMIC-II. As a result, the use of missing severity scores as an exclusion criteria may result in distinct patients being excluded.