Reproducibility in critical care: a mortality prediction case study

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

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

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

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.

1. Introduction

Intensive care units (ICUs) provides 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 ¡ 15 years old [1]. 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[2]. 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 *xgboost* v0.6 (Chen and Guestrin, 2016) and logistic regression (LR) as implemented in *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 (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 [3]. 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. 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).

Study	Window, W (hours)	Inclusion criteria
Caballero Barajas and Akella (2015)	24	Age>18, Random fixed size subsample
Caballero Barajas and Akella (2015)	48	Age>18, Random fixed size subsample
Caballero Barajas and Akella (2015)	72	Age>18, Random fixed size subsample
Calvert et al. (2016b)	5*	Age>18, In MICU, >1 obs. for all features, LOS $\geq$ 17hr, ICD-9 codes indicating alcohol withdrawal
Calvert et al. (2016a)	5*	Age>18, In MICU, >1 obs. for all features, 500hr $\geq$ LOS $\geq$ 17hr
Celi et al. (2012)	72	ICD-9 code 584.9
Celi et al. (2012)	24	ICD-9 code 430 or 852
Che et al. (2016) (b)	48	PhysioNet 2012 Challenge dataset
Ding et al. (2016)	48	PhysioNet 2012 Challenge dataset
Ghassemi et al. (2014)	12	Age>18, >100 words across all notes
Ghassemi et al. (2014)	24	Age>18, >100 words across all notes
Ghassemi et al. (2015)	24	Age>18, >100 words across all notes, >6 notes
Grnarova et al. (2016)	Entire stay	Age>18, stays with only one hospital admission
Harutyunyan et al. (2017)	48	Age>18, only one ICU stay during the hospital admission
Hoogendoorn et al. (2016)	24	>18, 1 obs. for BUN/Hct/GCS/HR/IV medication, LOS $\geq$ 24hr
Johnson et al. (2012)	48	PhysioNet 2012 Challenge dataset
Johnson et al. (2014)	48	PhysioNet 2012 Challenge dataset
Joshi and Szolovits (2012)	24	As in Hug and Szolovits (2009)
Lee and Maslove (2017)	24	Not missing data
Lehman et al. (2012)	24	Have SAPS-I, LOS $\geq$ 24hr, first ICU stay only
Pirracchio et al. (2015)	24	Age>15
Ripoll et al. (2014)	24	No missing data, only septic patients

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.

Study	Window, W (hours)	Inclusion criteria
Che et al. (2016)[1] (a)	48	None described
Hug and Szolovits (2009)[2]	24	>1 obs. for HR/GCS/Hct/BUN, Not NSICU/CSICU, first ICU stay, full code, no eventual brain death
Joshi et al. (2016)[2]	LOS $\geq$ 48hr	
Luo et al. (2016)[2]	12**	As in Hug and Szolovits (2009)
Luo and Rumshisky (2016)[2]	Entire stay	Have a discharge summary, have SAPS-II
Ghassemi et al. (2014)[3]	12	Age>18, >100 words across all notes
Grnarova et al. (2016)[3]	Entire stay	Age>18, stays with only one hospital admission
Lee et al. (2015)[3]	24	Only ICU stays with complete SAPS data
Lee and Maslove (2017)[3]	24	Only ICU stays with complete SAPS data
Lee (2017)[3]	24	Only ICU stays with complete SAPS data
Wojtusiak et al. (2017)[3]	Entire stay	Age $\geq$ 65, Alive at hospital discharge
Luo and Rumshisky (2016)[4]	Entire stay	Have a discharge summary, have SAPS-II
Ghassemi et al. (2014)[5]	12	Age>18, >100 words across all notes
Ghassemi et al. (2015)[5]	24	Age>18, >100 words across all notes, >6 notes
Grnarova et al. (2016)[5]	Entire stay	Age>18, stays with only one hospital admission
Lee and Maslove (2017)[6]	24	Only ICU stays with complete SAPS data

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.

	Sample size		Outcome (%)			AUROC		
Cohort	Study	Repro.	Study	Repro.	Model	Study	GB	LR
Caballero Barajas and Akella (2015), W=24	11,648	11,648	-	13.01	NonLin	0.8657	0.906	0.88616
Caballero Barajas and Akella (2015), W=48	11,648	11,648	-	13.01	NonLin	0.7985	0.9227	0.9034
Caballero Barajas and Akella (2015), W=72	11,648	11,648	-	13.01	NonLin	0.7385	0.9314	0.9144
Calvert et al. (2016b)	3,054	1,985	12.84	13.8	NonLin	0.934	0.9565	0.9025
Calvert et al. (2016a)	9,683	18,396	10.68	14.71	NonLin	0.88	0.9333	0.9110
Celi et al. (2012), AKI	1,400	4,741	30.7	23.92	NonLin	0.875	0.8812	0.8706
Celi et al. (2012), SAH	223	350	25.6	24.86	NonLin	0.958	0.8929	0.8289
Che et al. (2016) (b)	4,000	4,000	13.85	14.35	NonLin	0.8424	0.8461	0.8273
Ding et al. (2016)	4,000	4,000	13.85	14.35	NonLin	0.8177	0.8461	0.8273
Ghassemi et al. (2014), W=12	19,308	28,172	10.84	12.2	Lin	0.84	0.8846	0.8609
Ghassemi et al. (2014), W=24	19,308	23,442	10.80	12.92	Lin	0.841	0.8841	0.8651
Ghassemi et al. (2015)	10,202	21,969	-	13.51	NonLin	0.812	0.8781	0.8591
Grnarova et al. (2016)	31,244	29,572	13.82	12.49	NonLin	0.963	0.9819	0.9765
Harutyunyan et al. (2017)	42,276	45,493	-	10.54	NonLin	0.8625	0.9406	0.9286
Hoogendoorn et al. (2016)	13,923	17,545	-	14.97	NonLin	0.841	0.8797	0.8618
Johnson et al. (2012)	4,000	4,000	-	14.35	NonLin	0.8602	0.8461	0.8273
Johnson et al. (2014)	4,000	4,000	-	14.35	Lin	0.8457	0.8461	0.8273
Joshi and Szolovits (2012)	10,066	10,696	12.0	4.14	Lin	0.89	0.8872	0.8716
Lee and Maslove (2017)	17,490	20,961	17.73	17.86	Lin	0.775	0.8655	0.8488
Lehman et al. (2012)	14,739	21,738	14.6	12.32	Lin	0.82	0.888	0.8694
Pirracchio et al. (2015)	24,508	28,795	12.2	12.72	NonLin	0.88	0.9070	0.8897
Ripoll et al. (2014)	2,002	2,251	21.10	39.63	NonLin	0.8223	0.7900	0.7647

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). [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.

	Sample size		Outcome (%)			AUROC		
Cohort	Study	Repro.	Study	Repro.	Model	Study	GB	LR
Che et al. (2016)[1] (a)	19,714	26,508	8.7	19.0	NonLin	0.8527	0.8675	0.7500
Hug and Szolovits (2009)[2]	10,066	10,696	17.0	6.35	Lin	0.879	0.8545	0.8479
Luo et al. (2016)[2]	7,863	8,931	17.0	6.45	NonLin	0.848	0.8260	0.8249
Joshi et al. (2016)[2]	17,000	26,508	-	14.95	Lin	0.72	0.8759	0.8539
Ghassemi et al. (2014)[3], W=12	19,308	28,172	3.23	16.92	Lin	0.761	0.8687	0.8447
Grnarova et al. (2016)[3]	31,244	29,572	3.70	16.36	NonLin	0.858	0.9590	0.9528
Lee et al. (2015)[3]	17,490	20,961	15.1	12.69	Lin	0.784	0.8828	0.8643
Lee and Maslove (2017)[3]	17,490	20,961	23.56	31.57	Lin	0.762	0.8442	0.8240
Lee (2017)[3]	17,152	23,443	15.1	17.94	Lin	0.815	0.8674	0.8492
Luo and Rumshisky (2016)[3]	18,412	27,747	3.4	17.05	NonLin	0.86	0.9289	0.9193
Wojtusiak et al. (2017)[3]	21,651	22,699		7.74	NonLin	0.734	0.7957	0.7901
Luo and Rumshisky (2016)[4]	18,412	27,747	9.5	25.17	NonLin	0.842	0.8921	0.8770
Ghassemi et al. (2014)[5], W=12	19,308	28,172	3.34	29.75	Lin	0.743	0.8450	0.8201
Ghassemi et al. (2015)[5]	10,202	21,969	-	32.35	NonLin	0.686	0.8403	0.8194
Grnarova et al. (2016)[5]	31,244	29,572	12.06	24.63	NonLin	0.853	0.9117	0.9009
Lee and Maslove (2017)[6]	17,152	23,443	43.82	17.94	Lin	0.83	0.8674	0.8492

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

Karla L Caballero Barajas and Ram Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–78. ACM, 2015.

Jacob Calvert, Qingqing Mao, Jana L Hoffman, Melissa Jay, Thomas Desautels, Hamid Mohamadlou, Uli Chettipally, and Ritankar Das. Using electronic health record collected clinical variables to predict medical intensive care unit mortality. *Annals of Medicine and Surgery*, 11:52–57, 2016a.

Jacob Calvert, Qingqing Mao, Angela J Rogers, Christopher Barton, Melissa Jay, Thomas Desautels, Hamid Mohamadlou, Jasmine Jan, and Ritankar Das. A computational approach to mortality prediction of alcohol use disorder inpatients. *Computers in biology and medicine*, 75:74–79, 2016b.

Leo A Celi, Sean Galvin, Guido Davidzon, Joon Lee, Daniel Scott, and Roger Mark. A database-driven decision support system: customized mortality prediction. *Journal of personalized medicine*, 2(4):138–148, 2012.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865*, 2016.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

Yangyang Ding, Xuejian Li, and Youqing Wang. Mortality prediction for icu patients using just-in-time learning and extreme learning machine. In *Intelligent Control and Automation (WCICA), 2016 12th World Congress on*, pages 939–944. IEEE, 2016.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84. ACM, 2014.

Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2015, page 446. NIH Public Access, 2015.

Paulina Grnarova, Florian Schmidt, Stephanie L Hyland, and Carsten Eickhoff. Neural document embeddings for intensive care patient mortality prediction. *arXiv preprint arXiv:1612.00467*, 2016.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.

Mark Hoogendoorn, Ali el Hassouni, Kwongyen Mok, Marzyeh Ghassemi, and Peter Szolovits. Prediction using patient comparison vs. modeling: A case study for mortality prediction. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 2464–2467. IEEE, 2016.

Caleb W Hug. *Detecting hazardous intensive care patient episodes using real-time mortality models*. PhD thesis, 2009.

Caleb W Hug and Peter Szolovits. Icu acuity: real-time models versus daily models. In *AMIA*, 2009.

Alistair EW Johnson. alistairewj/reproducibility-mimic: Reproducibility in mimic v0.3.0. Jul 2017. doi: 10.5281/zenodo.842922.

Alistair EW Johnson, Nic Dunkley, Louis Mayaud, Athanasios Tsanas, Andrew A Kramer, and Gari D Clifford. Patient specific predictions in the intensive care unit using a bayesian ensemble. In *Computing in Cardiology (CinC), 2012*, pages 249–252. IEEE, 2012.

Alistair EW Johnson, Andrew A Kramer, and Gari D Clifford. Data preprocessing and mortality prediction: The physionet/cinc 2012 challenge revisited. In *Computing in Cardiology Conference (CinC), 2014*, pages 157–160. IEEE, 2014.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 35, 2016.

Alistair EW Johnson, Leo A Celi, David J Stone, and Tom J Pollard. The MIMIC code repository: Enabling reproducibility in critical care research. *JAMIA*, 2017. doi: 10.1093/jamia/ocx084.

Rohit Joshi and Peter Szolovits. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1276. American Medical Informatics Association, 2012.

Shalmali Joshi, Suriya Gunasekar, David Sontag, and Joydeep Ghosh. Identifiable phenotyping using constrained non–negative matrix factorization. *arXiv preprint arXiv:1608.00704*, 2016.

RL Kane, TA Shamliyan, C Mueller, S Duval, and T J Wilt. The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis. *Medical Care*, 45(12):1195–1204, December 2007.

Robinder G Khemani, David Conti, Todd A Alonzo, Robert D Bart, and Christopher JL Newth. Effect of tidal volume in children with acute hypoxemic respiratory failure. *Intensive care medicine*, 35(8):1428–1437, 2009.

Joon Lee. Patient-specific predictive modeling using random forests: An observational study for the critically ill. *JMIR Medical Informatics*, 5(1), 2017.

Joon Lee and David M Maslove. Customization of a severity of illness score using local electronic medical record data. *Journal of intensive care medicine*, 32(1):38–47, 2017.

Joon Lee, David M Maslove, and Joel A Dubin. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PloS one*, 10(5):e0127428, 2015.

Li-Wei H Lehman, Mohammed Saeed, William J Long, Joon Lee, and Roger G Mark. Risk stratification of icu patients using topic models inferred from unstructured progress notes. In *AMIA*. Citeseer, 2012.

Yen-Fu Luo and Anna Rumshisky. Interpretable topic features for post-icu mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 827. American Medical Informatics Association, 2016.

Yuan Luo, Yu Xin, Rohit Joshi, Leo Celi, and Peter Szolovits. Predicting icu mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements. In *AAAI*, pages 42–50, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Romain Pirracchio, Maya L Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J van der Laan. Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. *The Lancet Respiratory Medicine*, 3(1):42–52, 2015.

Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. Variational adversarial deep domain adaptation for health care time series analysis. *ICLR*, 2017.

Vicent J Ribas Ripoll, Alfredo Vellido, Enrique Romero, and Juan Carlos Ruiz-Rodríguez. Sepsis mortality prediction with the quotient basis kernel. *Artificial intelligence in medicine*, 61(1):45–52, 2014.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology (CinC), 2012*, pages 245–248. IEEE, 2012.

Janusz Wojtusiak, Eman Elashkar, and Reyhaneh Mogharab Nia. C-lace: Computational model to predict 30-day post-hospitalization mortality. *BIOSTEC 2017*, page 169, 2017.

Appendix A. - Study Flow diagram

Figure 1 shows a flow diagram of the literature review.

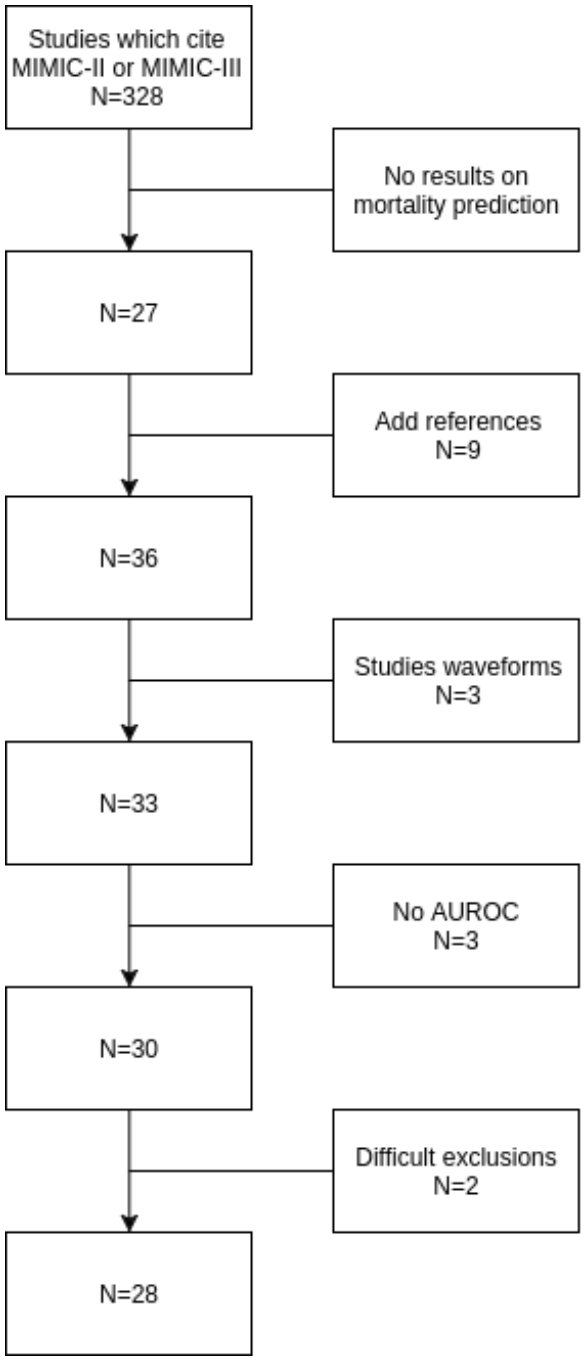

Figure 1: Study identification and exclusion flow diagram.

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.

Time window	Feature extracted	Variables
$[t_{i,w} - W, t_{i,w}]$	Minimum, Maximum, First, Last	Heart rate, Systolic/Diastolic/Mean blood pressure, Respiratory rate, Temperature, Peripheral Oxygen Saturation, Glucose
$[t_{i,w} - W, t_{i,w}]$	Minimum	Glasgow coma scale
$[t_{i,w} - W, t_{i,w}]$	Last	Glasgow coma scale, Glasgow coma scale components (motor, verbal, eyes), unable to collect verbal score
$[t_{i,w} - W - 24, t_{i,w}]$	First, last	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
$[t_{i,w} - W - 24, t_{i,w}]$	First, last	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
$[t_{i,w} - W - 24, t_{i,w}]$	Sum	Urine output

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.