1 Introduction

This study is a technical supplement to “AI gone astray: How subtle shifts in patient data send popular algorithms reeling, undermining patient safety.” from STAT News [1], which investigates the effect of time drift on clinically deployed machine learning models. We use MIMIC-IV, a publicly available dataset, to train models that replicate commercial approaches by Dascena and Epic to predict the onset of sepsis, a deadly and yet treatable condition. We observe some of these models degrade over time; most notably an RNN built on Epic features degrades from a 0.729 AUC to a 0.525 AUC over a decade, leading us to investigate technical and clinical drift as root causes of this performance drop.

2 Methods

Dataset  We investigate time drift using the MIMIC-IV database [1], which includes electronic health records of over 50,000 patients admitted to the intensive care units at the Beth Israel Deaconess Medical Center (BIDMC) between the years 2008-2019. We filter for patients over the age of 15, with an ICU stay between 24 hours and 10 days, and take each patient’s first ICU stay (see Figure 1). After all filtering, we end up with 50k patients in our dataset.
Sepsis is a potentially life-threatening condition where a body’s reaction to an infection triggers damage in organ systems. Following recent work \cite{2,3}, we predict the onset of Sepsis-3 within 6 hours given 24 hours of ICU data. Sepsis-3 defines sepsis onset as an increase in SOFA-score of ≥ 2 points within a window of 48 hours before and 24 hours after a Suspicion of Infection, which occurs when there are concomitant orders of antibiotics and microbiological samples taken. We follow the implementation of Moor et al. \cite{2}, and use the SOFA value in the first hour of a patient’s ICU stay as the baseline score for later comparisons.

**Sepsis Prediction Task**  

Figure 1: Cohort filtration process for MIMIC-IV

**Cohort Construction**  
See Figure 2 for our cohort construction procedure. For patients with a positive sepsis label, the first instance of sepsis onset was taken to be the sepsis onset time. Following prior work by the Dascena group, \cite{4}, we assign a random "onset" time for control patients with no sepsis onset during their stay. For both positive and control samples, we extract 24 hours of ICU data preceding a 6 hour gap period before their onset time. We eliminate all patients with an onset.

Figure 2: Top: data construction process for a positive sepsis patient, taking their first sepsis onset time during their ICU stay. Bottom: data construction process for a control patient with no sepsis onset during their ICU stay, assigning a random "onset" time for data extraction purposes.
time before 6 hours. For patients who have less than 24 hours of data after extraction, we left pad all non-existent hours with zeros.

**Modeling Commercial Tools** We model two commercial sepsis detection systems: Dascena’s and Epic Systems’. Their specific model implementations are not publicly available, so we use their known features to build our own models on MIMIC-IV and experiment with different model architectures. Dascena’s sepsis models include a combination of six vital signs – systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, peripheral oxygen saturation (SpO2), and temperature – and patient age. Epic’s sepsis models include more than 40 high level features, generally divided into 6 categories: Demographics, Vital Signs, Recent Lab Results, Chronic Illness Diagnoses, Medication Orders, and Active Drains, Airways or Wounds.

These models leveraged two types of patient features: static demographic features, and time-varying features. For the former, one value per feature is included for each patient. For the latter, we follow prior work and aggregate multiple irregularly-measured values into hourly buckets for each hour of a patient’s ICU stay. Vital sign measurements are aggregated by taking the average of all measurements in the hour, while events such as medication, infusions or drainages are aggregated by taking the sum across the hour. For each feature, we simply take their raw representation, as identified by each feature’s unique item ID from the MIMIC-IV dataset. To account for data missingness in time-varying features, we implement simple imputation, which involves forward-filling each feature for each patient by hour, then concatenating binary flags for whether a feature was originally missing and another value for the time it was last recorded. A detailed list of the features we used and their Item ID’s in MIMIC-IV can be found in Appendix A.

### 2.1 Model Details

To investigate the impact of model architecture on performance degradation, we experimented with both Logistic Regression models and Elman RNNs. Reference our code repository for specific model implementation and training details.

![Figure 3: Model architectures for Logistic Regression and RNN models](image)

**Logistic Regression** Our Logistic Regression baseline is illustrated in Figure 3. Following Nestor et al.’s work, we first flatten all hourly features, treating each measurement at each hour as a separate column. We then concatenate the result with the patient’s static features to form the overall feature vector for the patient. In the context of Dascena features, static features include age, and hourly
features include all vital sign measurements. In the context of Epic features, static features include
demographic features and diagnoses codes, and hourly features include all vital sign measurements,
medications, and lab measurements.

**RNN** To capture recurrent features, we implement an Elman RNN [7]. As depicted in Figure 3,
static features are first fed through a separate 4-layer MLP (of 32 neurons in each layer, preceded
by batch-normalization and succeeded by ReLU activation), while hourly features are fed directly
through the Elman RNN. Then, the output hidden representations from the MLP are concatenated
with the output hidden representations of the RNN before feeding everything through a final linear
layer.

3 Experiments

3.1 Time Drift Experiments

We seek to investigate how different models replicating commercial feature sets would age in clinical
settings. To understand why model performance may change over time, we investigate technical and
clinical sources of data shift.

**Year Agnostic** Similar to common model development practices, we report numbers from models
trained and tested on the full set of MIMIC-IV patient data from 2008-2019, with patients randomly
assigned to either training, validation or testing regardless of the year of care.

**Year Buckets** To measure temporal model drift, we train on a subset of patient data in the 2008-
2010 bucket, then test on patient data from the unseen portion of the 2008-2010 bucket as well as the
entirety of the data in the subsequent three year buckets 2011-2013, 2014-2016, and 2017-2019.

3.2 Investigating Cause of Time Drift

We investigate two causes of model drift: technical drift, and data drift. To investigate technical drift,
we analyzed the effect of Beth Israel’s switch in use of diagnoses codes from ICD-9 to ICD-10 in
October 2015. We investigate the impact of these changing codes on model performance by removing
all features from the Epic feature set that rely on ICD codes (HIV, obesity, coronary artery disease,
congestive heart failure, chronic obstructive pulmonary disease, chronic kidney disease, chronic liver
disease, diabetes, and hypertension), and retraining a new model with all other hyper-parameters kept
constant. To investigate data shift over time, we plotted sepsis onset times, microbiology samples,
and antibiotic usage trends over the years.

4 Results

We ran each combination of feature set, model, and training regime three times to measure the models’
performance. The average AUCs from the three runs are reported in Tables 1 and 2. Overall, we
observe a large drop in performance for models trained on the Epic feature set, especially the RNN.
The models trained on the Dascena model also experienced some performance degradation, but on a
much smaller scale.
4.1 Time Drift Experiments

|                      | Year-Agnostic | 2008-2010 | 2011-2013 | 2014-2016 | 2017-2019 |
|----------------------|---------------|-----------|-----------|-----------|-----------|
| **Dascena Features** |               |           |           |           |           |
| RNN                  | 0.764         | 0.723     | 0.726     | 0.730     | 0.707     |
| Logistic             | 0.731         | 0.687     | 0.696     | 0.682     | 0.640     |
| **Epic Features**    |               |           |           |           |           |
| RNN                  | 0.787         | 0.729     | 0.723     | 0.578     | 0.525     |
| Logistic             | 0.724         | 0.706     | 0.701     | 0.679     | 0.626     |

Table 1: Time agnostic and year bucket results (average test AUCs of three runs) from RNN and Logistic Regression models trained on the Epic and Dascena feature sets.

We observe in Table 1 that the models trained on Epic’s feature set dropped from 0.729 to 0.525 for the RNN model, and from 0.706 to 0.626 for the Logistic Regression model. This amounts to a 0.08 to 0.20 difference in test AUC over the years. These results indicate that several Epic features vary over time. In contrast, models built on Dascena features, which contain mainly vital signs, perform significantly better over time. We found a small temporal drop in AUC from 0.723 to 0.707 for the RNN model and a slightly larger AUC drop from 0.687 to 0.640 for the Logistic Regression model. In addition, we note that the choice in model architecture impacts performance drop, as we found that there is a much larger drop for the RNN model than the Logistic Regression model trained on Epic’s feature set, while the opposite was true for models trained on the Dascena feature set. We also include results for two additional time drift experiments, Length of Stay and ICU Mortality prediction, in Appendix B.

4.2 Time Drift Investigations

|                      | Year-Agnostic | 2008-2010 | 2011-2013 | 2014-2016 | 2017-2019 |
|----------------------|---------------|-----------|-----------|-----------|-----------|
| RNN                  | 0.783         | 0.716     | 0.731     | 0.721     | 0.671     |
| Logistic             | 0.669         | 0.715     | 0.702     | 0.692     | 0.633     |

Table 2: Average test AUCs from models trained on the Epic feature subset without ICD codes.

**Technical Change Investigation** In Table 2, we report the results of models trained on a subset of the Epic feature set without ICD codes. We found that without the ICD codes, the AUC drop across the entire time period is now 0.08 for the Logistic Regression model and 0.045 for the RNN. Compared to the RNN model trained on the full Epic feature set, this represents a significant improvement of around 0.15 for the last year bucket. Because models trained on the first year bucket only observed ICD-9 codes, they were unable to leverage ICD-10 features when the transition occurred.

We notice that the introduction of the ICD codes did not contribute to higher overall model performance, as seen by the year-agnostic numbers in Table 1 and Table 2. However, the inclusion of these features instead led to significant model degradation.
Figure 4: Sepsis onset time in the ICU by year bucket groups.

Figure 5: Microbiology samples across year buckets and time into ICU stay.

Figure 6: Antibiotics administered across year buckets and time into ICU stay.

Figure 7: Microbiological culture types with the largest drop in number of samples taken across year bucket groups.
Table 3: Specimen types with greatest absolute change between 2008-2010 and 2017-2019 year buckets. “Change” and “% Change” refers to the difference between the first and the last year buckets.

| Specimen Type                  | 2008-2010 | 2011-2013 | 2014-2016 | 2017-2019 | Change  | % Change |
|--------------------------------|-----------|-----------|-----------|-----------|---------|----------|
| SWAB                           | 92304     | 79169     | 68363     | 39677     | -52627  | -57%     |
| MRSA SCREEN                    | 39086     | 34375     | 14766     | 5657      | -33429  | -86%     |
| BLOOD CULTURE                  | 171630    | 158924    | 163711    | 144620    | -27010  | -16%     |
| SPUTUM                         | 52437     | 39027     | 40078     | 32135     | -20302  | -39%     |
| SEROLOGY/BLOOD                 | 35713     | 35489     | 32866     | 16187     | -19526  | -55%     |
| STOOL                          | 47316     | 38649     | 38166     | 28502     | -18814  | -40%     |
| IMMUNOLOGY                     | 13524     | 14336     | 11085     | 209       | -13315  | -98%     |
| URINE                          | 221335    | 211626    | 263363    | 234576    | +13241  | +6%      |
| BONE MARROW - CYTOGENETICS     | 7442      | 3643      | 5         | 0         | -7442   | -100%    |
| TISSUE                         | 21455     | 26311     | 32957     | 28481     | +7026   | +33%     |
| Immunology (CMV)               | 5921      | 5278      | 4285      | 42        | -5879   | -99%     |
| CATHETER OR LINE               | 6862      | 3653      | 2014      | 1397      | -5465   | -80%     |
| CSF:SPINAL FLUID               | 10725     | 8527      | 8601      | 6333      | -4392   | -41%     |
| BRONCHIAL WASHINGS             | 2214      | 2577      | 5087      | 6442      | +4228   | +191%    |
| Influenza A/B by DFA           | 4214      | 2771      | 532       | 0         | -4214   | -100%    |
| ABSCESS                        | 9597      | 11657     | 12734     | 13796     | +4199   | +44%     |
| Blood (LYME)                   | 0         | 4         | 2414      | 3815      | +3815   | +100%    |
| PLEURAL FLUID                  | 7560      | 7560      | 8834      | 9826      | +2266   | -30%     |
| BRONCHOALVEOLAR LAVAGE         | 12772     | 10930     | 13205     | 10523     | -2249   | -18%     |
| Staph aureus Screen            | 3609      | 9151      | 11949     | 5608      | 1999    | +55%     |

Figure 8: Amount of cultures drawn for each hour of the day, per year-bucket.

**Data Change Investigation** We plot all data changes in Figures 4-8 and Table 3. While we see no major changes in antibiotics administered, we see a large increase in average sepsis onset time from 9 hours to 12 hours by the last year bucket, accompanied by a drop in microbiology samples taken within the first 10 hours of a patient’s ICU stay. Figure 7 shows a select few of the microbiological sample types with the largest changes across the years, and Table 3 shows the full numbers in detail. Since the definition of Sepsis-3 onset depends on a suspicion of infection, which is influenced by the timing of microbiology samples, we observe that these two changes are related. Finally, Figure 8 shows that the drop in number of samples occurs mainly during daytime hours (between 9 AM and 5 PM).

Clinical sources we interviewed explained that changing hospital demographics, caused by multiple new hospital acquisitions and partnerships, in recent years could have been the cause of changing clinical practices around microbiology sampling and other ICU procedures.

5 Conclusion

By replicating commercially produced sepsis prediction models on the MIMIC-IV dataset, we were able to observe the extent of model degradation over time. We found a significant AUC drop of 0.08 to 0.20 for models trained on the Epic feature set, while a smaller drop of 0.02 to 0.05 was observed
for models trained on the more stable Dascena feature set. We concluded that the technical shift of changing ICD codes from ICD-9 to ICD-10 in 2015 was a major cause of degradation in the former models, as the performance drop diminished by 0.15 after the models were retrained without ICD codes. Finally, we observed a data shift in later years from changing microbiology sampling practices in the ICU, which had an impact on delaying sepsis onset times in the last year bucket.

**Limitations**  There are several limitations to this study. First, because the Epic and Dascena models are not publicly available, we trained our models leveraging their published feature sets on MIMIC-IV. These features can be found in Appendix A. Our results do not directly reflect the performance of Epic or Dascena’s commercial models. Second, this study does not account for possible model retraining procedures a hospital could undertake to diminish the impact of model degradation in clinical settings.

**Acknowledgements**

We are grateful to BIDMC, the MIMIC team, Leo Anthony Celi and Brian Gow for their support of this project.

**Data and Code Availability**

The MIMIC-IV dataset is freely available provided you have completed CITI certification. It can be downloaded from [https://physionet.org/content/mimiciv/1.0/](https://physionet.org/content/mimiciv/1.0/). All code used to reproduce the results in this report (including instructions on how to do so) is openly available at [https://github.com/mariehane/ai-gone-astray](https://github.com/mariehane/ai-gone-astray).

**References**

[1] Alistair Johnson et al. *MIMIC-IV*. Version Number: 0.4 Type: dataset. DOI: 10.13026/A3WN-HQ05. URL: [https://physionet.org/content/mimiciv/0.4/](https://physionet.org/content/mimiciv/0.4/) (visited on 02/19/2022) (cited on p. 1).

[2] Michael Moor et al. “Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping”. en. In: *Proceedings of the 4th Machine Learning for Healthcare Conference*. ISSN: 2640-3498. PMLR, Oct. 2019, pp. 2–26. URL: [https://proceedings.mlr.press/v106/moor19a.html](https://proceedings.mlr.press/v106/moor19a.html) (visited on 02/19/2022) (cited on p. 2).

[3] Michael Moor et al. “Predicting sepsis in multi-site, multi-national intensive care cohorts using deep learning”. In: *arXiv:2107.05230 [cs]* (July 2021). arXiv: 2107.05230. URL: [http://arxiv.org/abs/2107.05230](http://arxiv.org/abs/2107.05230) (visited on 02/19/2022) (cited on p. 2).

[4] Christopher Barton et al. “Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs”. In: *Computers in biology and medicine* 109 (June 2019), pp. 79–84. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2019.04.027. URL: [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6556419/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6556419/) (visited on 02/19/2022) (cited on p. 2).

[5] Bret Nestor et al. “Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks”. In: *arXiv:1908.00690 [cs, stat]* (Aug. 2019). arXiv: 1908.00690. URL: [http://arxiv.org/abs/1908.00690](http://arxiv.org/abs/1908.00690) (visited on 02/19/2022) (cited on p. 12).

[6] Zhengping Che et al. “Recurrent Neural Networks for Multivariate Time Series with Missing Values”. en. In: *Scientific Reports* 8.1 (Apr. 2018). Number: 1 Publisher: Nature Publishing Group, p. 6085. ISSN: 2045-2322. DOI: 10.1038/s41598-018-24271-9. URL: [https://www.nature.com/articles/s41598-018-24271-9](https://www.nature.com/articles/s41598-018-24271-9) (visited on 02/19/2022) (cited on p. 3).

[7] Jeffrey L. Elman. “Finding Structure in Time”. en. In: *Cognitive Science* 14.2 (1990). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cog.1402_1, pp. 179–211. ISSN: 1551-6709. DOI: 10.1002/cog.1402_1. URL: [https://onlinelibrary.wiley.com/doi/abs/10.1002/cog.1402_1](https://onlinelibrary.wiley.com/doi/abs/10.1002/cog.1402_1) (visited on 02/24/2022) (cited on p. 4).
A Feature Lists and Origin Categories

Below are the features we found in MIMIC-IV that best approximate the feature categories used by Dascena and Epic’s models, given that their code is not publicly available. For Epic’s feature set, we were unable to find certain feature results in the MIMIC database, including medicines administered by category and the exact mapping of ICD codes for all diagnoses used.

A.1 Dascena Features

| High-level feature | Itemid | Low-level label | Origin |
|--------------------|--------|----------------|--------|
| diastolic blood pressure | 8441 | nbp [diastolic] | chartevents |
| | 8368 | arterial bp [diastolic] | chartevents |
| | 220180 | non invasive blood pressure diastolic | chartevents |
| | 220051 | arterial blood pressure diastolic | chartevents |
| | 225310 | art bp diastolic | chartevents |
| | 8555 | arterial bp #2 [diastolic] | chartevents |
| | 8440 | manual bp [diastolic] | chartevents |
| | 224643 | manual blood pressure diastolic left | chartevents |
| systolic blood pressure | 455 | nbp [systolic] | chartevents |
| | 51 | arterial bp [systolic] | chartevents |
| | 220179 | non invasive blood pressure systolic | chartevents |
| | 220050 | arterial blood pressure systolic | chartevents |
| | 225309 | art bp systolic | chartevents |
| | 6701 | arterial bp #2 [systolic] | chartevents |
| | 442 | manual bp [systolic] | chartevents |
| | 224167 | manual blood pressure systolic left | chartevents |
| | 227243 | manual blood pressure systolic right | chartevents |
| heart rate | 211 | heart rate | chartevents |
| | 220045 | heart rate | chartevents |
| temperature | 678 | temperature f | chartevents |
| | 223761 | temperature fahrenheit | chartevents |
| | 679 | temperature f (calc) | chartevents |
| | 676 | temperature c | chartevents |
| | 223762 | temperature celsius | chartevents |
| respiratory rate | 618 | respiratory rate | chartevents |
| | 220210 | respiratory rate | chartevents |
| | 615 | resp rate (total) | chartevents |
| | 614 | resp rate (spont) | chartevents |
| | 224689 | respiratory rate (spontaneous) | chartevents |
| | 224690 | respiratory rate (total) | chartevents |
| | 651 | spon rr (mech.) | chartevents |
| | 224422 | spon rr | chartevents |
| oxygen saturation | 646 | spo2 | chartevents |
| | 220277 | o2 saturation pulseoxymetry | chartevents |
| | 834 | sao2 | chartevents |
| | 220227 | arterial o2 saturation | chartevents |

Table 4: Full Dascena feature set

A.2 Epic Features

| Feature | Origin |
|---------|--------|
| age     | patients |
| ethnicity | admissions |
| marital status | admissions |
| gender | patients |

Table 5: Static demographical features used by Epic’s model
| High-level feature | Itemid | Low-level label | Origin |
|--------------------|--------|-----------------|--------|
| creatinine         | 791    | creatinine (0-1.3) | chartevents |
|                    | 1525   | creatinine      | chartevents |
|                    | 220615 | creatinine      | chartevents |
| heart rate         | 211    | heart rate      | chartevents |
|                    | 220345 | heart rate      | chartevents |
| hematocrit         | 813    | hematocrit      | chartevents |
|                    | 220545 | hematocrit (serum) | chartevents |
| hemoglobin         | 814    | hemoglobin      | chartevents |
|                    | 220228 | hemoglobin      | chartevents |
| platelets          | 828    | platelets       | chartevents |
|                    | 227457 | platelet count  | chartevents |
| red blood cell count | 833   | rbc             | chartevents |
| respiratory rate   | 614    | resp rate (spont) | chartevents |
|                    | 615    | resp rate (total) | chartevents |
|                    | 618    | respiratory rate | chartevents |
|                    | 651    | spon rr (mech.) | chartevents |
|                    | 220210 | respiratory rate | chartevents |
|                    | 224422 | respiratory rate (spontaneous) | chartevents |
|                    | 224689 | respiratory rate (total) | chartevents |
| respiratory rate set | 619   | respiratory rate set | chartevents |
|                    | 224688 | respiratory rate (set) | chartevents |
| temperature        | 676    | temperature c   | chartevents |
|                    | 677    | temperature c (calc) | chartevents |
|                    | 678    | temperature f   | chartevents |
|                    | 679    | temperature f (calc) | chartevents |
|                    | 223761 | temperature fahrenheit | chartevents |
|                    | 223762 | temperature celsius | chartevents |
| white blood cell count | 861   | wbc (4-11,000) | chartevents |
|                    | 1127   | wbc (4-11,000)  | chartevents |
|                    | 1542   | wbc             | chartevents |
|                    | 220546 | wbc             | chartevents |
| Band neutrophils   | 51144  | Bands           | labevents |
| Base excess         | 50802  | Base Excess     | labevents |
| Lymphocyte          | 51244  | Lymphocytes     | labevents |
| Mean corpuscular hemoglobin concentration | 51249 | MCHC | labevents |
| Monocytes           | 51254  | Monocytes       | labevents |
| Neutrophils         | 51256  | Neutrophils     | labevents |
| Nucleated red blood cell count | 51257 | Nucleated Red Cells | labevents |
| Red Blood Cell morphology | 52171 | RBC Morphology | labevents |
| Red Blood Cell distribution width | 52204 | RBCDist | labevents |
| reticulocyte count  | 51282  | Reticulocyte Count, Absolute | labevents |
| Segmented neutrophil count | 51232 | Hypersegmented Neutrophils | labevents |
### Table 6: Time-varying Epic features.

| High-level feature                  | Itemid   | Low-level label       | Origin                  |
|-------------------------------------|----------|-----------------------|-------------------------|
| Peripherally inserted central catheters | 224264   | PICC Line             | procedureevents         |
| Central venous catheters            | 225311   | Tunneled (Hickman) Line | procedureevents         |
| Drains                              | 225447   | percutaneous drain insertion | procedureevents         |
|                                     | 225456   | ventricular drain     | procedureevents         |
|                                     | 226475   | intraventricular drain inserted | procedureevents         |
|                                     | 22923    | subdural drain        | procedureevents         |
|                                     | 22924    | lumbar drain          | procedureevents         |
| Feeding tube                        | 224007   | gi #1 intub site      | chartevents             |
|                                     | 224441   | gi #2 intub site      | chartevents             |
|                                     | 224442   | gi #3 intub site      | chartevents             |
|                                     | 224472   | incision site #1      | chartevents             |
|                                     | 224473   | incision site #2      | chartevents             |
|                                     | 224474   | incision site #3      | chartevents             |
|                                     | 224475   | incision site #4      | chartevents             |
|                                     | 224476   | incision site #5      | chartevents             |
|                                     | 224477   | incision site #6      | chartevents             |
|                                     | 224559   | incision #1- location | chartevents             |
|                                     | 224560   | incision #2- location | chartevents             |
|                                     | 224561   | incision #3- location | chartevents             |
|                                     | 224562   | incision #4- location | chartevents             |
|                                     | 224563   | incision #5- location | chartevents             |
|                                     | 224564   | incision #6- location | chartevents             |
|                                     | 229015   | incision #7- location | chartevents             |
|                                     | 229016   | incision #8- location | chartevents             |
|                                     | 229017   | incision #9- location | chartevents             |
|                                     | 229018   | incision #10- location| chartevents             |
| Pressure Ulcers                     | 228506   | pressure ulcer #1- location | chartevents         |
|                                     | 228507   | pressure ulcer #2- location | chartevents         |
|                                     | 228508   | pressure ulcer #3- location | chartevents         |
|                                     | 228509   | pressure ulcer #4- location | chartevents         |
|                                     | 228510   | pressure ulcer #5- location | chartevents         |
|                                     | 228511   | pressure ulcer #6- location | chartevents         |
|                                     | 228512   | pressure ulcer #7- location | chartevents         |
|                                     | 228513   | pressure ulcer #8- location | chartevents         |
|                                     | 228514   | pressure ulcer #9- location | chartevents         |
|                                     | 228515   | pressure ulcer #10- location | chartevents        |
| Active penicillin orders            | 008880   | Penicillin V Potassium | prescriptions          |
|                                     | 043350   | Penicillin G Benzathine | prescriptions          |
| Active vancomycin orders            | 043952   | Vancomycin            | prescriptions          |
|                                     | 009331   | Vancomycin            | prescriptions          |
|                                     | 009328   | Vancomycin            | prescriptions          |
|                                     | 009329   | Vancomycin            | prescriptions          |
|                                     | 087111   | Vancomycin            | prescriptions          |
|                                     | 029011   | Vancomycin            | prescriptions          |

### Table 7: Epic features based on ICD-codes.

| Feature                              | ICD Code | ICD Version | Origin        |
|--------------------------------------|----------|-------------|---------------|
| Diabetes                             | E11      | ICD-10      | diagnoses_icd |
|                                      | E10      | ICD-10      | diagnoses_icd |
| Hypertension                         | I10      | ICD-10      | diagnoses_icd |
| HIV                                  | 42       | ICD-9       | diagnoses_icd |
|                                      | B20      | ICD-10      | diagnoses_icd |
| Obesity                              | 27800    | ICD-9       | diagnoses_icd |
|                                      | E66      | ICD-10      | diagnoses_icd |
| Coronary Artery Disease              | 41400    | ICD-9       | diagnoses_icd |
|                                      | 1251     | ICD-10      | diagnoses_icd |
| Congestive Heart Failure             | I502     | ICD-10      | diagnoses_icd |
|                                      | I503     | ICD-10      | diagnoses_icd |
|                                      | I504     | ICD-10      | diagnoses_icd |
| Chronic Obstructive Pulmonary Disease (COPD) | I44  | ICD-10      | diagnoses_icd |
| Chronic Kidney Disease               | I13      | ICD-10      | diagnoses_icd |
|                                      | I12      | ICD-10      | diagnoses_icd |
| Chronic Liver Disease                | 5719     | ICD-9       | diagnoses_icd |
|                                      | K76      | ICD-10      | diagnoses_icd |
B Other Prediction Tasks

To validate our results of temporal drift with Nestor et al.’s [5] experiments on MIMIC-III, we followed a similar setup to run experiments on Length of Stay and ICU Mortality prediction tasks on MIMIC-IV, with results in Table 8 and Table 9. The length of stay prediction task involves taking the first 24 hours of data of a patient’s stay in the ICU, and predicting whether this patient’s specific stay in the ICU would last >= 3 days (a binary prediction task). The ICU mortality task uses the same feature set as the LOS prediction task, and likewise uses the first 24 hours of data of a patient’s ICU stay to predict whether a patient would die in the ICU during this particular stay. The features used for both prediction tasks were the same 8 demographic variables and 141 vital signs from the chartevents table used by Nestor et al.

B.1 Length-of-stay (LOS) Results

| Year Buckets | Model | Year-Agnostic | 2008-2010 | 2011-2013 | 2014-2016 | 2017-2019 |
|--------------|-------|---------------|-----------|-----------|-----------|-----------|
|              | RNN   | 0.658         | 0.661     | 0.635     | 0.623     | 0.564     |
|              | Logistic | 0.601       | 0.620     | 0.600     | 0.585     | 0.549     |

Table 8: Results from Length-of-stay prediction on MIMIC-IV, trained on raw data representations

B.2 ICU Mortality Results

| Year Buckets | Model | Year-Agnostic | 2008-2010 | 2011-2013 | 2014-2016 | 2017-2019 |
|--------------|-------|---------------|-----------|-----------|-----------|-----------|
|              | RNN   | 0.830         | 0.745     | 0.760     | 0.776     | 0.777     |
|              | Logistic | 0.680       | 0.696     | 0.684     | 0.693     | 0.690     |

Table 9: Results from ICU Mortality prediction on MIMIC-IV, trained on raw data representations