Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation

Bret Nestor
University of Toronto
bretnestor@cs.toronto.edu

Matthew B. A. McDermott
MIT
mmd@mit.edu

Geeticka Chauhan
MIT
geeticka@mit.edu

Tristan Naumann
Microsoft Research
tristan@microsoft.com

Michael C. Hughes
Tufts University
mhughes@cs.tufts.edu

Anna Goldenberg
Hospital for Sick Children, University of Toronto, Vector Institute
anna.goldenberg@utoronto.ca

Marzyeh Ghassemi
University of Toronto, Vector Institute
marzyeh@cs.toronto.edu

Abstract

Machine learning for healthcare often trains models on de-identified datasets with randomly-shifted calendar dates, ignoring the fact that data were generated under hospital operation practices that change over time. These changing practices induce definitive changes in observed data which confound evaluations which do not account for dates and limit the generalisability of date-agnostic models. In this work, we establish the magnitude of this problem on MIMIC, a public hospital dataset, and showcase a simple solution. We augment MIMIC with the year in which care was provided and show that a model trained using standard feature representations will significantly degrade in quality over time. We find a deterioration of 0.3 AUC when evaluating mortality prediction on data from 10 years later. We find a similar deterioration of 0.15 AUC for length-of-stay. In contrast, we demonstrate that clinically-oriented aggregates of raw features significantly mitigate future deterioration. Our suggested aggregated representations, when retrained yearly, have prediction quality comparable to year-agnostic models.

1 Introduction

Training predictive machine learning systems for clinical outcomes has been made possible by the shift towards electronic health records (EHR) in modern healthcare systems. The best example of this is the widely-used MIMIC-III dataset [1], which is a public *de-identified* dataset from a single hospital's intensive care units (ICU).

A key part of de-identification is obscuring any calendar dates related to care, e.g., MIMIC-III dates are shifted into the future by a random offset for each individual patient, resulting in new dates "between the years 2100 and 2200" [1]. Not accounting for such shifts is problematic because EHR data is fundamentally generated as a byproduct of care. Care practices evolve over time via concept drift [2] and are serious in their own right [3]. MIMIC notably reflects such a change, through the EHR system update from Carevue to Metavision in the Beth Israel Deaconess Medical Center in 2008 [1]. However, even without an EHR change, a *random date shift* may also cause models to unintentionally *train* with data generated from newer care practices than they are tested on.

In Machine Learning for Health (ML4H) Workshop at NeurIPS 2018. https://arxiv.org/abs/1811.07216

Previous work has trained models from MIMIC-III data to predict outcomes such as mortality [4; 5], length-of-stay [6] or billing codes [7]. However, it is standard machine learning practices to sample a random set of patients for training and test splits, and select a large number of raw features as input. Our goal in this paper is two-fold: 1) to investigate whether date randomisation interferes with evaluation of predictive models trained with standard practices, and 2) to identify the extent that standard health tasks are meaningful over evolving care practices.

We show that standard machine learning practices on the date-randomised data from MIMIC over-estimates predictive model performance on ICU mortality prediction - a common task. We further demonstrate that simple feature aggregation can mitigate degradation in predictive power over time. We identify that predictive performance saturates quickly in mortality prediction, and further demonstrate that other tasks show a similar performance pattern, and improvement to generalisation as a result of feature aggregation.

2 Data

We use MIMIC III, a public dataset with EHR data from over $58,900$ hospital admissions of nearly $38,600$ adults at Beth Israel Deaconess Medical Center from 2001 to 2012 [1]. We consider the first intensive care unit (ICU) stay of patients older than 15 who were in the ICU for at least 36 hours ($21,877$ unique ICU stays). We extract patient demographic data, as well as physiological measurements queried from the `CHARTEVENTS` table, which contains the measurements captured at patient bedside. We use the same steps taken to pre-process data as found in [8; 9]. In addition to the publicly-available MIMIC data, a Limited Data Use Agreement was granted to provide the admitting year for these patients.

We use the first 24 hours of data for each patient, and collect physiological measurements into hourly buckets via averaging. Several works have focused on imputation methods for healthcare data [5; 10; 11; 12; 13; 14]. We use simple imputation to assign three sub-features to the data: the imputed (forward-filled) measurement of the feature, a binary indicator of whether or not that feature was observed at that time, and the number of hours since the feature was last observed[5].

3 Methods

Data Representation For each patient, we investigate the impact of two common data representation strategies on tasks.

1. `Item-ID` representation: First, we allow measurements of each variable's own raw `itemid`[1], which is a common strategy in prediction.
2. `Clinically Aggregated` representation: While prior work has sought to learn mappings between similar features[15; 16], we use simple aggregations of highly-present `itemid`s [6] to create "expert driven" condensed categories[2].

Models We use a random forest (RF) classifier for all tasks with simple imputation to handle missing data, as described in [5]. Depending on whether a), the test years overlap with the training years, or b), the training years completely precede the test years, then variation will be introduced by random train/test splits (80% and 20%) and random initialisations, respectively.

In all cases, 5-fold cross validation was applied to the training data, using a random search to find best parameters for maximum area under the receiver-operator curve (AUROC) on the validation split. It is important to note that since the variation is introduced using samples containing overlapping data, these samples cannot be treated as independent. However, any differences in the data should explain differences in the way that aggregation (and the confounding missingness) affects AUROC scores. We use a Wilcoxon signed-rank test [17] to test for significance between models.

Tasks We test on two common baseline clinical machine learning tasks: mortality prediction and length-of-stay (LOS) predicted, realized as a classification task by splitting patients between high LOS (≥ 3 days) and low LOS (< 3 days).

[1] The `itemid` is the internal MIMIC identifier for different lab measures, vitals, etc.

[2] An example of this aggregation is given in Appendix A

Experiments In addition to measuring baseline, year-agnostic performance, we train models for both tasks under 3 regimes, in all cases profiling the performance difference between using the raw `Item-ID` representation vs. our `Clinically Aggregated` representation.

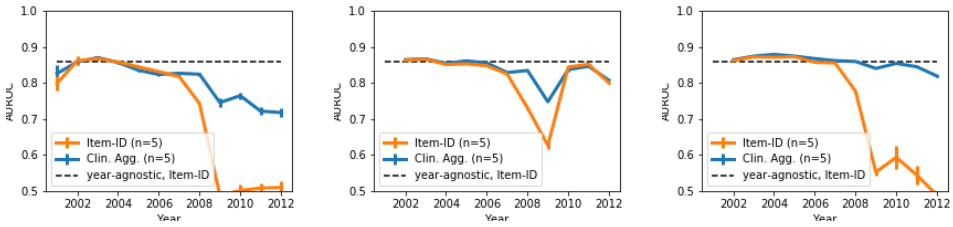
1. **Year-Specific one-time training**: Train models on the data from 2001 and 2002. If a sudden `itemid` shift occurs in recording a vital, this model will not recover.
2. **Year-Specific continuous training**: Train a model on **all** previous years e.g., data from 2001-2005 will be used to train a model that is tested on data from 2006.
3. **Year-Specific short-term training**: Train models on the data from **only** the previous year, e.g., data from 2005 will be used to train a model that is tested on data from 2006.

In addition to these temporal drift experiences, we assess how much performance over time changes as we reduce the size of our overall train set.

4 Results and Discussion

4.1 Performance comparison between feature representations for clinical prediction

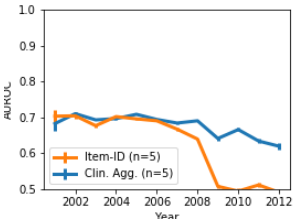
We assess mortality prediction and LOS prediction for both of our data representations across all training regimes mentioned in Section 3 in Figure 1. Overall, we note that the Clinical Aggregate representation is much more robust to the performance degradation over time observed under the `Item-ID` representation. This resembles the findings in [15] though our clinically determined groupings appear to offer a lower drop in performance across that shift in practice than do their learned representations.



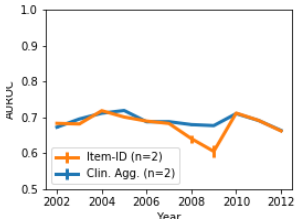
(a) Mortality AUC, models trained on 2001-2002 data.

(b) Mortality AUC, models trained yearly on prior year only.

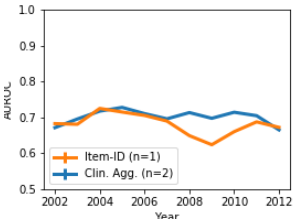
(c) Mortality AUC, models trained yearly on all prior data.



(d) LOS AUC, models trained on 2001-2002 data.



(e) LOS AUC, models trained yearly on prior year only.



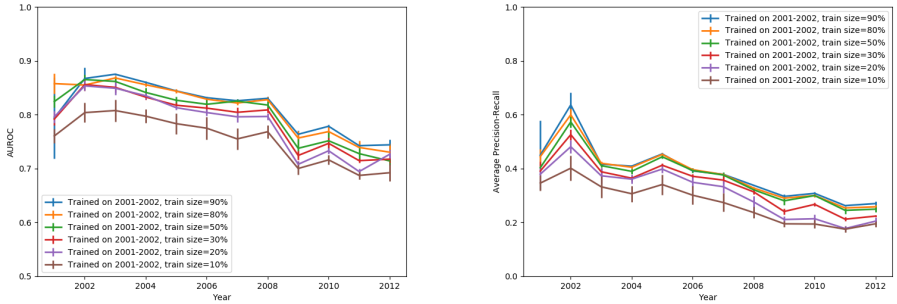
(f) LOS AUC, models trained yearly on all prior data.

Figure 1: Performance of RF classifiers using `Item-Id` and `Clinically Aggregated` representations on mortality (top) and LOS prediction (bottom). Error bars indicate $\pm$ standard error.

In Figures 1a and 1d we demonstrate that models trained on the basic `Item-ID` representation using 2001-2002 data degrade after the hospital system change and do not perform up to the expectations of baseline training paradigms. In Figures 1b and 1e, we further show that training models on the previous year only causes a shock in performance during system change, and Figures 1c and 1f identify that only the `Clinically Aggregated` representation trained on all prior data is able to sustain the expected performance throughout the system change.

4.2 Models Saturate Quickly on Mortality Prediction, Impacting Generalisation

We note that predictive performance in RF models trained with `Clinically Aggregated` representations on only 2001-2002 data performs at a higher level than expected. Specifically, an RF model trained on 90% of 2001-2002 data (1982 patients) is able to predict mortality ten years later (2012) with an AUROC of 0.744 $\pm$ 0.019 (AUPR 0.270 $\pm$ 0.008). Surprisingly, when the training data was reduced to only 10% (220 patients) this was observed to drop to AUROC of 0.692 $\pm$ 0.032 (AUPR of 0.195 $\pm$ 0.012) (Figure 2).



(a) AUROC (b) AUPRC

Figure 2: Generalisability of RF models trained on partial data from 2001-2002 alone in ICU mortality prediction. As training data from is reduced from 1,982 to 220 patients (90% to 10% training samples), performance decreases by only 7% for predicting patient outcomes 10 years into the future.

Such prediction results imply that mortality may be a trivial prediction task. Using feature ablations, we demonstrate that one feature, Glasgow coma scale, alone is able to sustain AUROCs greater than 0.77 for the duration in which it is measured (Figure 3).

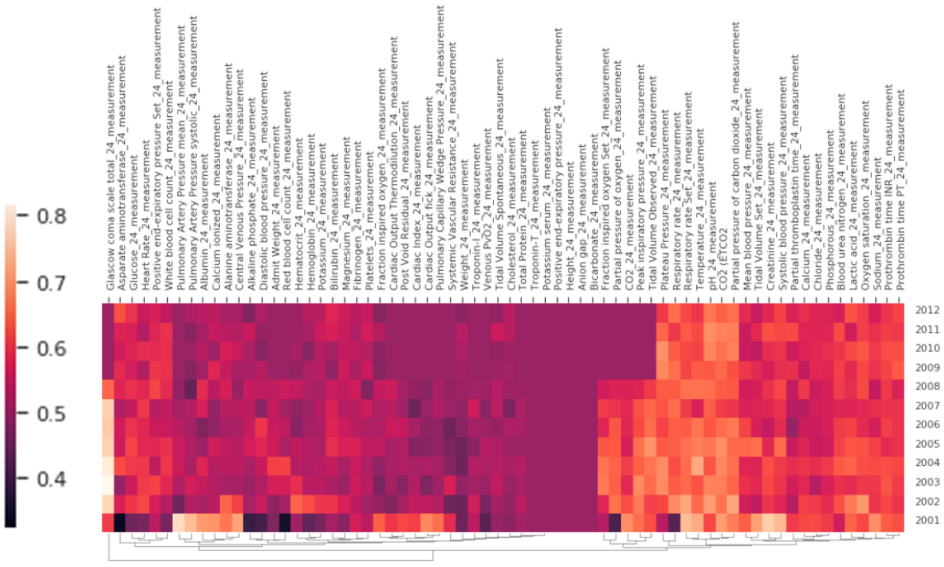


Figure 3: Random forest classifiers trained on 2001-2002 data of a single feature measured over 24 hours using the `Item-ID` representation (n=5 per feature)

5 Conclusion

While it is essential to compare models on standardised tasks, we should not expect to train models that translate into the clinic on MIMIC features without careful consideration of how data was generated [18]. We demonstrate that only using a continuous training regime of models, with a `Clinically Aggregated` representation can generate yearly performances that are comparable to those reported when training year-agnostic models.

References

[1] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, 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*, 3:160035, 2016.

[2] I. Žliobaitė. Learning under Concept Drift: an Overview. *ArXiv e-prints*, October 2010.

[3] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[4] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 75–84. ACM, 2014.

[5] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1), 2018.

[6] 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.

[7] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795. ACM, 2017.

[8] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 322–337, Boston, Massachusetts, 18–19 Aug 2017. PMLR.

[9] M.B.A. McDermott, T. Yan, T. Naumann, N. Hunt, H. Suresh, P. Szolovits, and M. Ghassemi. Semi-supervised Biomedical Translation with Cycle Wasserstein Regression GANs. In *Association for the Advancement of Artificial Intelligence*, New Orleans, LA, 2018.

[10] Volker Tresp and Thomas Briegel. A Solution for Missing Data in Recurrent Neural Networks with an Application to Blood Glucose Prediction. In *Advances in Neural Information Processing Systems*, 1997.

[11] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing Data Imputation using Generative Adversarial Nets. In *International Conference on Machine Learning*, 2018.

[12] Kristel J.M. Janssen, A. Rogier T. Donders, Frank E. Harrell, Yvonne Vergouwe, Qingxia Chen, Diederick E. Grobbee, and Karel G.M. Moons. Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*, 63(7):721–727, 2010.

[13] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *BMJ*, 361(k1479), 2018.

[14] Jau-Huei Lin and Peter J Haug. Exploiting missing clinical data in bayesian network modeling for predicting medical problems. *Journal of biomedical informatics*, 41(1):1–14, 2008.

[15] Jen J Gong, Tristan Naumann, Peter Szolovits, and John V Guttag. Predicting Clinical Outcomes Across Changing Electronic Health Record Systems. In *the 23rd ACM SIGKDD International Conference*, pages 1497–1505, New York, New York, USA, 2017. ACM Press.

[16] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, P. J. Liu, X. Liu, M. Sun, P. Sundberg, H. Yee, K. Zhang, G. E. Duggan, G. Flores, M. Hardt, J. Irvine, Q. Le, K. Litsch, J. Marcus, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. Howell, C. Cui, G. Corrado, and J. Dean. Scalable and accurate deep learning for electronic health records. *ArXiv e-prints*, January 2018.

[17] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[18] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, and Rajesh Ranganath. Opportunities in machine learning for healthcare. *arXiv preprint arXiv:1806.00388*, 2018.

Appendices

Appendix A Feature aggregation description

For example, in MIMIC III, `CHARTEVENTS` with `itemids` *861, 1127, 1542, and 220546* are averaged into one feature in the `Clinically Aggregated` feature vector called *White blood cell count*. Note that `itemids` *861, 1127, and 1542* were recorded in the 2001-2008 system and `itemid` *220546* was recorded in the 2008-2012 system.

Appendix B Significance test for representation generalisability

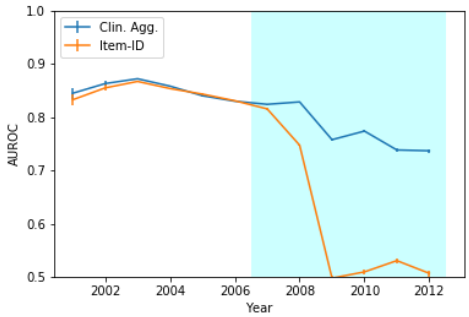


Figure 4: Area under the ROC for the task of classifying patient mortality in ICU, after observing 24 hours of data. Shaded regions indicate significant differences between `Clinically Aggregated` representation and `Item-ID` representation (Wilcoxon signed-rank test, $p<0.01$, $n=20$)