Instability in clinical risk stratification models using deep learning

Daniel Lopez-Martinez[*]　　　　　　　　　　　　DLMOCDM@GOOGLE.COM
Alex Yakubovich　　　　　　　　　　　　　　　　ALEXYAK@GOOGLE.COM
Martin Seneviratne　　　　　　　　　　　　　　MARTSEN@GOOGLE.COM
Adam D. Lelkes　　　　　　　　　　　　　　　　LELKES@GOOGLE.COM
Akshit Tyagi　　　　　　　　　　　　　　　　AKSHITTYAGI@GOOGLE.COM
Jonas Kemp　　　　　　　　　　　　　　　JONASBKEMP@GOOGLE.COM
Google Research, Mountain View, CA, USA

Ethan Steinberg　　　　　　　　　　　　　　ETHANID@STANFORD.EDU
N. Lance Downing　　　　　　　　　　　　LDOWNING@STANFORD.EDU
Ron C. Li　　　　　　　　　　　　　　　　　RONL@STANFORD.EDU
Keith E. Morse　　　　　　　　　　KEITH.MORSE@STANFORD.EDU
Nigam H. Shah　　　　　　　　　　　　　NIGAM@STANFORD.EDU
Department of Medicine, Stanford University, Stanford, CA, USA

Ming-Jun Chen　　　　　　　　　　　　MINGJUNCHEN@GOOGLE.COM
Google Research, Mountain View, CA, USA

Abstract

While it has been well known in the ML community that deep learning models suffer from instability, the consequences for healthcare deployments are under characterised. We study the stability of different model architectures trained on electronic health records, using a set of outpatient prediction tasks as a case study.

We show that repeated training runs of the same deep learning model on the same training data can result in significantly different outcomes at a patient level even though global performance metrics remain stable.

We propose two stability metrics for measuring the effect of randomness of model training, as well as mitigation strategies for improving model stability.

Keywords: Model Stability, Nondeterminism, Neural Networks, Model Ensembles, Electronic Health Records

---

[*] Corresponding author.

1. Introduction

Machine learning methods have shown great promise in a wide range of healthcare applications, from diagnosing conditions from imaging, to personalizing care, to improving administrative workflows (see e.g. Davenport and Kalakota (2019)). One particular use case that has attracted a lot of research interest is *risk stratification*; i.e., assigning each patient a risk status using a machine learning model trained to predict a patient's individual risk of some adverse event. Previously studied risk prediction tasks cover a range of inpatient and outpatient risks, from acute inpatient predictions such as acute kidney injury (Mohamadlou et al., 2018) to longer term outpatient predictions such as readmissions (Huang et al., 2021).

These risk stratification use cases are typically motivated by the promise of improving overall patient care by helping clinicians prioritize the highest risk patients. Therefore,

they can be used to inform the allocation of scarce and valuable resources, such as a care manager's time, among a group of patients. Whenever algorithmic tools are used for resource allocation, especially in such an important and sensitive area as healthcare, ensuring the robustness and fairness of the algorithmic decision-making process is of utmost importance.

In this paper, we want to draw the community's attention to a particular source of concern in this respect: model instability. As in many other areas of machine learning applications, deep learning methods often demonstrate state-of-the-art performance for risk prediction tasks. It is well known in the machine learning community that deep learning models are more sensitive to random phenomena during model training, and their outputs are often less stable than those of traditional, simpler models. However, the practical implications of model instability for risk stratification haven't been well studied.

As a case study for this phenomenon, we train feed-forward neural network model on EHR data to predict all-cause emergency admissions, along with a set of disease-specific admission events. These models are trained an data from the EHR of an academic medical center. Importantly, we evaluate the impact of instability in a pragmatic fashion, by simulating the prospective deployment setting of the model—namely, as a method to select high-risk patients for a care manager to contact. We compare deep learning (DL) models against simpler logistic regression (LR) models on both performance and stability metrics. We conclude by proposing an ensembling approach to mitigate against DL instability and suggest directions for future methodological work.

While we used unplanned hospital admission predictions as our case study, the concept of model stability and our methods to evaluate it are generalizable to any other models where the ordinal nature of the risk scores is important—essentially any clinical workflow where resources (e.g. capacity to intervene or follow up) are restricted.

2. Related work

Risk stratification models have been built for a number of prediction tasks, such as forecasting mortality (Luo et al., 2022), patient deterioration (Mann et al., 2021), unplanned hospital admissions (King et al., 2022), or the need for medical procedures (Lopez-Martinez et al., 2022). However, the performance of these models is usually evaluated using traditional metrics of predictive performance, such as the area under the curves (AUC) of the receiver operating characteristic (ROC-AUC) and the precision-recall curve (PR-AUC). Although having merit, recent work has highlighted the need for additional metrics to evaluate models.

For example, Mišić et al. (2021) has noted that traditional metrics lack interpretability and provide an incomplete picture of the overall benefit of the model as they do not take into account how providers may use them. For example, if used to allocate scarce medical resources, resource constraints should inform model evaluation. King et al. (2022) has also noted that building a model for real-world implementation involves several additional challenges to those encountered when simply optimising a model's technical performance, such as evaluating the drift in model performance over time in deployment simulations. Model fairness (Rajkomar et al., 2018a; Mhasawade et al., 2021) should also be key aspect of model evaluation, to avoid disparities in resource allocation. Therefore, if risk-stratification models are to be used operationally, model evaluation should go beyond the traditional metrics.

Reproducibility is also a key aspect of the assessment of a model, and a minimal prerequisite for any machine learning model applied to human health (McKinney et al. (2020); Beam et al. (2020); Liu et al. (2022)). Lack of reproducibility undermines the credibility of a prediction, and diminishes user trust in the machine learning model. Moreover, regulatory agencies are implementing rules governing medical use of machine learning (Vollmer et al. (2020)), and may reasonably expect being able to duplicate the results of model's evaluation using the same elements as were used in the prior study.

Unfortunately, even models trained with the same training algorithm and dataset can have different model parameters and yield different predictions. This is due to nondeterminism during training (Nagarajan and Warnell; Summers and Dinneen, 2021; Madhyastha and Jain, 2019), which includes:

- **Parameter initialization** Prior to training, the weights of a NN are initialized randomly (often from a Gaussian or uniform distribution centered at zero) (Glorot and Bengio, 2010).

- **Stochastic optimization** NNs are trained with stochastic gradient descent (SGD) or other similar stochastic optimization techniques, which typically sample small batches of examples from the training set at random in each iteration.

- **Stochastic regularization layers** Regularization techniques such as dropout or noisy activation functions (Gulcehre et al., 2016).

In addition to randomness due to stochastic training algorithms, the hardware and software tooling used for model training can also result in differences across models trained with the same data and algorithm.

These often trade determinisim and precision for speed, introducing nondeterminism e.g. by the random completion order of parallel tasks (Zhuang et al., 2022).

In the EHR literature, model stability issues have already been identified. D'Amour et al. (2020) and Dusenberry et al. (2020) have noted that changing random seeds in training led to substantially different test set predictions. Therefore, Dusenberry et al. (2020) highlighted that model uncertainty is a key factor one should pay attention to while adopting a DL approach to predict events from EHR data. Despite these insights, the instability of deep learning models with EHR data, and specifically its impact in the real-world use of risk stratification models, remains poorly understood. For example, DL training instability has been shown to lead to variability in model fairness (Qian et al., 2021), but this has not been studied in clinical models.

3. Methods

3.1. Dataset

We used a limited dataset provided by the STAnford medicine Research data Repository (STARR-OMOP) (Datta et al., 2020), which consists of Stanford's EHR data in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Data from Lucile Packard Children's Hospital (LPCH), which is a pediatric institution, was excluded from this study. The dataset was utilized in accordance with HIPAA and Stanford Medicine Research IT processes, including IRB review.

To leverage existing FHIR-based modeling infrastructure, we manually mapped the dataset from the OMOP CDM to the HL7 Fast Healthcare Interoperability Resources (FHIR) standard (version R4) (Bender and Sartipi (2013)).

3.2. Model development

To explore the stability of risk stratification methods, we implemented 4 outpatient deterioration tasks as a case study.

3.2.1. Task definition

We predicted the 30-day risk of all-cause emergency hospitalizations and three condition-specific potentially preventable unplanned hospitalizations: chronic obstructive pulmonary disease (COPD), diabetes and heart failure (HF).

Emergency admissions were identified by examining the class and admit source of the medical encounter. The 3 disease-specific unplanned hospitalizations were defined using the inclusion and exclusion criteria of the Agency for Healthcare Research and Quality's Patient Quality Indicators (PQIs) for ambulatory care-sensitive conditions (Agency for Healthcare Research and Quality, 2016, 2018), and identified by examining the diagnosis and procedure codes corresponding to the PQI definition (Quan et al., 2005).

3.2.2. Label extraction

Following Rajkomar et al. (2018b), we represented each patient's history as an ordered timeline of FHIR resources. As in Lopez-Martinez et al. (2022), for each patient we generated a prediction on the first day of every month within the study period, if and only if the patient had at least one FHIR resource within the preceding year. Each example in our dataset corresponded to one such monthly prediction: we included all resources occurring prior to the prediction time as features, and we computed a binary label denoting whether the patient had an emergency admission within the first 30 days following the prediction time.

3.2.3. Feature extraction

For each patient, we extracted information about the demographics, conditions, procedures, medications, observations and encounters.

The demographics features were gender and age. The condition-based features contained all International Classification of Diseases (ICD) codes associated with each encounter. The procedure-based features included all Current Procedural Terminology (CPT) codes, as well as ICD procedure codes. The medication-based features contained RxNorm codes (Liu et al., 2005) for medication requests. The observation-based features included indicators for measurements of patients' vital signs and lab results and were constructed by concatenating the observation code (LOINC (McDonald et al., 2003) and SNOMED (Cornet and de Keizer, 2008) codes) and the unit of measurement. The encounter-based feature was the class of encounter.

Following Lopez-Martinez et al. (2022), we bucketed the feature values into one of four disjoint time intervals indicating how long before the prediction time that feature value occurred: less than 30 days before, 30-90 days before, 90 days-1 year before, or 1-10 years before. The demographic features (age and gender) were not bucketed. Each feature was represented as a multi-hot vector.

3.2.4. Model architecture

Using the aforementioned features, we designed independent logistic regression (LR) and feedforward neural network (FNN) machine learning models for each of the 4 tasks. The FNN consisted of three hidden layers of 128, 64, and 32 units respectively, with ReLU activations and dropout after each hidden layer, followed by a final prediction layer. Weights for the FNN model were randomly initialized using Glorot uniform initialization

(Glorot and Bengio, 2010) on each training run; weights for the LR model were simply initialized to zero. All models were trained using Adam (Kingma and Ba, 2014), a stochastic gradient-based optimizer method, with batch size 512. Predictions were rendered on the 1st day of every month on those patients who had at least one health interaction within the preceding year and were aged 18 years or older.

All models were implemented and trained with TensorFlow in Python Abadi et al. (2016). Model hyperparameters, including L1/L2 regularization coefficients, dropout rates, and the parameters of an exponential learning rate decay, were tuned using Google Vizier Golovin et al. (2017) to minimize loss on the validation set.

3.2.5. Model calibration

While it is desirable for our models to output well-calibrated risk probabilities, our LR and FNN models, like most supervised learning methods, produce scores that can be used to rank the predictions from highest to lowest risk, but are not accurate probability estimates.

To address this limitation, we used isotonic regression (Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2005) to calibrate these models and transform ranking scores into probability estimates. Specifically, we fit a non-parametric isotonic (monotonically increasing) step-wise function m to the original model scores f_i such that $y_i = m(f_i) + \epsilon_i$, where y_i denotes the true targets. Isotonic regression works by finding the isotonic function $\hat{m}$ such that:

$$\hat{m} = \arg\min_z \sum (y_i - z(f_i))^2 \qquad (1)$$

subject to $z(f_i) \geq z(f_j)$ whenever $f_i \geq f_j$.

All risk predictions reported in this papers are calibrated probabilities obtained from after applying isotonic regression.

3.3. Model evaluation

Model evaluation was done on an fixed independent held out test set comprising a 10% random selection of the available patients. No patients in the test set were used for model training or validation.

Performance was evaluated using two approaches. In the first one, we aggregated all test set predictions to compute the ROC-AUC and PR-AUC. Note that given the low prevalence of triggers with positive labels, the PR-AUC metric is more informative than ROC-AUC (Saito and Rehmsmeier, 2015). In the second approach, we simulated the real-word deployment of the model by evaluating the test set predictions of each month independently (from January 2009 to December 2021), and then averaging these monthly metrics. This last evaluation approach allows us to introduce capacity constraints by examining the top K highest risk patients each month only.

3.4. Stability analysis

To quantify model stability, in addition to the standard deviation of the ROC-AUC and PR-AUC metrics, we implemented the following metrics, which are computed using $N = 10$ model runs with exactly the same training data, configurations and model hyperparameters, but different random seeds.

3.4.1. Top K Jaccard index

In risk stratification methods, users may care about the K highest-risk patients, as they are often limited in the amount of patients they can act upon. Therefore, it is reasonable to expect that the list of top K patients should be the same across different models.

To assess the similarity of the N patient lists, we use the pairwise Jaccard index which calculates the average Jaccard index across all pairwise comparisons. The Jaccard index

is given by

$$j(X_i^K, X_j^K) = \frac{|X_i^K \cap X_j^K|}{|X_i^K \cup X_j^K|} \qquad (2)$$

where X_i^K and X_j^K denote the top K predictions in models i and j respectively.

For N models, there exist $\binom{N}{2}$ pairwise comparisons. Therefore, the top K pairwise Jaccard index is defined as follows:

$$J = \frac{1}{\binom{N}{2}} \sum_{i<j} j(X_i^K, X_j^K) \qquad (3)$$

The choice of K depends on the specific use case.

3.4.2. RANK CORRELATION COEFFICIENT

Risk scores are often used to prioritize patients, so that those with higher risk scores are acted upon earlier. In these scenarios, the ordinal nature of the patient lists is important. Users would expect models to produce similarly ranked patient lists.

Kendall's τ is commonly used to compare the ordinal correlation between two ranked lists (Kendall, 1938, 1945).

It is based on computing the number of concordant ($|C|$) and discordant ($|D|$) pairs:

$$\tau = \frac{|C| - |D|}{\binom{N}{2}} \qquad (4)$$

where a pair $((x_i, y_i), (x_j, y_j))$ is defined to be concordant if the sort order of the x's is the same as the sort order of the y's, otherwise it is discordant.

We also consider a weighted version of Kendall's tau with hyperbolic weighting, in which correlations between top ranked patients have more influence on the metric (Vigna (2015)).

In this formulation, a datapoint of rank r is assigned a weight of $(r + 1)^{-1}$. The weight of exchanging two datapoints with ranks r and s is the sum of their weights, i.e. $(r + 1)^{-1} + (s + 1)^{-1}$.

Weighting captures the assumption that the relative ordering of very low-risk patients is less important since interventions are typically aimed at high-risk patients.

3.5. Ensembling

Ensembling is a popular approach for improving prediction performance across many different tasks, and it also reduces run-to-run variablity. Averaging the outputs of multiple models reduces the random noise which might be causing the unnecessary variance in the predictive distribution (Huang et al., 2017; Xie et al., 2013). It has also been shown to give better-calibrated estimates for uncertainty in the predictive distributions (Lakshminarayanan et al., 2016). In our experiments we use a simple arithmetic mean of model outputs as a baseline; more complex averaging schemes could potentially yield further improvement.

4. Results

4.1. Dataset

A total of 1,576,215 individuals were present in the dataset. Of these, 53.66% were females. White was the most predominant race (44%), followed by asian (18.13%), black (3.85%), hispanic (2.02%) and native american (1.12%).

From these patients, we extracted a total of 66,977,919, 6,412,742, 5,588,138, and 2,400,145 predictions for the emergency, COPD, diabetes and heart failure tasks respectively. These predictions spanned a time period from January 1st 2009 to December 1st 2021. The differences in the number of predictions for each task can be explained by the different cohort inclusion criteria for prediction rendering.

Task	# predictions	# positives	ROC-AUC		PR-AUC	
			LR	FNN	LR	FNN
Emergency	60,259,528	104,123 (0.17%)	0.9052	**0.9116**	0.0734	**0.0858**
COPD	5,759,224	1,228 (0.02%)	0.8894	**0.9234**	0.0047	**0.0069**
Diabetes	5,034,065	701 (0.01%)	0.8273	**0.8919**	**0.0177**	0.0057
Heart failure	2,161,112	2,369 (0.11%)	0.8717	**0.8789**	0.0160	**0.0170**

Table 1: Number of predictions and positive label prevalence in the development (train and validation) set, together with model performance on the aggregated test set predictions for single ($N = 1$) LR and FNN models. For each metric, the best performing model is highlighted in bold. ROC: receiver operating characteristic; PR: precision-recall curve; AUC: area under the curve.

Using this data, we generated training, validation and test sets by randomly selecting 80%, 10% and 10% of patients respectively. Positive label prevalence in the development set (training and validation sets) for the 4 tasks is shown in Table 1.

4.2. Model performance

Table 1 shows the performance of the LR and FNN models for all 4 tasks in terms of ROC-AUC and PR-AUC. Note that given the imbalance of the data (positive labels were overwhelmingly underrepresented), PR-AUC is more informative (Saito and Rehmsmeier (2015)).

The best performance was obtained by the FNN model architecture, with the exception of the diabetes task, where it was outperformed by the linear model. In both cases, the emergency task achieved the best performance, and did significantly better than the 3 remaining tasks. This may be explained by the significantly higher label prevalence of the emergency task.

4.3. Stability analysis

In order to conduct the model stability analysis described in Sec.3.4, for each task and model architecture we trained $N = 10$ models using the same model hyperparameters and training data. Differences in model pa-

rameters and predictions across model runs may only be explained by nondeterminism in the training algorithm and hardware. Note that the choice of N was based on visual inspection of the convergence of the stability metrics as we increased N. We determined that increasing it beyond $N = 10$ would not have provided any additional value.

First, we computed the standard deviation of the ROC-AUC and PR-AUC metrics across the N model runs. This is depicted in Fig.1, which shows higher variability for the FNN architecture, with the exception of the PR-AUC metric of the diabetes model. Also, note that all PR-AUC standard deviation metrics are smaller than 10^{-2}, and these are difficult to contextualize and interpret. Hence, the standard deviations of the ROC-AUC and PR-AUC metrics may be inadequate for the assessment of model stability.

A better assessment of model stability may be provided by the pairwise Jaccard index $J(K)$, which is shown in Fig.2 for different values of K. By illustrating average differences in top K patient lists, it provides a more interpretable stability assessment. Of note is that the linear architecture achieves the best stability when compared to the FNN one, for all tasks and values of K. In addition to this, note that $J(K)$ decreases inversely with K. This suggests that stability is less of an issue for large values of K, but when
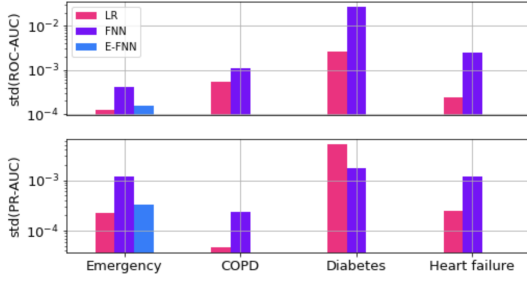
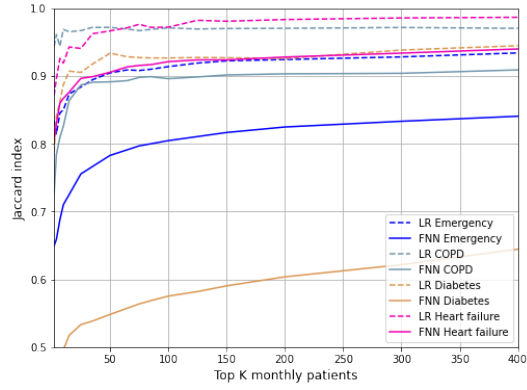Figure 1: Standard deviation of the ROC-AUC and PR-AUC metrics.

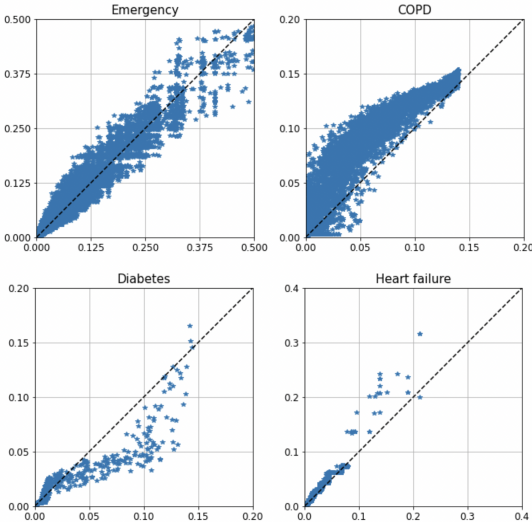Figure 2: Average pairwise Jaccard index $J(K)$ for all models.

Figure 3: Comparison of calibrated risks produced by two FNN models.

Task	Model	Unweighted τ	Weighted τ
Emergency	LR	0.9647 (0.0027)	0.9912 (0.0029)
	FNN	0.9314 (0.0026)	0.9817 (0.0032)
	EFNN	0.9796 (0.0017)	0.9947 (0.0008)
COPD	LR	0.9267 (0.0214)	0.9864 (0.0049)
	FNN	0.9540 (0.0013)	0.9831 (0.0020)
Diabetes	LR	0.9060 (0.0188)	0.9797 (0.0037)
	FNN	0.5883 (0.0700)	0.8312 (0.0704)
HF	LR	0.9870 (0.0010)	0.9950 (0.0007)
	FNN	0.8847 (0.0197)	0.9704 (0.0038)

Table 2: Average pairwise Kendall's rank correlation coefficient (τ), and standard deviations, with uniform and hyperbolic weighting.

resources are constrained and K represents a small subset of the patient population, patient lists may be impacted by instability.

Note that given that the test set is comprised of a random sample of 10% of the patients available in the STARR-OMOP dataset, a top K value of e.g. $K = 100$ in the test set would correspond to 1,000 monthly patients in the entire dataset.

A different insight into model stability is provided by Kendall's τ rank correlation coefficient, which, unlike the Jaccard index, takes into account the relative ordering of the patient risk scores. We consider both the standard Kendall's τ and one with hyberbolic weighting, where low-risk patients are downweighted (Vigna (2015)). We observed that the FNN model resulted in lower values of both weighted and unweighted τ with the exception of the unweighted τ of the COPD task, which confirms the previous insights that the LR model is more stable. In addition to this, we noted that using hyperbolic weighting resulted in larger values of τ, suggesting that ordinal relationships are more preserved for the highest ranked patients.

Finally, we explore the differences in calibrated risk scores. Fig.3 compares two randomly selected models for each task. Of note is that an individual's calibrated risk score

can vary significantly due to training variability. For the emergency model, across all pairwise comparisons we observed that the highest ranked patients experienced the largest variability (see Fig.3, top left, for an example). This behavior was occasionally present in the disease-specific tasks, but the risks of the mid-ranked patients varied more more as illustrated in the comparison of the COPD and diabetes tasks.

4.4. Ensemble

We trained $N = 10$ averaging ensembles, each one comprised of $M = 10$ models whose final outputs are averaged. Due to the cost of training $N \times M$ models, we focused our analysis on the FNN emergency model only.

Ensembling did not significantly impact the average ROC-AUC, which was 0.9111 for both the FNN and ensembled FNN (E-FNN), but did improve the average PR-AUC from 0.0843 to 0.0848, and the variance of these AUC metrics, as shown in Fig.1. Kendall's τ (Table 2) and the Jaccard index (Fig.4) also improved significantly, surpassing the linear architecture in stability.

4.5. Fairness analysis

Finally, we investigate whether the stability issues discussed in Sec.4.3 impact the fairness of the model, specifically, whether it differentially impacted the gender and race representation in the monthly lists of top K highest-risk patients.

For each patient and prediction time, we computed the range in representation across $N = 10$ models. This is illustrated in Fig.5 for the FNN emergency model and different values of K. It shows that in addition to leading to different patient selections and individual risk scores, instability can also lead to gender and race differences in patient selection. Note that while smaller differences were observed for the Asian and Black race
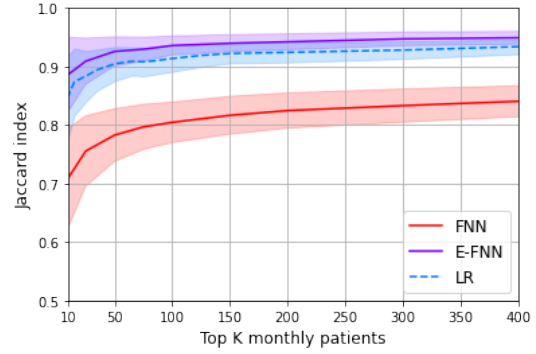

Figure 4: Average pairwise Jaccard index $J(K)$ for the LR, FNN and E-FNN emergency models. The shadow represents one standard deviation computed over all prediction times.
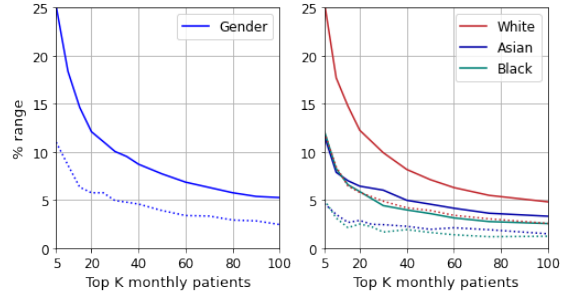

Figure 5: Range in gender (left) and race (right) composition of the top K patients for the N runs of the FNN (solid lines) and E-FNN (dotted lines) emergency models.

categories, that is explained by their reduced prevalence in the dataset (see Sec.4.1).

Finally, we also noted that the E-FNN reduced both gender and race variability. For example, for $K = 30$, the range in gender across the N FNN model runs was 10.04%, whereas the range across the N E-FNN model runs (where each E-FNN model was derived from $M = 10$ independent FNN models) was 4.94%. Similarly, the range in race composition decreased with the ensemble from 9.92% to 4.88%, 6.02% to 2.44%,

and 4.42% to 1.67% for white, asian and black races respectively.

5. Discussion

This work examined the instability of risk stratification models resulting from nondeterminism in the training process. Using a set of unplanned hospital admission prediction tasks as a case study, we confirmed previous observations that repeated training runs can lead to significantly different predictions at a patient level even if traditional model accuracy metrics remain unchanged.

We have proposed two similarity metrics to quantify this run-to-run variability among risk stratification models in a way that accounts for real-world usage under resource constraints. The first one, the Jaccard index, evaluates the consistency in patients selected for intervention when the model is used to allocate limited resources. The second one, Kendall's rank correlation coefficient, evaluates whether the ordinal relationships are preserved; therefore, it is broadly useful to evaluate models that prioritize patients for intervention. These stability metrics can also be used to select the most stable models.

We have also shown that the observed instability can lead to variance in the representation of subgroups among the top K predictions, which may have implications for equitable allocation of scare resource and could in turn diminish trust in the models.

Finally, we have evaluated a simple strategy (model ensembling) to mitigate model instability. There are other potential strategies, such as implementing techniques for addressing data imbalance (e.g. synthetic minority oversampling or weighted loss functions), decaying the learning rate, replacing the ReLU activation function with SmeLU (Shamir et al., 2020) etc., that can also improve the model stability. However, producing a throughout review of strategies to improve stability is not in the scope of this work; instead, we want to provide an example to show how stability issue can be mitigated. Our averaging ensemble improved all stability metrics considered while preserving overall model performance. However, this benefit of creating an ensemble comes with an extra cost of multiple model runs in memory and disk space. Future work should consider more efficient ensemble approaches, such as batch ensembling (Wen et al., 2020).

The acceptable level of instability will vary among different application use cases depending on a complex set of factors such as capacity to intervene, subgroup representation, and frequency of training runs; we do not make any claim as to whether any specific models we have examined are stable enough for particular applications. However, we believe that model instability has been an overlooked factor that should receive more consideration when deciding whether to deploy a model to a healthcare setting. In order for potential users to be able to make informed decisions, researchers need to transparently report and discuss the stability of their proposed methods. We hope that our work will lead to more transparency and that it will start a discussion about the importance of model stability in the research community.

Acknowledgments

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath

Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

Agency for Healthcare Research and Quality. Prevention quality indicators technical specifications - version 6.0, september 2016. https://www.qualityindicators.ahrq.gov/modules/pqi_techspec.aspx, 2016.

Agency for Healthcare Research and Quality. Prevention quality indicators technical specifications updates - version v2018 and v2018.0.1 (ICD 10-CM/PCS). https://www.qualityindicators.ahrq.gov/Modules/PQI_TechSpec_ICD10_v2018.aspx, 2018.

Andrew L Beam, Arjun K Manrai, and Marzyeh Ghassemi. Challenges to the reproducibility of machine learning models in health care. *JAMA*, 323(4):305–306, January 2020.

Duane Bender and Kamran Sartipi. HL7 FHIR: An agile and RESTful approach to healthcare information exchange, 2013.

Ronald Cornet and Nicolette de Keizer. Forty years of SNOMED: a literature review. *BMC Med. Inform. Decis. Mak.*, 8 Suppl 1:S2, October 2008.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D Sculley. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.*, 2020.

Somalee Datta, Jose Posada, Garrick Olson, Wencheng Li, Ciaran O'Reilly, Deepa Balraj, Joseph Mesterhazy, Joseph Pallas, Priyamvada Desai, and Nigam Shah. A new paradigm for accelerating clinical data science at stanford medicine. March 2020.

Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future Healthc J*, 6(2):94–98, June 2019.

Michael W Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M Dai. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 204–213, New York, NY, USA, April 2020. Association for Computing Machinery.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR*, 9:249–256, 2010.

Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D Sculley. Google vizier: A service for

Black-Box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1487–1495, New York, NY, USA, August 2017. Association for Computing Machinery.

Caglar Gulcehre, Marcin Moczulski, Misha Denil, and Yoshua Bengio. Noisy activation functions. March 2016.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get M for free. April 2017.

Yinan Huang, Ashna Talwar, Satabdi Chatterjee, and Rajender R Aparasu. Application of machine learning in predicting hospital readmissions: a scoping review of the literature. *BMC Med. Res. Methodol.*, 21 (1):96, May 2021.

M G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

M G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239, November 1945.

Zella King, Joseph Farrington, Martin Utley, Enoch Kung, Samer Elkhodair, Steve Harris, Richard Sekula, Jonathan Gillham, Kezhi Li, and Sonya Crowe. Machine learning for real-time aggregated prediction of hospital admission for emergency patients. *NPJ Digit Med*, 5(1):104, July 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. December 2016.

S Liu, Wei Ma, R Moore, V Ganesan, and S Nelson. RxNorm: prescription for electronic drug information exchange. *IT Prof.*, 7(5):17–23, September 2005.

Zhichao Liu, Ting Li, Skylar Connor, Shraddha Thakkar, Ruth Roberts, and Weida Tong. Best practice and reproducible science are required to advance artificial intelligence in real-world applications. *Brief. Bioinform.*, 23(4), July 2022.

Daniel Lopez-Martinez, Christina Chen, and Ming-Jun Chen. Machine learning for dynamically predicting the onset of renal replacement therapy in chronic kidney disease patients using claims data. In *Applications of Medical Artificial Intelligence*, pages 18–28. Springer Nature Switzerland, 2022.

Cida Luo, Yi Zhu, Zhou Zhu, Ranxi Li, Guoqin Chen, and Zhang Wang. A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. *J. Transl. Med.*, 20(1):136, March 2022.

Pranava Madhyastha and Rishabh Jain. On model stability as a function of random seed. In *The SIGNLL Conference on Computational Natural Language Learning*, November 2019.

Kay D Mann, Norm M Good, Farhad Fatehi, Sankalp Khanna, Victoria Campbell, Roger Conway, Clair Sullivan, Andrew Staib, Christopher Joyce, and David Cook. Predicting patient deterioration: A review of tools in the digital hospital setting. *J. Med. Internet Res.*, 23(9):e28209, September 2021.

Clement J McDonald, Stanley M Huff, Jeffrey G Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, Warren

Williams, James Case, and Pat Maloney. LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clin. Chem.*, 49(4):624–633, April 2003.

Scott Mayer McKinney, Alan Karthikesalingam, Daniel Tse, Christopher J Kelly, Yun Liu, Greg S Corrado, and Shravya Shetty. Reply to: Transparency and reproducibility in artificial intelligence, 2020.

Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8): 659–666, July 2021.

Velibor V Mišić, Kumar Rajaram, and Eilon Gabel. A simulation-based evaluation of machine learning models for clinical decision support: application and analysis using hospital readmission. *NPJ Digit Med*, 4(1):98, June 2021.

Hamid Mohamadlou, Anna Lynn-Palevsky, Christopher Barton, Uli Chettipally, Lisa Shieh, Jacob Calvert, Nicholas R Saber, and Ritankar Das. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. *Can J Kidney Health Dis*, 5: 2054358118776326, June 2018.

Prabhat Nagarajan and Garrett Warnell. The impact of nondeterminism on reproducibility in deep reinforcement learning.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning, 2005.

Shangshu Qian, Viet Hung Pham, Thibaud Lutellier, Zeou Hu, Jungwon Kim, Lin Tan, Yaoliang Yu, Jiahao Chen, and Sameena Shah. Are my deep learning systems fair? an empirical study of Fixed-Seed training. In *Advances in Neural Information Processing Systems*, volume 34, pages 30211–30227, 2021.

Hude Quan, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L Duncan Saunders, Cynthia A Beck, Thomas E Feasby, and William A Ghali. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care*, 43(11):1130–1139, November 2005.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.*, 169(12):866–872, December 2018a.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H Shah, Atul J Butte, Michael D Howell, Claire Cui, Greg S Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*, 1:18, May 2018b.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, 10(3):e0118432, March 2015.

Gil I. Shamir, Dong Lin, and Lorenzo Coviello. Smooth activations and reproducibility in deep networks. 2020. doi: 10. 48550/ARXIV.2010.09931. URL https://arxiv.org/abs/2010.09931.

Cecilia Summers and Michael J Dinneen. Nondeterminism and instability in neural network optimization. In *International Conference on Machine Learning (ICML)*, July 2021.

Sebastiano Vigna. A weighted correlation index for rankings with ties. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1166–1176, Republic and Canton of Geneva, CHE, May 2015. International World Wide Web Conferences Steering Committee.

Sebastian Vollmer, Bilal A Mateen, Gergo Bohner, Franz J Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, Katherine S L McAllister, Puja Myles, David Granger, Mark Birse, Richard Branson, Karel G M Moons, Gary S Collins, John P A Ioannidis, Chris Holmes, and Harry Hemingway. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*, 368:l6927, March 2020.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning. February 2020.

Jingjing Xie, Bing Xu, and Zhang Chuang. Horizontal and vertical ensemble with deep representation for classification. June 2013.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates, 2002.

Donglin Zhuang, Xingyao Zhang, Shuaiwen Leon Song, and Sara Hooker. Randomness in neural network training: Characterizing the impact of tooling. In *Proceedings of the 5th MLSys Conference*, 2022.