Preparing a Clinical Support Model for Silent Mode in General Internal Medicine

Bret Nestor* **bretnestor@cs.toronto.edu**
University of Toronto
Vector Institute
SickKids Research Institute
LKS CHART

Liam G. McCoy* **liam.mccoy@mail.utoronto.ca**
University of Toronto
Vector Institute
LKS CHART

Amol Verma amol.verma@mail.utoronto.ca
University of Toronto
LKS CHART

Chloe Pou-Prom chloe.pou-prom@unityhealth.to
LKS CHART

Joshua Murray joshua.murray@unityhealth.to
LKS CHART

Sebnem Kuzulugil sebnem.kuzulugil@unityhealth.to
LKS CHART

David Dai david.dai@unityhealth.to
LKS CHART

Muhammad Mamdani muhammad.mamdani@unityhealth.to
University of Toronto
LKS CHART

Anna Goldenberg anna.goldenberg@vectorinstitute.ai
SickKids Research Institute
University of Toronto
Vector Institute

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

Abstract

The general internal medicine (GIM) ward oversees the recovery of ill patients, excluding those who require intensive attention. Clinicians provide full recoveries, or when appropriate, end-of-life care. We hope to eliminate unexpected deaths in the GIM ward, promptly transfer patients who require escalated care to the intensive care unit, and proactively

* Equal Contribution

address deteriorating health to minimise ICU transfers. We describe a clinical decision support system which accesses labs, vitals, administered medications, clinical orders, and specialty consults. Using an ensemble of linear, gated recurrent unit (GRU) and GRU-decay (GRU-D) models, we are able to achieve a positive predictive value of 0.71 while successfully identifying 40% of patients who will experience a future adverse event. We believe that this tool will be useful in shift scheduling and discharging patients, in addition to warning clinicians of risk of decompensation. We note the lessons we learned in transitioning from a high performing model to deployment in silent mode, and all results reported in this paper report on data immediately preceding silent mode.

1. Introduction

Since medical data has been digitised, experts have been turning to clinical decision augmentation to reduce the burden on clinicians, provide earlier disease identification, and improve the quality of care (Arcadu et al., 2019; Iizuka et al., 2020). The availability of electronic health records have enabled clinicians and computer scientists to develop early warning systems (EWS) to assess the risks of important clinical events. An EWS may consist of simple combinations of vitals such as the Modified Early Warning Score (MEWS) (Subbe et al., 2001) or the National Early Warning System (NEWS)(McGinley and Pearse, 2012), or they may use complex combinations of features like the Hamilton early warning system (Xu et al., 2015; Fernando et al., 2019). These models perform remarkably well within 24 hours, for how simple they are. For example, NEWS received an AUROC of 0.87 when detecting any cardiac arrest, ICU transfers, or death within the next 24 hours with a 1% positive class frequency (Smith et al., 2013). While these models are comprehensible, they do offer space for improvement. Clinicians want to know as early as possible and as often as necessary when an adverse event will happen. In addition we want to avoid uneccesary tests, so we want to make decisions with information that is already available from the medical record system.

Many of the improvements on these early warning systems have focused on ICU settings due to the introduction of benchmark datasets for electronic health records (Johnson et al., 2016; Pollard et al., 2018). The Targeted Real-time Early Warning Score (TREWS) boasted an AUC of 0.83 which surpassed the MEWS benchmark score of 0.73 with a 14.1% positive class frequency (Henry et al., 2015). On the ICU cohort, this model had a median time-to-event of 28.2 hours.

Driven by the need for more accurate sepsis models outside of the ICU, Sepsis Watch employed multi-task Gaussian processes and recurrent neural networks to predict the risk of sepsis in the emergency department (Sendak et al., 2019). Marcus et al. (2019) use longitudinal patient data to identify patients who could benefit from prophylactic prevention of HIV.

However, the development of predictive models for clinical outcomes among GIM patients may be significantly more challenging, as this population tends to be difficult to diagnose and treat. Approximately 7% of these patients will deteriorate clinically and either require transfer to the intensive care unit (ICU) or die during the course of admission (Verma et al., 2019). The leading cause of unplanned transfer to the ICU is a failure to recognise clinical deterioration in time to respond (van Galen et al., 2016). Consequently, there is

reasonable interest in using ML to support prediction and improve clinical monitoring among hospitalised patients (Rajkomar et al., 2018).

Prior work in existing early warning systems (Burch et al., 2008; McGinley and Pearse, 2012) is intended ED triaging or for ICU use where measurements are more abundant than in the GIM context. As such, their adoption into clinical practice in GIM wards has been extremely limited. There is a considerable need for a highly accurate, automated risk prediction system that alerts clinicians to high-risk patients in advance of clinical deterioration in the GIM context (Verma et al., 2019).

We implement an ensemble of models, including linear models and recurrent neural networks (RNNs)(Suresh et al., 2017). Prior work has indicated that high-capacity models work well in such tasks, but may require extensive tuning in order to achieve improved performance over lower-capacity models when there is a large class imbalance as experienced in the GIM context (Che et al., 2018). We further work to demonstrate the performance stability of models over time, e.g., as hospital policies change and populations shift (Nestor et al., 2019).

Beyond this we seek to make the procedure of transitioning models to silent deployment familiar and comfortable to the machine learning community. While efforts have been made to make deep learning models transportable (Rajkomar et al., 2018), it is still largely up to individual healthcare providers to forge their own way to implementing machine learning models in their practice. Because of the nuances between EHR systems, patient cohorts, healthcare objectives, and other underlying non-stationarity factors, the translation of machine learning models into practice has straggled (Subbaswamy and Saria, 2019). In addition, we have been warned that machine learning models must be integrated into clinical workflow in order to garner support from clinicians (Sendak et al., 2019; Wiens et al., 2019; Sendak et al., 2020). Objectives that are pursued in benchmark machine learning for health tasks may not coincide with what is needed in the clinic. In addition Wiens et al. (2019) highlight that diligence must be provided to ensure models are ethically sound, and that results from the model should be thoughtfully reported before moving into deployment phase. Not only does communication have to be clear within the machine learning community but between clinicians and patients as well (Sendak et al., 2020).

In consideration of these lessons, this paper walks through essential steps that were taken between knowing the AUC of our state-of-the-art models and establishing a protocol for raising alarms for patients in practice. Special emphasis is placed on the nuances of result reporting to bridge the gap between clinical requirements and optimisation objectives.

2. Clinical Setting and Objective

Our objective is to improve the quality of patient care by identifying the deterioration of patients in the general internal medicine ward 24 hours before an adverse event. This will allow clinicians to make timely clinical decisions, such as administering precautionary procedures, ordering additional labs, transferring to a step-up unit or ICU, or discussing compassionate care measures. We define an adverse event as an unexpected death in the GIM ward, transfer to the ICU, issuance of palliative orders, or transfer to a palliative care facility. The labels are denoted by 1, for any prediction occurring within 24 hours of an adverse event, or a 0 otherwise (healthy discharge) as shown in Table 1. We consulted with clinicians and

patients to determine a reasonable goal for the model. Clinicians said they want no more than 2 false alarms per true alarm. We concluded that we should ensure our model has a positive predictive value (PPV) of at least 0.4. This means that each time an alarm is raised from our model, 40% of these patients will actually experience an adverse outcome within the next 24 hours. This requirement is to ensure the alarm remains actionable (Tonekaboni et al., 2019) and do not contribute to alarm fatigue (Murphy et al., 2016; Mitka, 2013). A PPV of 0.4 satisfies the clinicians' request for no more than 2 false alarms per true alarm (a minimum PPV of 33%) and provides an additional factor of safety for when the model is evaluated on a prospective test distribution. After meeting this requirement we would like to detect as many deteriorating patients as possible (increase the sensitivity).

Outcome	t_{6h}	t_{12h}	t_{18h}	t_{24h}	t_{30h}	t_{36h}
Palliative orders at 35^{th} hr	0	1	1	1	1	1
ICU transfer at 13^{th} hr	1	1	1	-	-	-
Discharge at 29^{th} hr	0	0	0	0	0	-

Table 1: Example of how binary labels are created for adverse events in the general internal medicine ward.

2.1. Data Description

The data in this paper is from the general internal medicine ward at a large urban hospital collected from 2011 to 2019. The ward has ∼3000 patient encounters per year. The demographic variables for each year of the data is summarised in Table 2.

The patient cohort had a median age of 67, with 57.5% of the patients being female. 14.7% of patients were admitted to the GIM ward from the ICU. 2.8% of patients were transferred to palliative care. 12% of patients do not have housing. 99.7% of patients were covered by universal healthcare.

We use the electronic health records from these patients to create static and time-varying features for our models. One advantage to implementing models on private data is that we can pull any electronically recorded feature in the hospital. Our models use some combination of demographic data, labs, vitals, clinical orders, medication administration, nurse notes, admission notes, and ICD10 diagnosis codes for training. The variables used in the models are characterised in Table 3 (see Appendix B for an expanded list). Consideration is given for which data is available once the patient is in the ward, for example, nurse notes are recorded more frequently so they are preferred to resident notes, which may be transcribed at the end of the day. In addition, ICD10 codes are only used as multitask training targets, so they are not required for inference.

3. Methods

Our model is an ensemble of Lasso models, GAM models, a GRU model (Cho et al., 2014), and a GRU-D model (Che et al., 2018) trained on different selections of the data (See Table 4). We chose Lasso models to eliminate redundant or uninformative features in the

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	All
Total	2117	3001	3406	3268	3014	2601	2941	2917	899	24164
Female	55.55	58.85	57.08	56.98	57.56	56.25	57.84	58.69	59.96	57.52
Male	44.45	41.15	42.92	43.02	42.44	43.75	42.16	41.31	40.04	42.48
Homeless	10.11	11.56	11.30	12.21	11.88	12.00	12.27	13.68	14.91	12.04
All Events	11.43	10.53	10.22	10.50	9.69	9.46	9.59	10.04	8.12	10.08
Unexp. Death	1.46	1.27	1.17	1.04	0.90	0.96	1.02	1.13	0.56	1.09
New Pal.	3.26	3.07	2.14	3.06	2.79	3.11	2.82	2.74	3.23	2.86
ICU Transfer	6.71	6.20	6.90	6.40	6.01	5.38	5.75	6.17	4.34	6.13
All Mortality	6.42	4.90	4.90	5.17	4.61	5.23	5.41	4.77	4.34	5.09
All Palliative	3.59	3.53	2.35	2.72	2.69	2.65	2.04	2.06	1.89	2.64
From ICU	13.75	13.06	14.27	14.81	15.66	15.03	15.27	15.56	15.68	14.73
From ED	75.96	84.31	82.35	81.76	80.99	78.47	*	*	*	*
Age Median	66.96	66.16	65.19	67.64	67.42	67.13	66.74	66.97	66.40	66.74
Age (95% CI)	26,93	26,94	26,93	25,93	25,94	26,93	26,93	26,94	28,94	26,93

Table 2: The patient demographics are displayed for each year. Outcome rows are bolded. Decimal values are expressing percentages per stay except for the "Age Median" row. Data from the patients are considered from the time they physically arrive in the GIM, which is typically after hospital arrival, and after admission to the GIM. "All Mortality" includes post GIM mortality. "All Palliative" includes patients who arrived as palliative patients, or became palliative in the hospital. * Transfer rates from emergency departments were not available due to a source system change.

	Demographics	Medications	Orders	Labs	Vitals	Notes
Features	3	232	165	108	17	1
Features Used/Stay	3	8.7	7.0	28.5	10.8	1
Mean Observations/Stay	-	73.4	17.8*	110.7	117.6	3.97
Median Observations/Stay	-	40	14*	73	78	3

Table 3: This table shows the quantity and number of features available from the electronic health record. This is an underestimate of the total number of labs and vitals measured as it is aggregated into 6 hour buckets. For example, heart rate may be measured twice within 6 hours, but it would only be recorded as being measured once. Demographic features are fully observed. *Clinical orders are represented as Boolean on/off values so these results represent the mean and median of number of times an order status changes. The table can be read as *"The average patient has 7.0 different types of clinical orders throughout their stay, and the patient can expect to have their clinical orders changed 17.8 times per stay"*

data that are too large to do other feature selection on. Generalised additive models were used to model more complex relationships in data that has continuous values. We chose to use multiple separate linear models for two reasons. First, our feature space is relatively large.

By restricting the number of features in a particular training objective, we intend to remove spurious correlations that could lead to over-fitting. Second, the ability of different portions of data to directly vote in our ensemble makes our model more robust. If the distribution of a certain categorical feature were to suddenly change, only 1-3 of 8 models will be voting in the ensemble. If we had not taken this precaution it could be that 1-3 of 4 models would be making incorrect votes. To model the time-series component of EHR data we use a GRU model that has access to all of the available data. Finally we included a GRU-D model which implicitly handles missingness. This model learns relationships between the continuous labs and vitals and their interactions with fully observed demographic variables. For both recurrent models we use early stopping (Che et al., 2018) and only use a prospective validation set to prioritise architectures that mitigate the effects of temporal shift (Nestor et al., 2019). More details on hyperparameter tuning can be found in Appendix A. EHR data is recorded as asynchronous events. The data is binned into 6 hour intervals, according to the *simple imputation* method (Che et al., 2018). There were no distinct performance changes between the bins of 2, 4, 6, or 8 hours except from the increase in AUC and AUPR that is caused by the decrease in task sparsity, and decrease in missingness (for 2, 4, 6, and 8 hour bins the rate of missingness is 97.3%, 94.7%, 92.4%, and 90.3%, respectively). Ultimately, we chose 6 hour bins because so that predictions could correspond to shift changes.

Model	Features Used
GAM	Labs
GAM	Topics
GAM	Vitals
LASSO	Medication Administration
LASSO	Orders
LASSO	All
GRU	All
GRU-D	Labs and Vitals

Table 4: Models and training data used in the ensemble. A detailed list of the feature categories can be found in Appendix B

3.1. Multitask Learning

In light of the performance improvements seen in the work of Harutyunyan et al. (2019) we introduce multitask learning objectives to predict the onset of sepsis, the onset of respiratory interventions, and the identification of the most responsible ICD10 code. The multitask learning objective is reserved for the *deep* models which have shared layers (GRU and GRU-D). The performance difference between the multitask objective and the single-task objective is demonstrated on a held-out prospective test set in Table 5. Presumably, the inclusion of a balanced task, such as ICD10 (Organization et al., 1992) classification, and highly correlated tasks (sepsis and respiratory interventions) assists the model in forming features that generalise (Ndirango and Lee, 2019).

Main Target	Multitask next 24h	Single Task next 24h
All Outcomes	0.950 (0.230)	0.943 (0.185)
ICU	0.931 (0.069)	0.921 (0.049)
Palliative	0.943 (0.142)	0.920 (0.064)
Unexpected	0.953 (0.014)	0.923 (0.007)

Table 5: Performance across tasks with and without multitask training on the GRU-D model: Scores are reported as: area under receiver-operator curve (area under precision-recall curve).

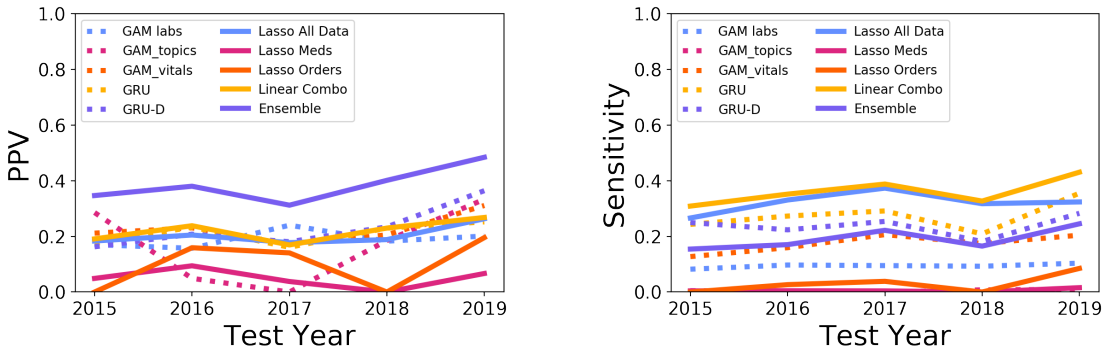
3.2. Ensembling and Implementation Details

We investigated multiple ways to combine the models. This includes linear combinations, majority voting, and random forest combinations. We demonstrate the performance of the linear combination scheme and the majority voting scheme in Figure 1. Ultimately we decided on majority-vote ensembling for its ease of auditing models. We are able to investigate which models vote, and keep track of the performance of each individual model. We set the ensemble to require 3 votes to make a prediction that an adverse event will occur. Using this information we find the thresholds at which individual models satisfy a model-specific PPV, and increment this model-specific PPV until the ensemble achieves a PPV of >0.4 (see Figure A1). This resulted in individual model achieving 0.2 PPV on the validation set. This combination of votes and model thresholds was sufficient to achieve >0.4 PPV for the ensemble on the validation set. One interesting finding is that ensembling simple models with complex models may improve overall performance. For example, the linear combinations of models outperform any one model on its own. This suggests that there is no one-size-fits-all model for EHR data. Restricted capacity models can still improve overall model performance (Zhang et al., 2019).

4. Measuring Successes and Failures

Quite simply, we measure success by how often we are able to avoid unnecessary deaths or provide palliative care as needed, while simultaneously avoiding unnecessary ICU transfers. However, we would like to quantify these with machine learning metrics.

Time-to-event As shown in Table 1, the task is formulated such that the label is "1" if we are within 24 hours of an adverse event or "0" otherwise. However, this poses several issues for evaluating machine learning models. GIM patients stay in the hospital for a relatively long time. Because of this the likelihood of an adverse event at any moment in time for any patient is incredibly rare. A model which randomly samples when patients will live and die from the task frequency distribution will still have an accuracy of 96.48% since our event only occurs 1.79% of the time. When questioning what the PPV of our model is, we are really asking, "What is the PPV of our model in the next 24 hours?". However, if an alarm sounds, and a patient is to experience an outcome, the outcome most likely will not occur until more than 24 hours have elapsed (see Table 6). All of these predictions may be timely and actionable, but are recognised as false positives in our evaluation criteria. On the other hand,



(*a*) The positive predictive values for each of the constituent models, and the ensembles as a whole on prospective test sets.

(*b*) The sensitivity for each of the constituent models, and the ensembles as a whole on prospective test sets

Figure 1: Model performance on prospective EHR data in general internal medicine for early detection of decompensation. A positive prediction is only correct if an outcome is to occur in the next 24 hours.

Target	Time-to-Event after Prediction		
	mean (h)	median (h)	sample size
All Outcomes	48.67	30	27
ICU Transfer	22.50	18	<10*
Palliative Orders	65.25	36	16
Unexpected Death	30.00	42	<10*

Table 6: Time to event after a prediction for true positive predictions in the test set. Times are reported as mean, median. Note that these distributions are not Gaussian. Unexpected death is defined as death without comfort measure orders. * The exact number is redacted.

predicting an adverse outcome several days in advance may not be actionable, and should be considered a false positive. Nonetheless, an alarm will occur more than 24 hours before an adverse event for over half of our patients who receive a true positive alarm (Table 6). It is easier to communicate with clinicians when it is phrased as, "When we predict that a patient is decompensating, how often do they actually decompensate *any time* after the alarm?". When we disregard the time until the event, our performance on the prospective test set appears to improve (See Table 7). The increase in PPV can be attributed to the reduction of false positives relative to the number of true positives. Some of these false positives can be attributed to predictions that are made more than 24 hours prior to an adverse event occurring. The sensitivity is increased when we evaluate on a per-stay basis because only 1 prediction is required from all of the prediction windows prior to an event for the prediction to be considered a true positive. This increases the number of true positives while simultaneously decreasing the number of false negatives.

Metric	Per Window	Per Stay
Count	21603	902
PPV	0.48	0.71
Sensitivity	0.25	0.40

Table 7: Ensemble performance on a held out prospective test set. The *per window* aggregation looks at the performance for each 6-hour window of each patients' stay. The *per stay* aggregation evaluates performance using the maximum number of ensemble votes at any time prior to the event. PPV is the positive predictive value, which is calculated as $True\ Positives/(True\ Positives + False\ Positives)$. Sensitivity is calculated as $True\ Positives/(True\ Positives + False\ Negatives)$.

This informs the design of our alarm protocol. When patients which have not been at-risk transition to being at-risk the care team is alerted by sending a push notification to the resident. When a patient was previously at risk and is consecutively predicted to be at risk, no alarm is given, but their EHR still indicates that they are at high risk. We chose to integrate model predictions in the EHR so that nurses can allocate staff more effectively at the start of shifts, and all clinicians can diligently monitor the patient. Out of the 26 patients who experience true alarms in our test set, patients receive an average of 4.46 model votes, resulting in 1.31 alarms (See Table 8). On the other hand the 12 patients who received false alarms only received an average of 1.33 model predictions, resulting in 1.25 alarms. It is worth noting that $PPV \neq 1 - false\ alarm\ rate$ in this alarm protocol. If we consider this alarm scenario, we would like to know how much time remains before event will occur. Table 6 shows the median time-to-event for ICU transfers is quite rapid as compared to palliative orders or unexpected death. This is promising because it demonstrates that clinicians can have conversations with patients about their wishes to pursue comfort care earlier than what is currently being done. Other patients who do not seek comfort measures, can be closely monitored and promptly transferred to the ICU if necessary. Of the 8.1% of patients that experience an adverse outcome, 40% of them (sensitivity) will receive this improvement in care. That corresponds to 2 patients per week in our clinical care setting. In order to understand how we can make these improvements more applicable for patients who decompensate, we must look at the failures of the model.

4.1. Analysis of False Positive Predictions

In this section, we describe anecdotal false positives from our model, specifically when we predicted a poor outcome and the patient did not get transferred, die or, otherwise experience a adverse event.

Models learn notions of generalised state, which may not apply universally to patient outcomes - particularly when care preferences come into play. Clinicians have a goal of preventing patients from deteriorating. Their interventions on critically ill patients, whom our model would predict as high risk of an adverse event, may actually prevent that adverse event from occurring. When false positives are observed, it is hard to determine if this is attributable to model failure, or if the clinicians really saved the life

Alarm style	Alarms/Stay	TP	Alarms/TP	FP	Alarms/FP
Regular	0.058	29	1.28	12	1.25
+1 vote	0.053	27	1.30	11	1.18
+1 vote & silencing	0.053	27	1.30	11	1.18

Table 8: Alarm characteristics on our test set for three alarming protocols. The *Regular* procedure involves sounding an alarm every time a patient switches from not at risk to at risk. The *+1 vote* protocol requires an extra vote to fire in the first 30 hours. The *+1 vote silencing* protocol requires an extra vote to fire in the first 30 hours and it silences all other alarms for the next 24 hours

of a deteriorating patient. Some of the 12 false positive patients in the test set were very ill, including patients who were transferred from the ICU, experiencing sepsis, intubated, unresponsive, and/or visited by the critical care response team around the time of the alert. Others seemed to be recovering well and were truly false alarms. Perhaps if we desired greater sensitivity in our model we could sacrifice some of the PPV. This would be equivalent to telling physicians more of what they already know about their patient. In some instances false positives can even be beneficial. They may even reaffirm the clinicians that they are performing life-saving work.

Predictions are accurate early-on, and generally perform with stability as stay lengths increase. Figure 2 shows the abundance of each outcome in the training set vs. the length of stay. Patient discharges are nearly exponentially distributed with respect to the length of stay (though patients are not as likely to be discharged in the first 24 hours since they were sick enough to admit in the first place). ICU transfers, new palliative orders, and unexpected deaths are approximately proportionate to discharges throughout the stay. The exception is that in the first few days of a patients' stay in the GIM ward an ICU transfer is more likely and palliative orders are less likely. Models are more precise but less sensitive in stays that conclude prior to 4 days (the median length of stay) than they are afterwards (See Table 9). We investigated increasing the number of votes in the ensemble during the first 30 hours to increase our model selectivity while hopefully not having much effect on sensitivity. After seeing several patient trajectories that oscillate around the alarm threshold we also implemented a 24 hour silencing period. If an alarm is sent for a patient, another alarm will not be sent in the next 24 hours even if the patient transitions back to low risk then again to high risk. These alarm scenarios are simulated for the 902 patients in our test set in Table 8. For these 902 patients spanning a 4 month period, our models had 27 actionable alarms (approximately 1 every 4 days) and 11 false alarms (approximately 1 every 11 days).

Unexpected death is a difficult label to confirm. Since we define unexpected death as a death in the GIM ward without comfort measures, it is entirely possible that palliative patients decline comfort measures and palliative treatment for a variety of reasons (Van Kleffens and Van Leeuwen, 2005). While we are capturing both palliative orders and unexpected deaths in our outcome, it is worth noting that the number of unexpected deaths we detect, may actually be cases of patients refusing comfort measures. It would be difficult to determine

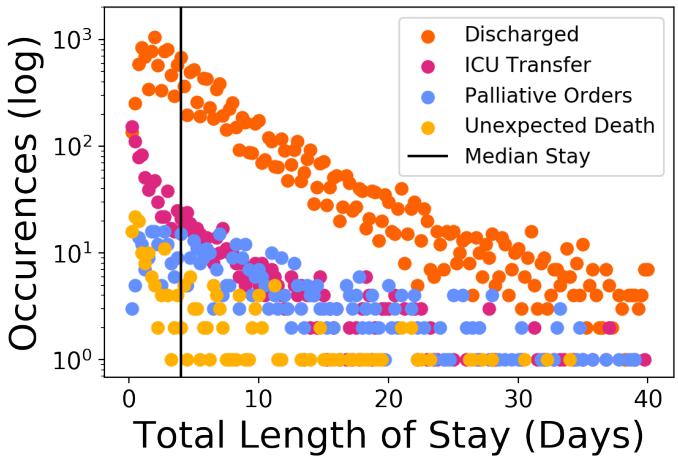


Figure 2: Occurrence of each outcome plotted against the time spent in the general internal medicine ward in 2011 through 2017. Counts are aggregated by 6 hour windows from the time a patient is admitted to the ward.

Metric	$<$ Median	$\geq$ Median
Population	443	459
PPV	0.833	0.655
Sensitivity	0.278	0.514

Table 9: Ensemble performance before and after the median stay. The positive predictive value (PPV) is reported for 443 patients who leave the general internal medicine before the median stay of 4 days and for the 459 patients who stay as long or longer than the median stay. The sensitivity indicates how many of the patients who experience an adverse outcome get a prediction from the model.

whether or not the death was truly unexpected without a thorough comprehension of the case. However, from clinician anecdote, deaths without comfort measure orders is a close proxy for unexpected death.

Models are robust to those who do not comply with medical advice. We suspected that some false positives may be due to sick patients leaving against medical advice. Between 2011 and 2018, 2% of patients left against medical advice. Despite this, the 11 false positives (in the +1 vote scheme) from our test set do not indicate that the patients left against medical advice. It is reasonable to conclude that patients who leave against medical advice do not contribute to the error in our model.

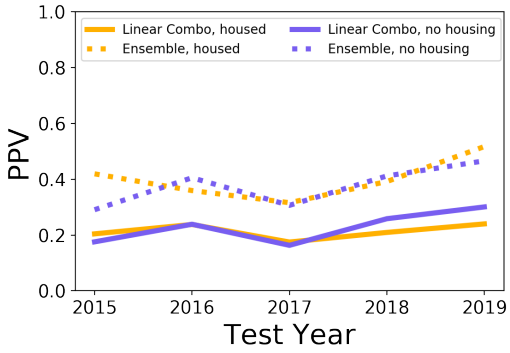
Predictions are generally fair across demographic attributes. Homelessness has been historically related to healthcare disparities, especially in the US (Kushel et al., 2001). However, even in regions with universal healthcare, disparities in healthcare for homeless individuals have been recorded (Cheallaigh et al., 2017). For example, in Ireland homeless patients are more likely to leave against medical advice(Cheallaigh et al., 2017). In Canada, homeless patients have higher readmission rates than patients with reliable housing (Saab et al., 2016). Figure 3a & 3b show the model PPV and sensitivity (per patient per hour) for homeless patients and housed patients when trained annually and tested on prospective datasets. Fortunately, the models tend to be relatively matched for both housed and homeless patients. Disparities in sensitivity tend to be made up for in PPV.

Afterwards we look at model predictions between sexes. Bias between sexes has been noted before in EHRs (Nestor et al., 2019; Gianfrancesco et al., 2018), clinical notes (Zhang et al., 2020), and medical images (Seyyed-Kalantari et al., 2020). Women tend to suffer from healthcare disparities (Hoffmann and Tarzian, 2001; Li et al., 2016). One approach to reduce bias in datasets is to collect more data so that protected groups are better represented in the training data(Chen et al., 2018). In our cohort 57.5% of the patients are female. We see relatively balanced performance between sexes in Figure 3c. When models falter in PPV for a particular sex in any year, they tend to have a higher sensitivity.

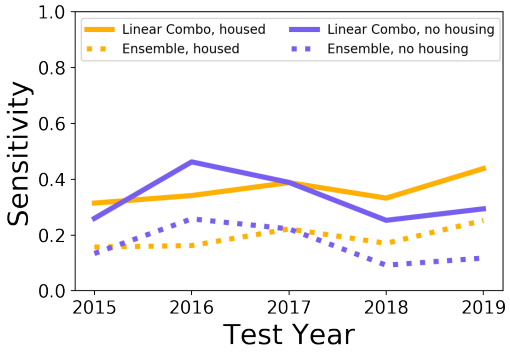
Low risk patients do not experience many adverse outcomes. We took advantage of our ensembling technique to investigate patients who are safe to discharge. Of the 902 patients in the test set, 829 patients ended their stays with no models voting in the ensemble. We deem these patients as *low-risk patients*. 97.7% of all patients who were discharged were at low risk in the last 6 hours of their stay. Coincidentally, the number of discharges was equal to the number of low risk patients (whose stays could end in discharge or an adverse event). 2.3% of patients were categorised as low risk immediately before experiencing an adverse event. Remarkably, each of these patients had received at least 1 alarm prior to experiencing the adverse event. This indicates that the number of models voting in an ensemble may be a good proxy for safe patient discharge, but should be trusted less when patients have previously experienced an alarm.

5. Generalised Insights

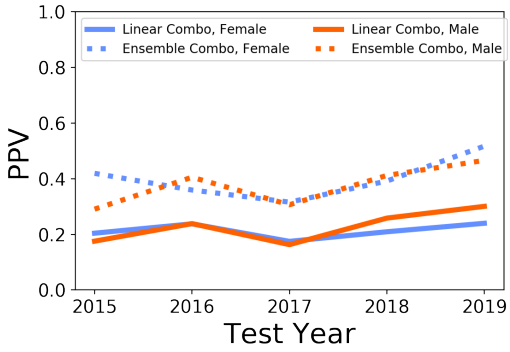
Being able to provide models with high AUC and AUPR is not sufficient in delivering models into clinical practice. We have provided an illustrated case study of the requirements of machine learning practitioners in preparation to trialling machine learning models (Sendak



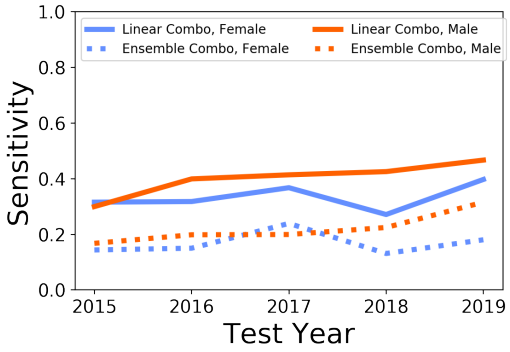
(*a*) The positive predictive value of linear combinations of the models (Linear Combo) and majority vote ensemble models (Ensemble) for both housed and homeless patients.



(*b*) The sensitivity of linear combinations of the models (Linear Combo) and majority vote ensemble models (Ensemble) for both housed and homeless patients.



(*c*) The positive predictive value of weighted combinations of the models (Linear Combo) and majority vote ensemble model (Ensemble) for female and male patients.



(*d*) The sensitivity of weighted combinations of the models (Linear Combo) and majority vote ensemble model (Ensemble) for female and male patients.

Figure 3: The performance of ensemble models on homeless and housed patients for each year (a & b) as well as for female and male patients for each year (c & d). Models are trained on historic data (i.e. 2011-2012) and validated on a prospective year of data (i.e. 2013) then tested on yet another prospective year of data (i.e. 2014).

et al., 2019). We emphasise the unique challenges in working with EHR data and our approaches to overcoming them. We must carefully consider how we communicate performance metrics. While clinicians may ask to have no more than a 60% false alarm rate, that does not always directly translate to a metric for actionable alarms. An alarm that occurs 27 hours before an adverse event might well be just as actionable as an alarm that occurs 23 hours before an event. Despite this the model would be reporting a false positive alarm. However, an alarm that occurs 5 days prior to an adverse event may not be actionable. It is equally as important to communicate the time-to-event from first alarm as it is to report the false alarm rate. In some cases, the alarm may be misleading. If a patient is obviously sick, then alarm is too late and it does not deliver any additional information. While this is not a false alarm, it is still a nuisance alarm. Conversely, patients who do not experience an adverse event may still be critically ill and may require urgent intervention. These patients are still experiencing a physiological decompensation, however they do not experience one of our proxy labels for decompensation. In summary, the GIM ward has historically experienced 5.6 adverse events per week (1 unexpected death every 11.5 days, 1 new palliative patient every 4.4 days, and 1 ICU transfer every 2 days). This model positively identifies one of these events once every 4 days, and falsely alarms clinicians once every 11 days.

We can use knowledge of the hospital operating procedures to inform alarm procedures. We opt to complete predictions every 6 hours because it corresponds with staff shift changes. Patients who are newly admitted require an additional vote to compensate for the fact that the recurrent models do not have much historical data to work with. Likewise patients who have received an alarm in the previous 24 hours will not receive an additional alarm because their condition will already be known to the care team.

6. Conclusion

Developing machine learning models for clinical use requires us to go beyond AUROC improvements on benchmark datasets. We are trying to walk the line between *"It hardly ever detects patients that deteriorate"* and *"Most of the patients it alarms do not need urgent treatment"*. Striking a balance between these two is essential in meeting the expectations of clinicians. In a future work we will investigate the overlap of the machine learning model vs. clinicians, and the mitigation of adverse events of the model during deployment vs. silent mode. We present this work as a reminder that there is essential work to be done to adapt machine learning for health research for clinical utility.

References

Filippo Arcadu, Fethallah Benmansour, Andreas Maunz, Jeff Willis, Zdenka Haskova, and Marco Prunotto. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ digital medicine*, 2(1):1–9, 2019.

Yoshua Bengio. Rmsprop and equilibrated adaptive learning rates for nonconvex optimization. *corr abs/1502.04390*, 2015.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

VC Burch, G Tarr, and C Morroni. Modified early warning score predicts the need for hospital admission and inhospital mortality. *Emergency Medicine Journal*, 25(10):674–678, 2008.

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): 6085, 2018.

Clíona Ní Cheallaigh, Sarah Cullivan, Jess Sears, Ann Marie Lawlee, Joe Browne, Jennifer Kieran, Ricardo Segurado, Austin O'Carroll, Fiona O'Reilly, Donnacha Creagh, et al. Usage of unscheduled hospital care by homeless individuals in dublin, ireland: a cross-sectional study. *BMJ open*, 7(11):e016420, 2017.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550, 2018.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Shannon M Fernando, Alison E Fox-Robichaud, Bram Rochwerg, Pierre Cardinal, Andrew JE Seely, Jeffrey J Perry, Daniel I McIsaac, Alexandre Tran, Steven Skitch, Benjamin Tam, et al. Prognostic accuracy of the hamilton early warning score (hews) and the national early warning score 2 (news2) among hospitalized patients assessed by a rapid response team. *Critical Care*, 23(1):60, 2019.

Agency for Healthcare Research and MD. Quality, Rockville. Clinical classifications software (ccs) for icd-10-pcs (beta version)., Nov 2019. URL www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11):1544–1547, 11 2018. ISSN 2168-6106. doi: 10.1001/jamainternmed.2018.3763. URL https://doi.org/10.1001/jamainternmed.2018.3763.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1): 96, 2019.

Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science Translational Medicine*, 7(299): 299ra122–299ra122, 2015.

Diane E Hoffmann and Anita J Tarzian. The girl who cried pain: a bias against women in the treatment of pain. *The Journal of Law, Medicine & Ethics*, 28:13–27, 2001.

Osamu Iizuka, Fahdi Kanavati, Kei Kato, Michael Rambeau, Koji Arihiro, and Masayuki Tsuneki. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific Reports*, 10(1):1–11, 2020.

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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Margot B Kushel, Eric Vittinghoff, and Jennifer S Haas. Factors associated with the health care utilization of homeless persons. *JAMA*, 285(2):200–206, 2001.

Shanshan Li, Gregg C Fonarow, Kenneth J Mukamal, Li Liang, Phillip J Schulte, Eric E Smith, Adam DeVore, Adrian F Hernandez, Eric D Peterson, and Deepak L Bhatt. Sex and race/ethnicity–related disparities in care and outcomes after hospitalization for coronary artery disease among older adults. *Circulation: Cardiovascular Quality and Outcomes*, 9 (2_suppl_1):S36–S44, 2016.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

Julia L Marcus, Leo B Hurley, Douglas S Krakower, Stacey Alexeeff, Michael J Silverberg, and Jonathan E Volk. Use of electronic health record data and machine learning to identify candidates for hiv pre-exposure prophylaxis: a modelling study. *The Lancet HIV*, 6(10): e688–e695, 2019.

Ann McGinley and Rupert M Pearse. A national early warning score for acutely ill patients. *BMJ*, 345, 2012. doi: 10.1136/bmj.e5310. URL https://www.bmj.com/content/345/bmj.e5310.

Mike Mitka. Joint Commission Warns of Alarm Fatigue: Multitude of Alarms From Monitoring Devices Problematic. *JAMA*, 309(22):2315–2316, 06 2013. ISSN 0098-7484. doi: 10.1001/jama.2013.6032. URL https://doi.org/10.1001/jama.2013.6032.

Daniel R Murphy, Ashley ND Meyer, Elise Russo, Dean F Sittig, Li Wei, and Hardeep Singh. The burden of inbox notifications in commercial electronic health records. *JAMA Internal Medicine*, 176(4):559–560, 2016.

Anthony Ndirango and Tyler Lee. Generalization in multitask deep neural classifiers: a statistical physics approach. In *Advances in Neural Information Processing Systems*, pages 15836–15845, 2019.

Bret Nestor, Matthew McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. *arXiv preprint arXiv:1908.00690*, 2019.

World Health Organization et al. *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization, 1992.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

Dima Saab, Rosane Nisenbaum, Irfan Dhalla, and Stephen W Hwang. Hospital readmissions in a community-based sample of homeless adults: a matched-cohort study. *Journal of General Internal Medicine*, 31(9):1011–1018, 2016.

Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. " the human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 99–109, 2020.

MP Sendak, W Ratliff, D Sarro, E Alderton, J Futoma, M Gao, M Nichols, M Revoir, F Yashar, C Miller, et al. Sepsis watch: A real-world integration of deep learning into routine clinical care. *JMIR Preprints*, 15182, 2019.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, and Ghassemi Marzyeh. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *arXiv preprint arXiv:2003.00827*, 2020.

Gary B Smith, David R Prytherch, Paul Meredith, Paul E Schmidt, and Peter I Featherstone. The ability of the national early warning score (news) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, 2013.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, 11 2019. ISSN 1465-4644. doi: 10.1093/biostatistics/kxz041. URL https://doi.org/10.1093/biostatistics/kxz041.

CP Subbe, M Kruger, P Rutherford, and L Gemmel. Validation of a modified early warning score in medical admissions. *QJM*, 94(10):521–526, 2001.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2017.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. *arXiv preprint arXiv:1905.05134*, 2019.

Louise S van Galen, Patricia W Struik, Babiche EJM Driesen, Hanneke Merten, Jeroen Ludikhuize, Johannes I van der Spoel, Mark HH Kramer, and Prabath WB Nanayakkara. Delayed recognition of deterioration of patients in general wards is mostly caused by human related monitoring failures: a root cause analysis of unplanned icu admissions. *PloS One*, 11(8), 2016.

T Van Kleffens and E Van Leeuwen. Physicians' evaluations of patients' decisions to refuse oncological treatment. *Journal of Medical Ethics*, 31(3):131–136, 2005.

Amol A Verma, Yishan Guo, Janice L Kwan, Lauren Lapointe-Shaw, Shail Rawal, Terence Tang, Adina Weinerman, and Fahad Razak. Characteristics of short general internal medicine hospital stays: a multicentre cross-sectional study. *CMAJ Open*, 7(1):E47, 2019.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0548-6. URL https://doi.org/10.1038/s41591-019-0548-6.

Michael Xu, Benjamin Tam, Lehana Thabane, and Alison Fox-Robichaud. A protocol for developing early warning score models from vital signs data in hospitals using ensembles of decision trees. *BMJ Open*, 5(9):e008699, 2015.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: Quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 110–120, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384448. URL https://doi.org/10.1145/3368555.3384448.

Kun Zhang, Yuan Xue, Gerardo Flores, Alvin Rajkomar, Claire Cui, and Andrew M Dai. Modelling EHR timeseries by restricting feature interaction. *arXiv preprint arXiv:1911.06410*, 2019.

Appendix A. Hyperparameter Search

All models were hyperparameter tuned from 2011-2017 and validated on 2018 data; this includes the value of k which is selected in the validation stages shown in Figure A1. This presents some bias in earlier model test set reporting as we have selected architectures that we know are suitable for 2011-2017 data.

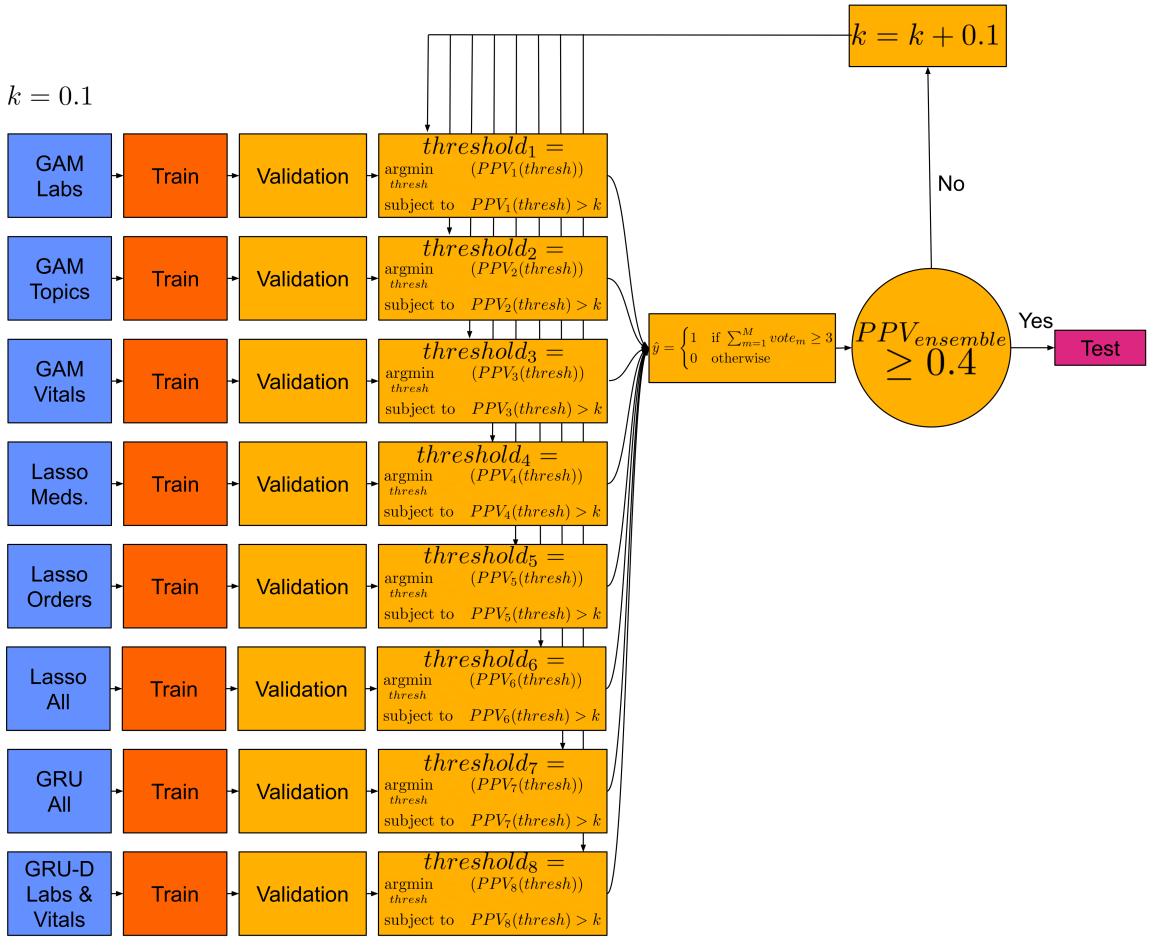


Figure A1: Training, validation, and testing protocol for the ensemble. The colours of the blocks represent which data is available at a particular step in the algorithm.

A.1. GRU

For the GRU parameters were randomly searched (Bergstra and Bengio, 2012) from the possible sets outlined in Table A1. The GRU was trained using the Adam optimizer (Kingma and Ba, 2014)with a learning rate of 0.002, betas of [0.9, 0.999], epsilon of [1e-08] and weight decay of 0. Upon finding the optimal hyperparameters, all training was done with early stopping after 5 epochs of no model improvements. Additional tasks were not included in hyperparameter searches, but are included in subsequent training.

Hyperparameter	Possible Set
Learning Rate	[0.005, 0.001, 0.0005, 0.0001]
Number of GRU Hidden Layers	[1,2,3]
GRU Hidden Layer Dimensions	[10, 20, 50, 75, 100, 150, 200]
Dropout Values	[0.005, 0.01, 0.05, 0.1, 0.25, 0.5
Maximum Number of Training Epochs	[50, 100, 150, 200,500]
Number of LDA Topics for Nurse Notes	[50, 100]

Table A1: GRU Hyperparameter Sets

A	GB
ASD	TASD

Table A2: Caption

A.2. GRU-D

GRU-D hidden layer dimension is selected from the range: [16, 256], The optimizer is selected from : [RMSProp(Bengio, 2015), Adam(Kingma and Ba, 2014), RAdam(Liu et al., 2019)] The learning rate is selected from the logrithmically distributed range of: [10e-5 to 10e-1] ICD10 is converted to a multclass task by discretising by: [the first code letter, ICD10 body systems, or HCUP groupings(for Healthcare Research and Quality, 2019)]

The GRU-D is hyperparameter tuned with early stopping after 5 epochs of no model improvements.

Appendix B. Model Features

Labs A1c., Abs.Basophils, Abs.Eosinophils, Abs.IG, Abs. Lymphocytes, Abs.Metamyelocytes M.Diff, Abs.Monocytes, Abs.Myelocytes M.Diff, Abs.Neutrophils, Acetaminophen Level, Albumin, ALP, ALT, Amylase, Anion Gap, Aortic Root, aPTT, APTT, AST, Base Excess Art., Base Excess Ven., Bicarbonate Arterial, Bicarbonate Venous, Bilirubin Direct, Bilirubin Total, Ca, Ionized (pH 7.4) corr, Ca, Ionized (pH 7.4) corr., Calcium, Calcium Ionized Arterial, Calcium, Ionized O.R. Arterial, Chloride, Chloride Random, CK, C-Reactive Protein, Creatinine, ESR, Ethanol, Ferritin, Globulin,calc., Glucose O.R. Arterial, Glucose POC (company 1), Glucose POC (company 1/company 2), Glucose POC (company 1/company 2/company 3), Glucose Random, Glucose, Syringe Arterial, HCT, Hematocrit O.R. Arterial calculation, HGB, H ion Arterial, H ion Venous, INR, Ionized Calcium, Iron Total, IV Septum, Lactate Arterial, Lactate Venous, LD, Left Atrium, Lipase, LV Diastole, LV Systole, Magnesium, MCH, MCHC, MCV, Meas. O2 Sat.Art, Meas. O2 Sat.Ven., MPV, nRBC, NT-proBNP, Osmolality, Osmolality Serum, pCO2 Arterial, pCO2 Venous, pH, pH Arterial, Phosphorus, pH Venous, PLT, pO2 Arterial, pO2 Venous, Posterior Wall, Potassium, Potassium O.R. Arterial, Potassium Random, Protein Total Serum, PT, RBC, RDW, Retics, Salicylate Level,

Saturation, Sodium, Sodium, Sodium O.R. Arterial, Specific Gravity, TIBC, Total CO2, Total Protein Serum, Troponin I Serum, TSH, Urea, Urobilinogen, Vitamin B12 Level, WBC

Vitals: Braden Score Total, Catheter, FiO2 %, Heart Rate, IV Dextrose 3.3%/NaCl 0.3%, IV Piggy Back 1, IV Sodium Chloride 0.9% (NaCL), O2 L/Min, O2 Saturation (%), Other Intake, Other Output, Pain Intensity at Rest, Pain Intensity at Rest (0-10) 1, Pain Intensity at Rest (0-10) 2, Pain Intensity with Movement, Pain Intensity With Movement (0-10) 1, Pain Intensity With Movement (0-10) 2, POC Glucose Result (mmol/L), Pulse, Resp FiO2 Percent, Respirations, sbp, Temperature (c)

Static Features: pre-GIM ICU stay, age, gender

Pre-GIM labs and vitals: These are a subset of what is available within the GIM: lab_co2, lab_cl, lab_na, lab_agap, lab_hct, lab_mch, lab_mchc, lab_rbc, lab_iwbcr, lab_mcv, lab_rdw, lab_hgb, lab_cr, lab_plt, lab_glur, lab_k, lab_mpv, lab_alymp, lab_aeos, lab_abaso, lab_amono, lab_aneut, lab_alb, lab_rpt, lab_rinr, lab_rptt, lab_tbil, lab_ast, lab_alt, lab_alp, lab_urea, lab_ca, lab_mg, lab_po4, lab_tni, lab_lip, lab_ck, lab_uuro, lab_spg, lab_ph, lab_amy, lab_vlact, lab_vhion, lab_mvsa, lab_vpo2, lab_vbe, lab_vtco2, lab_vpco2, lab_vph, vital_spulse, vital_sbpsystolic, vital_sbpdiastolic, vital_srespirations, vital_so2saturation, vital_stemperature

Orders for images (Binary): PORTABLE Chest, CT Thorax Abdomen Pelvis with Contrast, CT Thorax Abdomen Pelvis, XRAY Chest PA and LAT, US Abdomen Limited, CT Thorax Rule Out Pulmonary Embolus, MRI Head, XRAY Abdomen AP and LAT, CT Thorax, US Abdomen Complete, CT Perfusion Stroke, CT Abdomen and Pelvis, CT Head no Contrast, US Doppler Venous Lower Extremity Bilateral- R/O DVT, CT Abdomen, XRAY Pelvis AP, XRAY Abdomen, CT Abdomen Pelvis with Contrast, US Abdomen and Pelvis, Cardiolite and Persantine Scan with Rest, XRAY Lumbar Spine 3 Views, PORTABLE Abdomen, Interventional Radiological Procedure Request, CT Abdomen Pelvis Triphasic Study, MRI Spine, US Doppler Portal Vein, XRAY Chest, MRI Abdomen, CT Head with Contrast, XRAY Hip Unilateral Left AP and LAT, US Doppler Abdomen Renal Arterial or Venous, CT Head Angio (Circle of Willis), CT Thorax with Contrast, PORTABLE XRAY Abdomen AP and LAT, XRAY Deglutition Study, Bone Scan Whole Body, XRAY Knee Right AP and LAT, CT Neck, CT Extremity, XRAY Lumbar Puncture, US Face and Neck, CT Neck with Contrast, CT Head with and without Contrast, CT Head and Carotids, CT Abdomen Rule Out Renal Colic, XRAY Knee Left AP and LAT, XRAY Cervical Spine AP, Lateral and Open Mouth Odontoid View, US Guided Biopsy, XRAY Hip Unilateral Right AP and LAT, CT Pelvis, XRAY Foot Left 3 Views, XRAY ERCP, US Abdomen and Pelvis Limited, US Pelvis Transvaginal (First Trimester), XRAY Skull Orbits, CT Spine Lumbar without contrast, US Soft Tissue Unilateral, US Extremity Left (Soft Tissue), XRAY Chest 3 Views, US Extremity Right (Soft Tissue), Hemodialysis Catheter Insertion, Angiography Procedure Request, NVA Abscess Drain Abdomen, XRAY Foot Right 3 Views, XRAY Shoulder Right 3 Views, Imaging Order - CT Scan CT Head and Cervical Spine, XRAY Metastatic Survey, XRAY Hip Bilateral 4 Views, Cardiolite Persantine, V/Q Scan, XRAY Shoulder Left 3 Views, XRAY Ankle Right 3 Views, XRAY Knee Left 4 Views, XRAY NG Tube Insertion, US Doppler Liver Disease Screening, XRAY Knee Right 4 Views, XRAY Tib Fib Right 2 Views, XRAY Ankle Left 3 Views, XRAY Thoracic Spine 2 Views, MRA Brain,

CT Enterography, CT Spine Cervical without contrast, XRAY Knee Bilateral 4 Views, US Doppler Vein and Extremity Unilateral, US Doppler Vein and Extremity Bilateral, MRI Extremity Unilateral, MRI Pelvis, NVA Abscess Drain Thorax, MRA Carotids, CT Thoarax Low Dose, CT Abdominal Aneurysm, PICC Insertion Double Lumen, PICC Insertion Single Lumen, US Abdomen Limited and Pelvis Limited, MRI0733, MRI0759, MRI0701, MRI0700, MRI0824, MRI0832

Nutrition type: tube feed, regular other, oral, renal, diabetic, cardiac, NPO, Regular, clear fluids, nutrition supplement

Additional consultations: TELEMETRY ORDERS: telemetry

CONSULT ORDERS: consult_physio, consult_general, consult_stroke, consult_social, consult_speech, consult_dietitian, consult_chaplain, consult_physiotherapist, consult_acute, consult_gastroenterology, consult_respiratory, consult_occupational, consult_psychiatry, consult_wound, consult_physiotherapy, consult_geriatric, consult_pharmacist, consult_chiropodist, consult_addiction, consult_research

CARDIO ORDERS: cardio_ecg, cardio_vascularlab, cardio_echo, cardio_holter, cardio_peripheralvascular

RESP ORDERS: resp_oxygen, resp_pulmonaryfunctiontest, resp_bipapcpap, resp_respiratoryintervention, resp_chesttube, resp_ventilator

ACT ORDERS (Activity and limitations): act_sitter, act_constantcare, act_opcophysrestr, act_restrictions

CODE_ORDERS: No CPR: General Medical Care, No CPR: Comprehensive Comfort Care, Full Code, No CPR: Advanced Life Support

OPCO_ORDERS: opcociwacare, opcohephiaptt, opcoivinslow, opcoivinshi, opcochfdiuret

TRANS_ORDERS: trans_infusefrozenplasma, trans_transfusepackedredbloodcells, trans_transfuseplatelets trans_infusealbumin25, trans_transfusionother, trans_infuseivimmuneglobulin, trans_prothrombincomplexconcentratepcc

WOUND_ORDERS: wound_dressingswoundcare, wound_skincare

NEURO_ORDERS: neuro_eeg, neuro_emg

Demographic features: No housing

Medications (encoded by their index in the EHR system): med_01, med_02, med_03 ... med_n