Developing a prediction model with dynamic health data

Angie Shen[1], Reuben McCreanor[1], Matthew Phelan[2], Cara O'Brien[3], Armando Bedoya[3], Rebecca C. Steorts[1*], and Benjamin A. Goldstein[2*]

1. Department of Statistical Sciences
Duke University
Durham, NC 27705

yueqi.shen@duke.edu

reuben.mccreanor@duke.edu beka@stat.duke.edu

2. Department of Biostatistics & Bioinformatics
Duke University School of Medicine
Durham, NC 27705

matthew.phelan@duke.edu ben.goldstein@duke.edu

3. Department of Medicine
Duke University School of Medicine
Durham, NC 27705

cara.obrien@duke.edu armando.bedoya@duke.edu

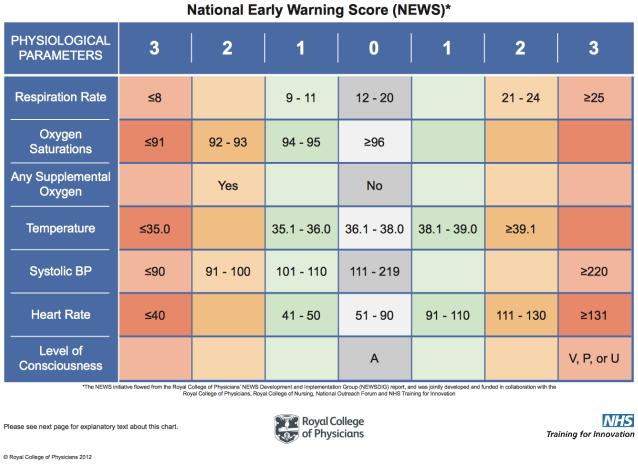
Abstract: Electronic Health Records (EHR) data constitute a relatively new data source that contain a running tally of a patient's clinical changes. As such, they are an appealing resource for clinical analysis, particularly risk prediction. While these data are potentially powerful, they inherently have a number of challenges such as many potential predictor variables, sparse and irregular measurements over time, and data that may be informatively not observed. As as result, developing robust risk models can be challenging. Using data from our institution's EHR system, we illustrate the various considerations necessary for developing a dynamic risk score for inpatient deterioration. We choose a computationally efficient time varying Cox model and show how the model can be adapted to incorporate different data complexities. We compare our results to an Early Warning Score currently implemented in the EHR system, showing that fitting a model with one's own institution's data results in better performance, even when using the same predictor variables.

1. Introduction

Early Warning Scores (EWSs) have become an important component of managing in-patient care. They provide a means to assess changes in a patient's clinical status, alerting clinicians to the need for intervention. One commonly used EWS is the National Early Warning Score (NEWS), which was designed to

*Drs. Goldstein and Steorts equally supervised this work.

detect risk of patient deterioration (Smith et al., 2013). Recently, our institution integrated automated calculation and reporting of the NEWS into its Electronic Health Record (EHR) system. NEWS was designed to be easily hand calculated and uses only a few predictor variables (see Figure 1). Moreover, as a general score, it is not optimized for our institution (or any particular) patient population. A recent internal evaluation showed that implementation of the NEWS had no meaningful impact on patient outcomes. This is not surprising, since the overall performance of the NEWS was quite low with the average Area Under the Curve (AUC) of 0.64 for the first seven days after admission. A review of EWSs has suggested that they have mixed performance (Alam et al., 2014).



National Early Warning Score (NEWS)*

PHYSIOLOGICAL PARAMETERS	3	2	1	0	1	2	3
Respiration Rate	≤8		9 - 11	12 - 20		21 - 24	≥25
Oxygen Saturations	≤91	92 - 93	94 - 95	≥96			
Any Supplemental Oxygen		Yes		No			
Temperature	≤35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥39.1	
Systolic BP	≤90	91 - 100	101 - 110	111 - 219			≥220
Heart Rate	≤40		41 - 50	51 - 90	91 - 110	111 - 130	≥131
Level of Consciousness				A			V, P, or U

*The NEWS initiative flowed from the Royal College of Physicians' NEWS Development and Implementation Group (NEWSDIG) report, and was jointly developed and funded in collaboration with the Royal College of Physicians, Royal College of Nursing, National Outreach Forum and NHS Training for Innovation

Please see next page for explanatory text about this chart.

Royal College of Physicians

NHS Training for Innovation

Fig 1: Calculation guide for the NEWS (Smith et al., 2013). Patients are assigned a set of points based on their clinical values. Typically, a score above 8 indicates risk of adverse outcomes.

EWSs such as the NEWS were not designed to fully utilize the capabilities of modern EHR systems. Instead, modern EHR systems are capable of collecting and analyzing patient data within a real time environment. For example, each time a blood pressure measurement is taken or a laboratory test is ordered this information is stored within a running catalog of the EHR system. With these data it is possible for each institution to develop its own robust risk score and not rely on simpler off-the-shelf scores. Other authors have illustrated how they can achieve increased performance by developing more sophisticated risk scores from single-center data (Henry et al., 2015; Kipnis et al., 2016).

As modern EHR systems have proliferated, there has been a steady rise in the use of EHR data for the development of risk prediction models (Goldstein et al., 2017a). While these studies have shown a number of strengths, few ($< 8\%$) of these studies considered modeling data in a time varying way. Such models are challenging to develop and implement in real time. As such, our goal in this paper is to demonstrate how to use EHR data to develop a robust risk model

for patient deterioration. We use real time EHR data (i.e. vital signs, laboratory measurements etc.) to develop a time updating risk score for our outcome. Our intention is that a patient's score will dynamically change as new information is incorporated into the health record.

One important question that arises is how one should best analyze EHR data. The primary analytic challenge is incorporating a longitudinal covariate pattern with a scalar outcome. In recent years, there has been significant development in analyzing such data, particularly around joint models for longitudinal and survival data (Rizopoulos, 2011) and Gaussian Processes to analyze individual trajectory data (Colopy et al., 2016). However, these approaches can be computationally expensive and not well suited for some EHR based data sets. Moreover, it is not clear how much added value they provide over simpler analytic approaches (Goldstein et al., 2017b; Wehbe et al., 2015). For this reason, we propose a simple time-varying covariate Cox regression model which takes in multiple covariates, is easily scalable, and estimates dynamic patient risk.

The rest of the paper proceeds as follows. Section 2 introduces the data motivating our analyses and the initial cleaning steps. Section 3 describes our analytic methods, consisting of a time-varying Cox model and multi-state models. Specifically, we propose four models which increase in complexity, incorporating more available covariates in each model. We also discuss how we set up the data in counting process and how we address the challenges of cleaning dynamic data, specifically how we handle missing values and labs. Section 4 presents the results of our analysis, including hazard ratios for different variables and comparative performance assessment. Section 5 discusses future work.

2. The Data

In this section, we describe the data, the outcome of interest, and the available predictors.

2.1. Available Data

The data are drawn from the Duke University Hospital (DUH) EHR system, an Epic Systems Corporation (EPIC) based health system, installed in late 2013. The EHR system is capable of capturing both historical (e.g. comorbidities, service history) and dynamic patient data (e.g. vitals measurements, laboratory orders). In July 2015, DUH integrated the NEWS and an associated alert system into the EHR system. While our internal evaluation of the NEWS implementation suggested little added value, to avoid potential biases due to intervention (Paxton, Niculescu-Mizil and Saria, 2013), we study patient encounters before the implementation of the alert system. Specifically, we abstracted data on patient hospital stays from July 1, 2014 – June 30, 2015. We focus on patients admitted to general medical wards, i.e. an environment where patients are relatively stable and not receiving constant monitoring as in an intensive care unit (ICU). Specifically, there were 28,932 individual patient hospitalizations during this time period.

2.1.1. Outcome of Interest

Our primary outcome, which we term *patient deterioration*, is a composite of inpatient mortality, transfer to the ICU, or call for Rapid Response Team (RRT). Our clinical collaborators chose this composite because it is believed to capture most forms of adverse outcomes and allow for the clinical team to make assessments regarding the best course of action (Smith et al., 2013). As we discuss in Section 3.1.2, we remove events that do not occur on one of the general medical wards of interest. For example, if a patient dies during a surgical procedure we do not consider this an event of interest. For analytical purposes, patients were censored either at the time of discharge or after 30 days of a hospital stay.

2.1.2. Available Predictor Variables

The NEWS only incorporates data on seven vital signs, while a typical EHR is rich regarding patient information. After consultation with our clinical collaborators we abstracted information on 3 demographics variables—age, sex, and race; 9 comorbidities from the medical history — diabetes, malignancy, chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), myocardial infarction (MI), stroke, HIV, transplant, surgery; and 10 laboratory tests—white blood cell count (WBC), magnesium (Mg), creatinine kinase (CK), blood urea nitrogen (BUN), creatine kinase - MB (CKMB), alanine aminotransferase (ALT), aspartate aminotransferase (AST), bilirubin, ammonia, D-dimer—to incorporate into the risk model.

2.2. Data Cleaning

In order to allow the data to be as close as possible to the original raw data, we employed minimal cleaning. First, we dealt with implausible vital measurements (0.15% of all vitals) by simply removing this small fraction of values. Second, two of the seven vitals indicators, consciousness level (Consciousness) and whether the patient is receiving supplemental oxygen (O2), are categorical variables. Consciousness can take five levels — alert, lethargic, responsive to pain stimulus, responsive to verbal stimulus, or unresponsive. For simplicity we combine them into two levels, alert and non-alert, as is done in the NEWS score. Also, there are two different variables indicating supplemental oxygen status. One is a binary variable indicating whether or not supplementary oxygen was given, while the other lists the type of intervention device such as face mask or nasal cannula. For simplicity, we assume that patients who have a device type listed did receive supplementary oxygen, and combine these two variables into a single binary variable. Third, we deal with the challenge of analyzing laboratory data. Due to data entry discrepancies, a single lab may have different names in the system. For example, "WBC" and "White Blood Cell Count" refer to the same lab. We identify all the common names that a lab may be be entered as in the system and extract them from the data. We also remove all implausible lab values ($< 1\%$) from the data set.

3. Data Preparation & Analytic Approach

One of the primary challenges of working with EHR data is converting the data into an analytic format that allow for modeling of the data. Since the analytic format dictates how the data need to be organized, we describe the analytic approach followed by the data processing steps. We then describe the specific models we fit and how we evaluated them (see Sections 3.2, 3.3, and 3.4).

3.1. Data Set-up

We discuss the conversion of our data into a format that one can model and then we discuss how we deal with missing values.

3.1.1. Conversion to an Analytic Format

As already mentioned, one challenge of working with EHR data is that they are not naturally collected in a format or data structure that is easy to analyze. This simple stream allows for the transfer of any piece of health information, in what are referred to as HL7 messages (Dolin et al., 2006). HL7 is an international standard for translating health information. Figure 2 shows a sample of how the raw data are extracted. Each row corresponds to a new measurement of one vital variable, recorded in irregular intervals, one measurement at a time.

```
“001”|”RESPIRATIONS”|”Resp”|”9”|8/21/2014 10:00:00|”16”
“001”|”PULSE”|”Pulse”|”8”|8/21/2014 11:00:00|”80”
“001”|”PULSE OXIMETRY”|”SpO2”|”10”|8/21/2014 11:00:00|”96”
“001”|”RESPIRATIONS”|”Resp”|”9”|8/21/2014 11:00:00|”13”
“002”|”TEMPERATURE”|”Temp”|”6”|8/12/2014 3:38:00|”97.5”
“002”|”RESPIRATIONS”|”Resp”|”9”|8/12/2014 3:38:00|”18”
“002”|”PULSE OXIMETRY”|”SpO2”|”10”|8/12/2014 3:38:00|”96”
“002”|”BLOOD PRESSURE”|”BP”|”5”|8/12/2014 3:38:00|”93/52”
“002”|”PULSE”|”Pulse”|”8”|8/12/2014 3:38:00|”58”
“002”|”RESPIRATIONS”|”Resp”|”9”|8/12/2014 7:07:00|”18”
“002”|”TEMPERATURE”|”Temp”|”6”|8/12/2014 7:07:00|”97.6”
“002”|”PULSE”|”Pulse”|”8”|8/12/2014 7:07:00|”52”
“002”|”PULSE OXIMETRY”|”SpO2”|”10”|8/12/2014 7:07:00|”97”
“002”|”BLOOD PRESSURE”|”BP”|”5”|8/12/2014 7:07:00|”103/65”
“002”|”TEMPERATURE”|”Temp”|”6”|8/13/2014 3:30:00|”98.2”
“002”|”PULSE”|”Pulse”|”8”|8/13/2014 3:30:00|”76”
“002”|”RESPIRATIONS”|”Resp”|”9”|8/13/2014 3:30:00|”18”
“002”|”BLOOD PRESSURE”|”BP”|”5”|8/13/2014 3:30:00|”110/51”
“002”|”PULSE OXIMETRY”|”SpO2”|”10”|8/13/2014 3:30:00|”98”
“002”|”TEMPERATURE”|”Temp”|”6”|8/13/2014 7:37:00|”98.1”
“002”|”PULSE”|”Pulse”|”8”|8/13/2014 7:37:00|”61”
“002”|”RESPIRATIONS”|”Resp”|”9”|8/13/2014 7:37:00|”18”
“002”|”BLOOD PRESSURE”|”BP”|”5”|8/13/2014 7:37:00|”109/50”
```

Fig 2: Raw vitals data as extracted from the EHR in HL7 format. The first column provides a patient ID. Columns 2 - 4 indicate the measurement being captured (long description, short description, code). Column 5 is a time stamp. The last column is a value. This format allows for the capture and transfer of all health information.

In order to relate the time varying covariates with our outcomes, we move the current data structure to a “counting process,” which involves creating a wider data set containing each of the unique predictor variables. Each time a new piece of clinical information is recorded, a new line is created for the patient, updating the new piece of information (e.g. blood pressure) and keeping the others the same. Note that not all clinical covariates are updated at the same time. Crucial to creating the counting process is the creation of non-overlapping *start* and *stop* time periods. As is typical in time varying survival analysis, we assume that the time intervals are closed at the start time and open at the stop time (with the section of the event at the end of the final stop time being closed). This discretization of times ensures that there is not repetition of a person. Figure 3 shows an example of the data after this transformation. More specifically, we have transformed all the covariates for person i at time j into a data matrix, and each time a covariate value changes we update this matrix.

	PAT_ENC_CSN_ID	StartTime	StopTime	timeOfEvent	event	Pulse	Resp	SpO2	Temp	SYS	Consciousness	O2
1	001	0	0.31	*NA*	0	70	16	100	37.000...	120	Alert	0
2	001	0.31	0.33	*NA*	0	91	14	97	37.000...	126	Alert	0
3	001	0.33	0.37	*NA*	0	88	16	96	36.611...	149	Alert	0
4	001	0.37	0.53	*NA*	0	88	16	96	36.611...	149	Alert	0
5	001	0.53	0.56	*NA*	0	87	20	97	36.500...	132	notAlert	0
6	001	0.56	0.58	*NA*	0	87	20	97	36.500...	132	notAlert	0
7	001	0.58	0.68	*NA*	0	87	18	97	36.500...	132	notAlert	1
8	001	0.68	0.73	*NA*	0	82	17	97	36.388...	110	notAlert	1
9	001	0.73	0.77	*NA*	0	82	20	96	36.388...	110	notAlert	1
10	001	0.77	0.79	0.79	1	82	16	96	36.388...	110	notAlert	1
11	002	0	0.02	*NA*	0	70	16	100	37.000...	120	Alert	0
12	002	0.02	0.02	*NA*	0	86	16	91	37.000...	135	Alert	0
13	002	0.02	0.03	*NA*	0	82	16	95	37.000...	134	Alert	0
14	002	0.03	0.04	*NA*	0	92	16	89	37.000...	134	Alert	0
15	002	0.04	0.04	*NA*	0	86	16	95	37.000...	126	Alert	0

Fig 3: Sample Data in transported in Counting Process format. Each time data are updated within the EHR system, a patient gets an additional row of data.

While we chose to create a new record each time a measurement was taken, we also considered binning data into regular time intervals, taking the mean value if more than one measurement was recorded, ultimately smoothing the data. We used cross validation to assess the following optimal bin size: 4-hour, 8-hour, or no binning. Our results suggested that there was no effect on the bin size with no-binning being nominally best.

3.1.2. Unobserved Values

In the data preparation process, we identified two types of missing values. The primary type occurs after a patient is first admitted into the hospital but before any measurements are taken. The secondary type occurs after the first measurement. For the vitals variables, 96% of the patients received a first measurement within the first 4 hours after admission to hospital. After measurements were taken regularly, the median time to next measurement was 1.9 hours. We use different imputation approaches to handle the aforementioned types types of missing data. Because vitals measurements during the first time bin in the counting process—from a patient's time of admission to time of the first observation—are unobserved, we impute these with values considered to be the most normal from the NEWS table, having observed that the median values of our vitals data are very similar to the median values of the normal range from the NEWS table. When a measurement is not updated, we use the Last Observation Carried Forward (LOCF) imputation approach, where missing values are replaced by the previous complete vital reading, and we assume vitals measurements then remain unchanged until there is a new reading.

On the other hand, laboratory tests are handled differently. First, some laboratory test are routine and typically ordered for all patients, while other tests (CK, CKMB, Bilirubin, Ammonia, D-dimer) are quite informative, i.e. a doctor only orders these tests if s/he suspects something is wrong. Therefore we did not want to impute in unobserved laboratory results. Second, there are two time stamps associated with a laboratory test, when the lab is ordered and when

the results are available. For this reason we considered coding the laboratory results in two ways. For those where the ordering is informative we created a counting variable for each time a laboratory test is ordered. The variable starts with the value 0 and increments by one whenever a new test is ordered. Then for all the labs we created a four level categorical variable of: No Value, Normal, High, Low. Our clinical collaborators specified clinical cut-offs for normal ranges. While there is potentially some loss of information in discretizing the laboratory results, this allows us to naturally handle time-varying capture of the labs without having to impute missing data.

3.2. Time-Varying Cox Model

To develop our risk model, we used a time-varying covariate Cox model (Cox, 1972).

The Cox proportional hazards model is a semi-parametric regression model with time varying covariates that allows the measurement of time to events. It is composed of a non-parametric hazard function and parametric vector of covariates. The hazard for individual i is

$$\lambda_i(t) = \lambda_0(t)e^{x_i(t)\beta}, \tag{1}$$

where x is a vector of the covariates, β is a vector of coefficients for the covariates, and $\lambda_0(t)$ is the baseline hazard, a non-negative function of time. Note that the covariates x are time varying throughout the remainder of the paper.

With the Cox model being well studied (see Therneau and Grambsch (2000)), there are many advantages of this model. For one such as the use of time-varying covariates allows for the easy integration of time-updated features. Moreover, the parametric component allows for flexible specification of the relationship between the covariates and the outcome. This can include non-linear effects, interactions, and dynamic (i.e. change over time) effects. The later, which we explicitly study allows one to approximate the more complex joint models. While the inclusion of such effects can result in an over parameterized model, Cox models are well suited for regularized methods (Simon et al., 2011). In addition, the Cox model is very computationally efficient for generating risk predictions, an important consideration from the clinical perspective given the goal is to implement this model in a real-time environment.

3.2.1. Multi-state Models

As motivated by our use case described in Section 2, we were only interested in patients when they were on a general medical floor. To capture this information, we extracted data on *where* in the hospital a patient was during her/his admission. When a patient left a unit of interest, (e.g. taken into surgery) we considered that s/he left the risk set. To account for this, we applied a multi-state model approach, which allows for a patient's location to consist of different

states, with transitions between the states (Putter, Fiocco and Geskus, 2007). We note, that multi-state models typically employ a Markov assumption.[1]

Putter et al. (2006) used this approach for developing a risk score for breast cancer patients. Analytically, this consists of estimating transition probabilities, as cause specific hazards, from one state to another. Our outcomes were similarly defined as 'states' with death being an absorbing state. This allowed us to define transitions of interest in assessing outcomes. For example, if a patient transitions from a surgery to the ICU, we would not consider this an event, while if a patient transitions from the floor to the ICU, this would be an event. We do this because the risk score is meant to operate on the general medical floor. A separate score could be considered for risk of death during a surgery. Another advantage of this approach is it allows us to consider discharge as a competing event as opposed to an independent censoring event. In the multi-state framework competing events are naturally handled via the cause specific hazard interpretation. We note that this makes any inference into risk factors challenging, though this is not our primary objective.

While Putter et al. (2006) established a framework for performing risk prediction from and to multiple states, in this paper, we are interested in transition from one specific state (medical ward) to one specific state (the composite of death, ICU transfer and RRT). Our task is simplified by considering only one particular transition. In principle, we did not have to create a composite endpoint and could have considered transition probabilities to each of these outcomes separately, however, this would have decreased our power to estimate effects and more importantly, clinically, is less important.

3.3. Analytic Approach

We now describe our analytic approach, namely providing model comparisons and model estimation in Sections 3.3.1 and 3.3.2.

3.3.1. Models Compared

Our interest was assessing the added value of both recalibrating the NEWS to our patient population and incorporating additional predictor variables. We considered four sets of predictor variables, namely:

1. the seven NEWS variables recalibrated to our patient population;
2. adding in additional demographics and comorbidities;
3. adding in laboratory tests;
4. adding in changes in vitals.

[1] That is, conditional on one's current state, the corresponding transition probability does not depend on one's previous state. We relax this assumption in the case of having a surgical procedure during the encounter. Specifically, we add a time varying indicator for whether a patient has had a surgery during the hospitalization.

These predictor sets represent increasing complexity of data to pull from the EHR system and then incorporate into a risk model. For (iv), we calculated the slope of the change for each continuous vital over the previous 4 hours. In addition, the standard NEWS was calculated as a benchmark.

3.3.2. Model Estimation

In order to perform model estimation, we randomly sampled 80% patients and included them in the training set, with the other 20% in the test set. We fit each of the four proposed models models and NEWS on the training data.

Using the training data we considered different ways of setting up the data. First, we assessed time-binning the data (see Section 3.1). Second, the NEWS table suggests there are U-shape relationships between four of the vitals (heart rate, blood pressure, temperature and respiration) and risk, with both low and high values conferring added risk. To account for this potential U-shape relationships, we include quadratic terms for those four vitals. Third, for model (4) we calculated changes over the previous 4, 8, 12 and 24 hours. Finally, we considered L_1 regularization for the largest model (Simon et al., 2011). For each of these we used 10-fold cross-validation on the training data assessing fit via the predictive partial log-likelihood (Verweij and Van Houwelingen, 1993). Results suggested that the optimal fit had no binning, incorporated quadratic terms, and used a 4 hour change value for vitals. Additionally, there was minimal added value to regularization, likely due to the relatively small ratio of predictors to observations.

3.4. Evaluation

To evaluate the models, we assessed each model's discrimination, or rather, the ability of a model to separate those who had an event and those who did not have an event via concordance (c)-statistics. There have been a number of proposals for how to calculate c-statistics for survival models with time varying predictors (Kamarudin, Cox and Kolamunnage-Dona, 2017). We use the approach of Heagerty and Zheng (2005), who propose the idea of an incident/dynamic sensitivity and specificity, which are defined as follows:

$$\text{sensitivity}^I(c,t): \ P(M_i > c | T_i = t) \tag{2}$$

$$\text{specificity}^D(c,t): \ P(M_i \leq c | T_i > t). \tag{3}$$

The "incident" sensitivity, equation 2, measures the expected fraction of subjects with a risk score, M_i, greater than c among the subpopulation of individuals who die at time t. On the other hand, the "dynamic" specificity, equation 3, measures the fraction of subjects with a risk score less than or equal to c among those who survive beyond time t. Since the sensitivity is calculated at the time of the event this method is well suited to time-varying risk scores. The global c-statistic is

$$P(M_j > M_k | T_j < T_k)$$

and c-statistics can be thought of as a weighted average under time specific ROC curves (Heagerty and Zheng, 2005).

To evaluate our risk models, we first calculate the c-statistic over time for each of the proposed variable sets as well as the NEWS. After identifying the best global risk model, we evaluated the model over different time horizons of 6, 12, 24, & 48 hours in order to assess what is the optimal time frame for which the model performs.[2]

A further advantage of our Cox model is the ability to generate individual risk predictions for patient risk relative to all other patients in the model. In order to do this, we compute the hazard ratio relative to the sample average for all of our predictor variables. Recall that the hazard for individual i is given by:

$$\lambda_i(t) = \lambda_0(t)e^{x_i(t)\beta}.$$

Thus, the hazard ratio between two individuals i and j with predictor variables x_i and x_j is independent of both the baseline hazard and time t. Therefore, the relative risk between two patients is given by:

$$\frac{e^{x_i(t)\beta}}{e^{x_j(t)\beta}}. \tag{4}$$

If we set $x_j(t)$ to the average values in the sample, then equation 4 represents the risk for person i relative to the average person, with a value above 1 indicating increased risk and a value below 1 decreased risk. To assess our model on the individual level, we chose 3 patients that ultimately had an event and plotted their relative risk score over time. We compared these plots to the relative risk score calculated by the NEWS. We also assessed globally how risk assessment varies between those that do and do not experience events.

4. Analysis

We provide an analysis of the proposed methodology. First, we give a more in depth description of the EHR data in Section 4.1. Next, we fit the four proposed models on the training data and calculate incident AUC predictions on the test data, making comparisons to the benchmark, the NEWS in Section 4.2. A full discussion of model performance is given in this section. In addition, we provide individual risk assessment in Section 4.3.

4.1. Data Description

Overall 5.4% of people had an event. The event rate was greatest during the first couple of days and by 30 days 98% of patients have either had an event or been discharged.

Supplemental tables 1–3 provide summaries of the extracted variables. Most patients had their first vitals measurement within four hours after admission

[2]We used the R package riskSetROC (Heagerty and Saha-Chaudhuri, 2012).

(Supp Table 1). Vitals were updated frequently but at irregular intervals ranging from several minutes to several hours, with a median interval of 1.9 hours. Note that there are meaningful differences in vital values and rates of different comorbidities (Supp Table 2) between those patients that did and did not experience events will and will not have events. Finally, with regards to laboratory test (Supp Table 3), we observed high variability between the frequency of which different labs are ordered, with some labs being ordered multiple times.

4.2. Overall Prediction Models

We fit each of the four proposed models models on the training data and calculated incident AUC of predictions on test data assessed over 30 days. We also calculated the NEWS on the test data. Figure 4 shows that the EHR based models significantly outperform the NEWS. Refitting the NEWS variables with our data improved the overall c-statistic from around 0.64 to 0.83. The inclusion of additional EHR covariates, slightly improves the overall fit with the labs based score having the best nominal c-statistic (0.84). However, there are not strong qualitative differences between the four models considered.

Model performance also varies over time. There is relatively strong performance at the time of admission though this quickly declines. As hospital stay increases, model performance improves. It is important to note that most events happen earlier in time when model performance is relatively weaker. One belief for the improved performance over time is that, as patients are stabilized, any deviation from "normal" values is more indicative of poorer health prognosis.

We also evaluate how well our model can predict future events based on current data. We look at 6-hour, 12-hour, 24-hour and 48-hour prediction performance for the model with vitals, demographics and comorbities and labs. Figure 5 shows that the model performs better in the near term.

Finally, we examine which covariates are most predictive of clinical risk. Figure 6 shows the coefficients for covariates. All covariates are standardized to have standard deviation of 1 so that the coefficients of our models are comparable. Not surprisingly, the strongest predictors are all the vital signs. This suggests that the NEWS likely identified many of the correct predictors and they are simply not optimized for our patient population. The other variables sets contribute less to the overall risk score, explaining why there is minimal improvement between the four models.

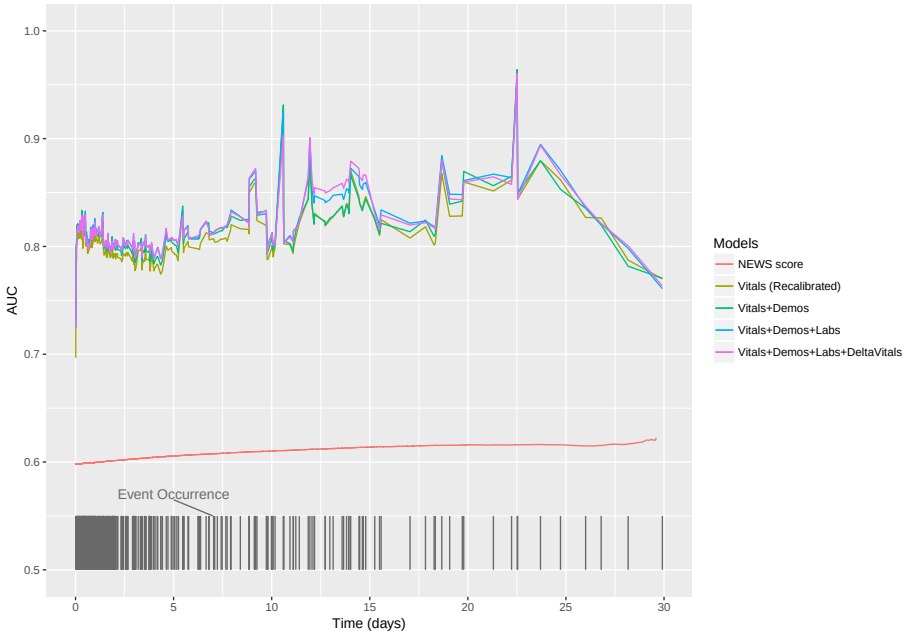


Fig 4: AUC for the 5 models assessed over time. The EHR based score performs meaningfully better than the NEWS. There is minimal difference between the 4 EHR variable sets.

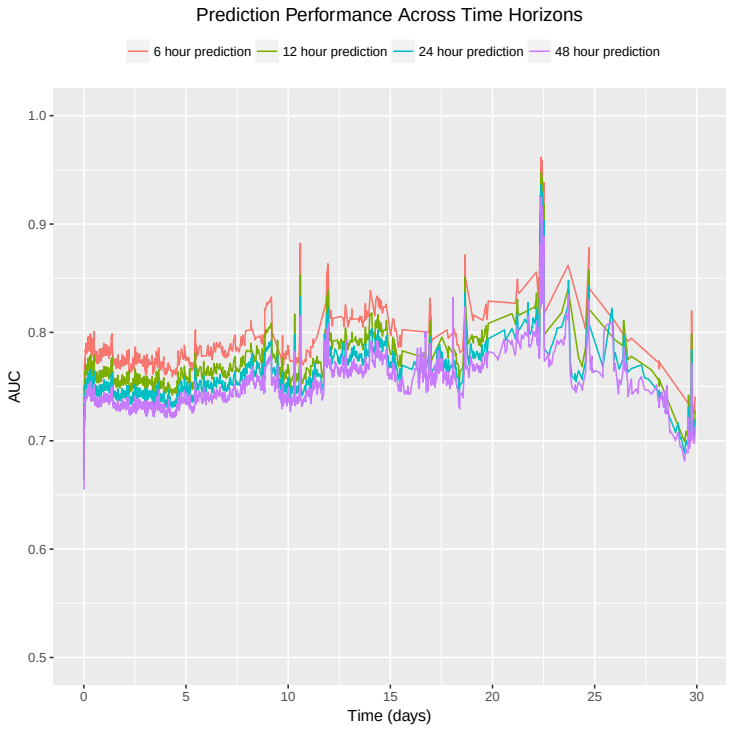


Fig 5: AUC for prediction at different time horizons under the model "Vitals+Demos+Labs." The risk model performs best in the near term (6 hrs).

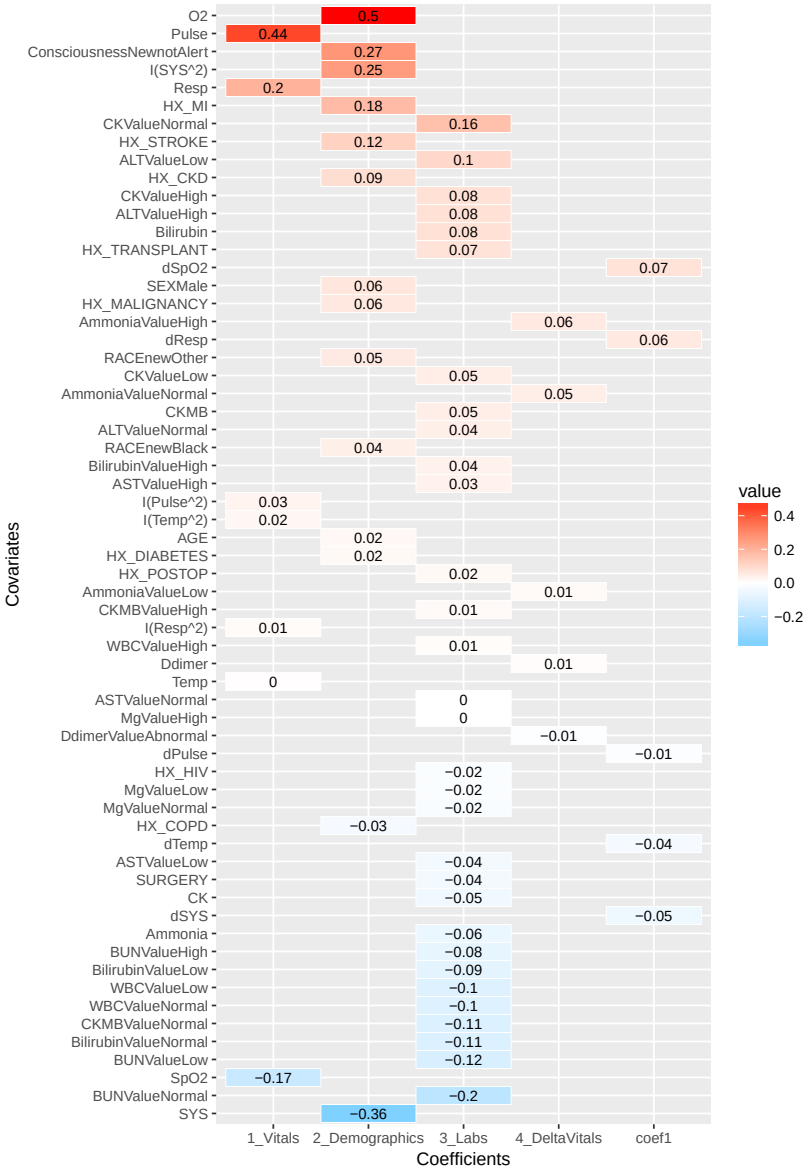


Fig 6: Cox regression coefficients estimated on the full model. All variables have been standardized to be comparable. Each column represents one of the four variable sets. Colors are on an intensity scale ranging from blue (protective factors) to white (no association) to red (risk factors). The NEWS variables have the strongest (most extreme) associations with the other variables being less impactful.

4.3. Individual Risk Assessment

In order to examine how the risk for individual patients varies between NEWS and our best performing score, we calculate the individual risk for each patient over their hospital stay as described in Section 3.4. In Figure 7, we illustrate the time varying relative risk for three individuals that ultimately had events. This allows us to examine how patient risk varies between models, and examine the underlying changes in patient features over time that drive the corresponding change in risk. In general, we find that for patients with events, the Cox model gives much higher relative risk scores, leading in turn to a higher predictive accuracy at all points during a patients hospital stay.

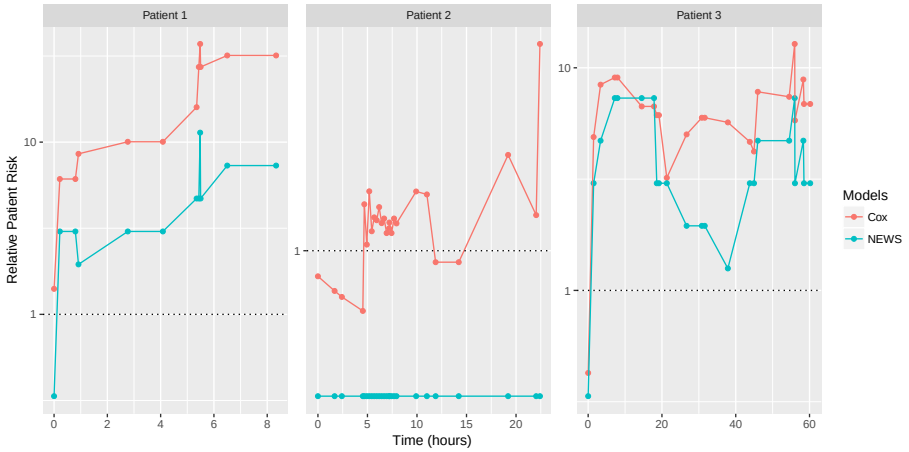


Fig 7: Relative patient risk for three individuals with events over their hospital stay for NEWS and the Cox model with Vitals+Demos+Labs. At each time point that patient's features are updated, a new individual risk score is calculated. These scores are plotted for each of the three patients allowing us to see an individual patient's risk changes throughout their entire stay.

Upon examining three individual patients we note different reasons for the superior performance of the Cox based risk score over the NEWS.

Patient 1: Both risk scores predict similar risk patterns. However, the Cox based scores has an amplified relative risk owing to the larger magnitude on the coefficients. We note that at hour six, there is a large decrease in oxygen saturations from 95% to 89% while all other factors remain the same.

Patient 2: During the observation period, this patient experiences fluctuations in heart rate, respiration rate, and systolic BP. However, these fluctuations remain within the normal intervals of the NEWS, giving a score of zero throughout the patient's entire stay. On the other hand, the

Cox model observes these changes as increases in risk relative compared with other patients in the observed data set.

Patient 3: This patient also has slightly elevated levels of risk in the Cox model. In particular, at the 20 hour mark, we note that the risk score for NEWS begins to decrease towards one. There are two key factors driving the difference between the Cox and NEWS models here. While this represents a period of relative stability, there is still a spike in systolic BP and smaller changes in heart rate. In addition, within this period a WBC lab is returned elevating the patients count from normal to high which is not taken into account by NEWS.

To further understand the differences in the NEWS and Cox based scores we stratified all people based on whether they had an event or were discharged alive. Then we calculated their risk score in the 24 hours prior to event/discharge. Figure 8 shows the time varying risk based on the Cox model and the NEWS. Overall, the Cox model estimates a meaningfully higher relative risk for those that have an event. Moreover, leading up to the event, the risk estimate begins to rise. This happens primarily in the 6 hours before the event, illustrating why the risk score has it's best performance at this horizon (see Figure 5). Conversely, the NEWS shows no such change prior to the event and there is no meaningful risk assessment difference between those with and without events, with people having events having a nominally *lower* risk score.

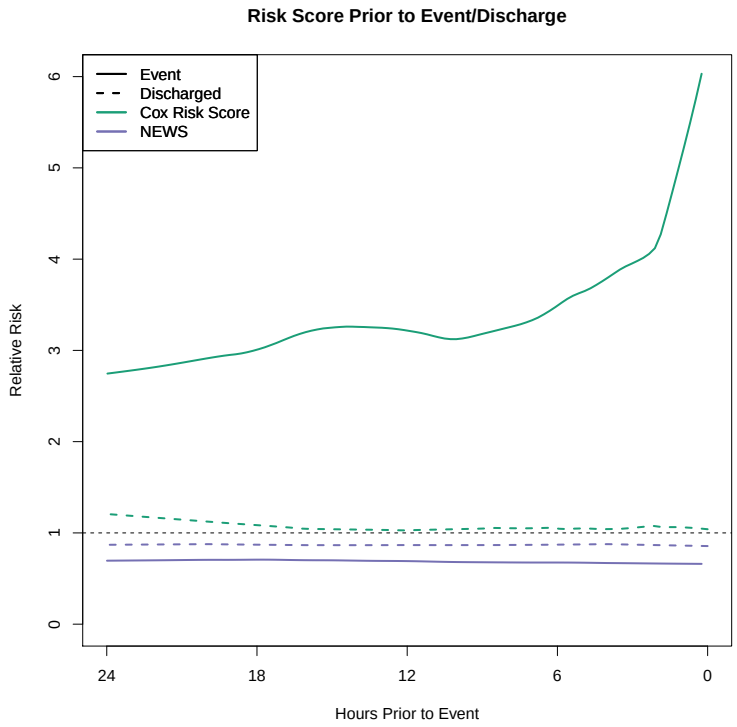


Fig 8: Relative risk in the 24 hours prior to deterioration event (solid lines) or discharge (dashed lines), where the dashed black line indicates reference risk of '1'. The Cox based score (solid and dashed teal lines) shows meaningfully different risk assessments between those experiencing and not experiencing events. In addition, the risk increases as one nears event. Conversely, the NEWS (solid and dashed purple lines) does not show any discrimination nor change with response that corresponds to the risk.

5. Discussion

EHR data are becoming more ubiquitous in clinical research, particularly risk assessment. However developing risk models with EHR data can be challenging since EHR data contain a number of challenges. These challenges include large sample sizes, dense and irregular longitudinal measurements, and complex data types with potentially missing values. In this paper we lay out many of these challenges and illustrate how one can apply a time-varying Cox model to flexibly derive a risk score.

By embedding our risk score within a Cox model we obtain increased flexibility over more complicated analytic approaches. We easily handle multiple longitudinal predictors, something that while theoretically possible with joint models and Gaussian Processes, is not computationally efficient (Banerjee, Dunson and Tokdar, 2011; Riedel and McCallum, 2011). Moreover, since a Cox model is well suited for regularization (Simon et al., 2011) it can also handle many predictors. This becomes useful because one can approximate the additional flexibility of joint models and Gaussian Processes by incorporating additional "slopes" or "changes" for the longitudinal predictors. While our results suggests that these variables did not improve risk assessment that flexibility is appealing.

The ability to flexibly code predictors is especially important with EHR data and becomes a means to handle missingness. Modeling predictors such as laboratory values, which are not collected on everyone is challenging. Since the lack of collection is likely informative — doctors only order a lab test if they think there is something wrong — imputation is not desirable. Therefore coding a lab test as "not ordered", "normal", "high" and "low" allows one to flexibly capture the clinical reality. It also allows one to capture the time delay between when a test is ordered and when the results arrive, an important consideration when one wants to implement a risk score in a real-time environment.

A final flexibility that a Cox model engenders, is the ability to allow people to leave and re-enter the risk set. In our use case we were only interested in patients that were at risk of deteriorating while on a general medical ward. Therefore, we have accounted for the time that a patient was in another location, e.g. a surgery, and therefore not part of the risk set. By taking a multistate modeling approach and defining the transitions of interest, this complexity was easily handled. In addition, our framework can account for any transition of interest.

Our work touches on some of the options for evaluating these types of risk models, both on aggregate and individual levels. On an aggregate level, one can consider risk performance both over-time and over different time horizons. Our model shows improved performance as a hospital stay progresses. This is likely because patients are stabilized after a couple of days so deterioration becomes more anomalous and therefore more predictable. Unfortunately, from a clinical perspective, most events happen earlier in the stay. Our evaluation also shows that prediction is best in the near term. While not surprising, this is also important clinically, as our clinical collaborators suggested they would ideally want 12 hours notice to effectively intervene on a patient. We have also illustrated how this model can be evaluated on an individual level. Our individual level analysis

showed that patients derive risk for different reasons. Therefore, it is important to have a model that can properly capture these changes. This is the primary limitation of a simpler model like the NEWS. While the general risk categories simplifies calculation, subtle changes are missed.

Finally, our results highlight the potential for using one's own data to develop a risk model as opposed to relying on off-the-shelf scores such as the NEWS. Even using the same variable set, we are able to derive meaningfully improved risk performance with a c-statistic of 0.83 vs 0.64. Interestingly, while we do find increased performance by adding in additional predictors such as demographics, comorbidities, and laboratory values, clearly the strongest predictors are those incorporated in the original NEWS. Therefore, even though others have found that additional risk factors are useful for assessing patient deterioration (Churpek, Adhikari and Edelson, 2016; Jo et al., 2016), this suggests that while the original authors likely identified the correct risk factors, the coefficients needed to be recalibrated for our environment. It is for this reason, that while we intend to implement this risk model in our clinical environment, we do not believe others should should do the same. Instead, the most transferable component of this work is not the model coefficients but the approach taken. Ideally, others would take a similar analytic approach and discover the optimal model for their clinical environment.

Acknowledgments

This work was funded by the Duke Center for Integrated Health Data Science. BAG was supported by National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) career development award K25 DK097279.

References

Alam, N., Hobbelink, E. L., van Tienhoven, A. J., van de Ven, P. M., Jansma, E. P. and Nanayakkara, P. W. (2014). The impact of the use of the Early Warning Score (EWS) on patient outcomes: a systematic review. *Resuscitation* **85** 587–594.

Banerjee, A., Dunson, D. and Tokdar, S. (2011). Efficient Gaussian Process Regression for Large Data Sets. *ArXiv e-prints*.

Churpek, M. M., Adhikari, R. and Edelson, D. P. (2016). The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation* **102** 1–5.

Colopy, G. W., Pimentel, M. A. F., Roberts, S. J. and Clifton, D. A. (2016). Bayesian Gaussian processes for identifying the deteriorating patient. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 5311-5314.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* **34** 187–220.

Dolin, R. H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F. M., Biron, P. V. and Shabo, A. (2006). HL7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association* **13** 30–39.

Goldstein, B. A., Navar, A. M., Pencina, M. J. and Ioannidis, J. P. (2017a). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* **24** 198–208.

Goldstein, B. A., Pomann, G. M., Winkelmayer, W. C. and Pencina, M. J. (2017b). A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis. *Stat Med*.

Heagerty, P. J. and Saha-Chaudhuri, P. (2012). risksetROC: Riskset ROC curve estimation from censored survival data R package version 1.0.4.

Heagerty, P. J. and Zheng, Y. (2005). Survival Model Predictive Accuracy and ROC curves. *Biometrics* **61** 92-105.

Henry, K. E., Hager, D. N., Pronovost, P. J. and Saria, S. (2015). A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine* **7** 299ra122–299ra122.

Jo, S., Yoon, J., Lee, J. B., Jin, Y., Jeong, T. and Park, B. (2016). Predictive value of the National Early Warning Score-Lactate for mortality and the need for critical care among general emergency department patients. *J Crit Care* **36** 60–68.

Kamarudin, A. N., Cox, T. and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med Res Methodol* **17** 53.

Kipnis, P., Turk, B. J., Wulf, D. A., LaGuardia, J. C., Liu, V., Churpek, M. M., Romero-Brufau, S. and Escobar, G. J. (2016). Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *J Biomed Inform* **64** 10–19.

Paxton, C., Niculescu-Mizil, A. and Saria, S. (2013). Developing predictive models using electronic medical records: challenges and pitfalls. In *AMIA Annual Symposium Proceedings* **2013** 1109. American Medical Informatics Association.

Putter, H., Fiocco, M. and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* **26** 2389–2430.

Putter, H., van der Hage, J., de Bock, G. H., Elgalta, R. and van de Velde, C. J. (2006). Estimation and prediction in a multi-state model for breast cancer. *Biometrical Journal* **48** 366-380.

Riedel, S. and McCallum, A. (2011). Fast and Robust Joint Models for Biomedical Event Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11* 1–12. Association for Computational Linguistics, Stroudsburg, PA, USA.

Rizopoulos, D. (2011). Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics* **67** 819-829.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* **39** 1-13.

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E. and Featherstone, P. I. (2013). 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** 465-470.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag New York.

Verweij, P. J. and Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Stat Med* **12** 2305-2314.

Wehbe, L., Ramdas, A., Steorts, R. C., Shalizi, C. R. et al. (2015). Regularized brain reading with shrinkage and smoothing. *The Annals of Applied Statistics* **9** 1997–2022.

Appendix A

Below we present additional description of the available data

Vitals	**Average number of observations per patient**	**Median (IQR)/Frequency Patients with event**	**Median (IQR)/Frequency Patients without event**
Heart Rate	35	90 (77, 106)	82 (70, 94)
Respiration Rate	32	20 (18, 22)	18 (16, 20)
Oxygen Saturations	27	96 (94, 98)	97 (95, 98)
Systolic BP	33	122 (108, 140)	124 (111, 140)
Temperature	25	36.8 (36.6, 37.0)	36.8 (36.6, 37.0)
O2 (receiving supplemental oxygen)	45	70.0%	51.6%
Consciousness (not alert)	13	39.1%	24.0%

Table 1: Vitals signs that are part of the primary NEWS. Patients had many readings during the exposure period reflecting the need for a time updating score. There are also meaningful differences between those that did and did not experience events for many of the vitals.

Demographics	**Percentage among patients with event**	**Percentage among patients without event**
Diabetes	36.4%	28.8%
Malignancy	35.2%	30.6%
CKD	28.0%	18.7%
COPD	18.3%	10.7%
MI	16.1%	9.0%
Stroke	9.5%	5.2%
HIV	1.3%	1.2%
Post-op	8.4%	7.2%
Transplant	10.2%	6.6%
Surgery	16.4%	25.7%
Sex = Female	45.0%	48.0%
Race = Black	31.0%	28.1%
Age (Average)	62	60

Table 2: Demographics variables. Those that ultimately experience events have much higher comorbidity rates.

Lab	Percent of patients ordered for	Median orders per patient	Average wait between order time and result time (hours)
WBC	93%	4	5.1
BUN	87%	4	5.2
Bilirubin	65%	2	4.7
Mg	62%	3	5.3
ALT	54%	1	4.5
AST	54%	1	4.5
CKMB	23%	2	3.1
CK	22%	1	3.0
Ammonia	5%	1	2.9
D-dimer	1%	1	8.8

Table 3: Lab Orders, Order Frequency and Wait Time. Some labs are ordered for all patients (i.e. White Blood Cell Count) while others are infrequently ordered (i.e. D-dime). Also there is a meaningful lag between when a lab is ordered and when the results are available.

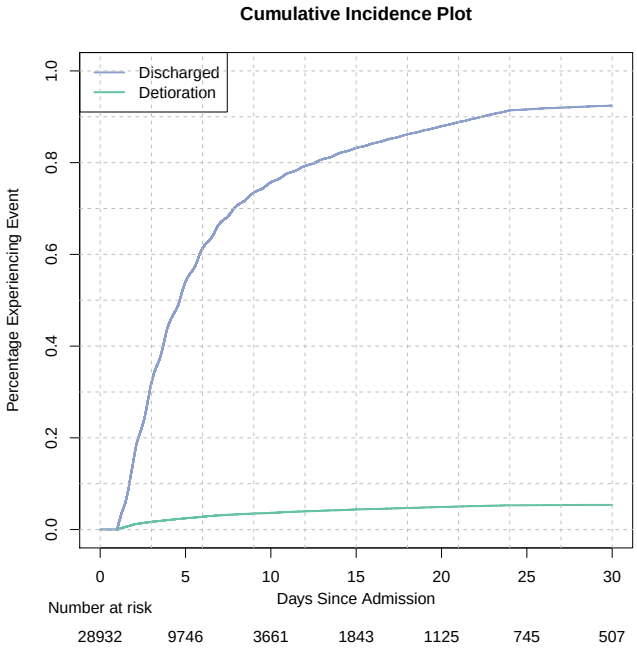


Fig 9: Cumulative incidence over first 30 days of hospitalization. Event rate is greatest over the first couple days of a hospitalizations. By 30 days 98% of patients have either had an event or been discharged.

Appendix B

In Section 4.3, we present the results for three patients with events. To further illustrate the the application of individual patient risk calculations and provide comparisons between the NEWS and Cox models, we calculate relative risk for a further subset of patients in our data. In Figures 10, 11, 12, and 13 we show the time varying relative risk for 200 randomly sampled patients both with and without events.

Fig 10: Relative patient risk for 50 randomly sampled individuals with events over their hospital stay for NEWS and the Cox model with Vitals+Demos+Labs. At each time point that patient's features are updated, a new individual risk score is calculated. These scores are plotted for each of the 50 patients allowing us to see an individual patient's risk changes throughout their entire stay.

Fig 11: Relative patient risk for 50 randomly sampled individuals with events over their hospital stay for NEWS and the Cox model with Vitals+Demos+Labs. At each time point that patient's features are updated, a new individual risk score is calculated. These scores are plotted for each of the 50 patients allowing us to see an individual patient's risk changes throughout their entire stay.

Fig 12: Relative patient risk for 50 randomly sampled individuals with events over their hospital stay for NEWS and the Cox model with Vitals+Demos+Labs. At each time point that patient's features are updated, a new individual risk score is calculated. These scores are plotted for each of the 50 patients allowing us to see an individual patient's risk changes throughout their entire stay.

Fig 13: Relative patient risk for 50 randomly sampled individuals with events over their hospital stay for NEWS and the Cox model with Vitals+Demos+Labs. At each time point that patient's features are updated, a new individual risk score is calculated. These scores are plotted for each of the 50 patients allowing us to see an individual patient's risk changes throughout their entire stay.