Explaining an increase in predicted risk for clinical alerts

Michaela Hardt*
Amazon
milaha@amazon.com

Alvin Rajkomar
Google, UCSF

Gerardo Flores
Google

Andrew Dai
Google

Michael Howell
Google

Greg Corrado
Google

Claire Cui
Google

Moritz Hardt*
UC Berkeley

ABSTRACT

Much work aims to explain a model's prediction on a static input. We consider explanations in a temporal setting where a stateful dynamical model produces a sequence of risk estimates given an input at each time step. When the estimated risk increases, the goal of the explanation is to attribute the increase to a few relevant inputs from the past.

While our formal setup and techniques are general, we carry out an in-depth case study in a clinical setting. The goal here is to alert a clinician when a patient's risk of deterioration rises. The clinician then has to decide whether to intervene and adjust the treatment. Given a potentially long sequence of new events since she last saw the patient, a concise explanation helps her to quickly triage the alert.

We develop methods to lift static attribution techniques to the dynamical setting, where we identify and address challenges specific to dynamics. We then experimentally assess the utility of different explanations of clinical alerts through expert evaluation.

1 INTRODUCTION

Routinely framed as a static prediction task, statistical risk assessment is often a dynamic problem. Risk estimates evolve over time as new measurements arrive and additional data become available. Alerts may be triggered once the risk or the increase in risk has reached a critical threshold. An alert can depend on numerous events in the past, making it difficult to quickly understand which events contributed to the increase in risk. The goal of our work is to provide tools to quickly assess an increase in predicted risk in dynamical risk assessment scenarios.

*work done at Google

Our main application is a clinical early warning system that alerts physicians to the deteriorating health of a hospitalized patient [1, 2]. Such a system can decrease mortality and length-of-stay [3]. It combines events such as measured vital signs, laboratory tests, and notes from doctors into a patient risk score. Once the risk score exceeds a threshold, an alert notifies the doctor. Alerts should have high precision – too many false alerts can lead to fatigue [4]. Alerts should also be informative; a clinician must be able to quickly assess what factors are contributing to the patient's increase in risk to identify interventions. This is not an easy task as there may have been dozens or hundreds of new events since the clinician last saw the patient. Simply reporting abnormal lab results or vitals is unhelpful in situations where abnormalities are common. For example, in the case of critically ill patients, abnormal lab results may be unrelated to the alert and are often nonspecific [5].

While much work on model interpretability has focused on identifying important features of a static input, these works do not address salient temporal effects, such as how the relevance of input changes over time as newer data become available. A recent survey of clinicians [6], describes temporal explanations as desirable by clinicians, yet unexplored in the ML literature. Our work begins to address these important questions with a detailed use case in the medical domain.

Our contributions. We broadly study explanations of an increase in risk in a dynamic setting. This is a departure from the predominantly studied static setting and we hope that it will spark more interest in the future. Within this broader context, we dive deep into an application and use our methods to explain alerts in a clinical early warning system, where we see our results as providing a set of valuable baselines for future work. More specifically, we make the following contributions towards explaining *changes* in prediction over time.

- Develop methods to lift commonly used static gradient-based attribution techniques [7–10] to the dynamical risk assessment setting.
- Analyze our methods in the simplified theoretical model of a linear dynamical system with a quadratic risk function to form an analytic understanding of the semantics and challenges.

- Implement our methods in the context of an early warning system to explain alerts, and open-source our code[1], see supplementary material.
- Evaluate our methods through expert judgment by medical students and an ICU doctor who were asked to assess the clinical utility and compare them to attention [11] and statistical methods [12, 13].

Organization. We describe the problem setup and its application to a clinical early warning system in Sec. 2. Related work is reviewed in Sec. 3. We then present our methods in Sec. 4 and analyze them theoretically in Sec. 5 in a simplified linear dynamical system gaining insights into the challenges that motivate some refinements. Sec. 6 details our experiments. We conclude in Sec. 7.

2 PROBLEM STATEMENT

We now introduce the general problem we address. At each time step t with $t = 0, 1, 2, \ldots, T$, a real-valued vector, $x_t \in \mathbb{R}^d$, is provided as input to a stateful model to produce an updated risk estimate $p_t \in \mathbb{R}$. We denote by h_t the hidden state of the model at time t. We then phrase the problem of explaining an increase in predicted risk as the identification of a few values in the sequence of inputs that are relevant to a risk increase. Clearly, the whole sequence of inputs perfectly explain the prediction, however, there may be too many inputs for a human to make sense of. Also some inputs may rather be indicative of a decrease in risk and would not make good explanations. The notion of relevancy is domain specific and depends on the consumers [14]. We therefore assess it through human evaluation in our experiments.

Notation. In this paper we consider attribution weights, $a \in \mathbb{R}^{d \times T}$, over all inputs and the explanations are then simply the input values with the highest weights. The weights may vary depending on which risk increase from p_{t_0} to p_{t_1} we are trying to explain. This can be made explicit with a superscript $a^{t_0 \to t_1}$. For instance, we may expect weights from the future $t > t_1$ to be zero. Extending explanations to chains of events is an open problem.

2.1 Use case of an early warning system

One promise of digital health is that electronic health record (EHR) data of hospitalized patients could be continuously monitored by electronic systems that alert physicians when a patient's health is worsening [1, 2]. Early warning systems for deterioration have already been shown to decrease mortality and length-of-stay [3]. Since many patients are hospitalized because they are at high risk, repeated alerts for persistently elevated risk may induce alarm fatigue, a known clinical problem. Instead, alerts are likely more relevant when a patient's risk has increased, especially over a short period of time, where there may exist possible interventions to blunt or reverse the increase in risk [13]. For example, we may seek to trigger alerts for patients whose mortality risk increased by at least 50% and is now above 0.2. In practice, a hospital can tune these alerts based on precision (or the number needed to evaluate) and recall (or the available resources).

Given an alert, a clinician has to decide quickly what to do about it in the presence of many competing inputs and responsibilities [15,
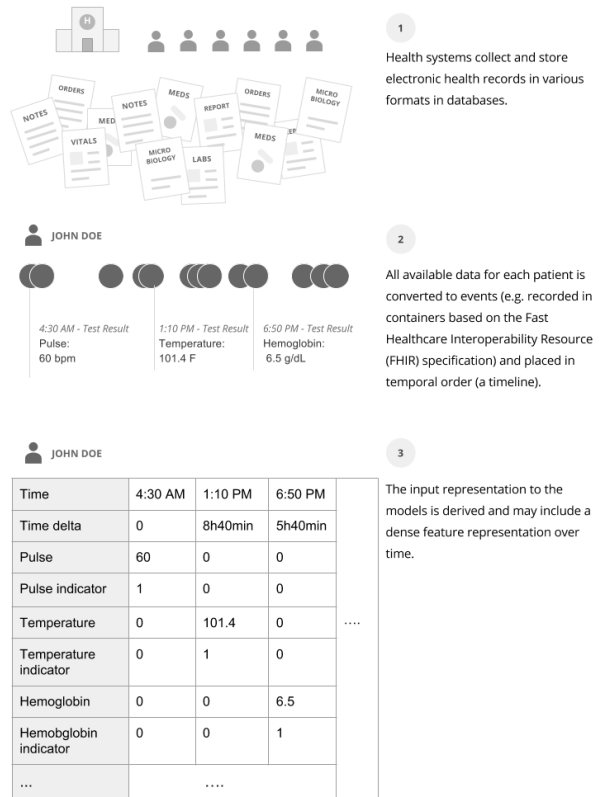
Figure 1: Data from a health system grouped by patient and ordered by time. Further input processing is model-dependent. In Step 3 we focus on lab results formatted for the input to a model prior to normalization.

Figure 2: Data from a patient in a health system ordered by time. We see the timeline of diagnoses, procedures, medications, etc. of a patient with the lab results at a particular time expanded.

---

[1] https://github.com/google-research/google-research/tree/master/explaining_risk_increase

	x_0	x_1	x_2	
Time delta	0	8h40min	5h40min	….
Pulse	60	0	59	
Pulse indicator	1	0	1	
Temperature	0	101.4	0	
Temperature indicator	0	1	0	
…				

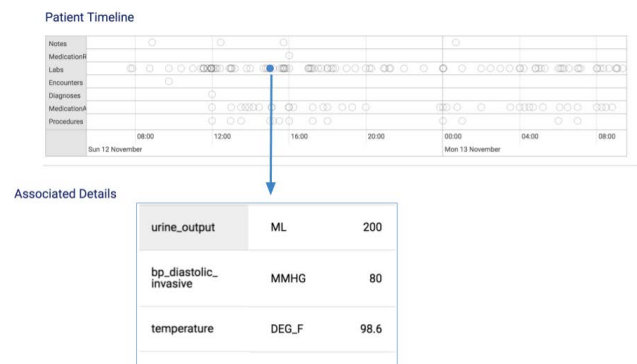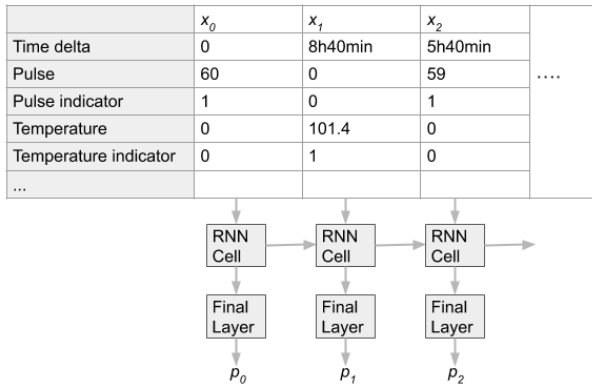**Figure 3: Input data to a risk prediction model. Specifically, a sparse representation of irregular lab results are used as input into a recurrent neural network (RNN). By maintaining an internal state, the prediction at a time depends on all previous lab results. Each time, a new lab result arrives, we create a sparse vector with the result at a position determined by the lab test and an indicator. This vector together with the internal state is fed into the RNN cell followed by a linear projection to obtain a risk score that is mapped to the range 0 to 1 using a sigmoid function. Additionally, the internal state is updated.**

16]. A concise explanation of an alert is helpful for triaging of alerts and clinical decision making. A good explanation should focus on the new information since the clinician last saw her patient.

Input Data. EHR data is heterogeneous and includes diagnoses, medications, procedures, notes from doctors, vitals and lab results is illustrated in Fig. 1. We can select all data from a single patient and order it by time to construct the patient's timeline see Fig. 3. While all of these features may be relevant for the risk assessment, they do not all make good explanations. Medications, for example, may be associated with a high risk without causing it. For instance, Norepinephrine is administered to patients in shock and therefore accompanies but actually decreases their high risk of dying. As a first step, we focus on lab results and vitals as explanations. For concreteness, we describe a commonly used input representation [17–19], see Fig. 3. We set $x_{t,i}$ to be the normalized result of a lab test i at time t (or 0 when the lab test was not taken). To distinguish a lab result of zero from an absent result we further include an indicator for each lab test. We also track the time between two consecutive events as a feature. The reason to focus on lab results is twofold. First, there is an abundance of lab results compared to any other type of medical event in the EHR, see Fig. 3, which makes it easy to miss important ones. They can be collected routinely but may only be reviewed by a clinician during the morning rounds or evening sign-outs. Second, they can give us insights into the physiological state of the patient, with a caveat discussed next.

Challenge of confounded measurements. The existence of a lab result is often due to a worried clinician ordering a test [20]. Patterns of lab tests can be as revealing as their results [18, 21]

and increasing the frequency of measurements leads to a risk increase [22]. Furthermore, results are affected by treatments as the following example illustrates.

Example. The tidal (lung) volume of 500ml of a patient was associated with a high risk despite the fact that 500ml is normal. This is partially because this test is only performed on intubated patients (generally at higher risk). But even after comparing only intubated patients the value of 500ml was still associated with an increased risk. The ventilation settings of a patient can be either specified by a target volume (commonly 500ml) or by pressure (in which case case the volume varies more widely). The first method is chosen more often for serious cases and once patients get better they may be switched to the second method. Hence, a value of 500ml is associated with more serious cases requiring intubations.

While our methods are general and do not depend on this specific input and domain, we will discuss their ability to deal with confounded irregular measurements.

3 RELATED WORK

The increasing use of complex models for decision making in critical areas including health care and the criminal justice system, has raised the question of model interpretability: We need to be able to check the soundness of the reasoning of our models [14]. For a general overview of interpretability see the comprehensive survey [23].

3.1 Explaining Predictions

Gradient-based methods have been applied mostly in image classification to produce a saliency map [7–10, 24–27]. Attention-mechanisms have been used with recurrent neural networks (RNNs) [11]. Techniques requiring only black-box access to a model use local additive models [28], ablate features through Shapley values [29], or maximize the mutual information of the predictions and the explanations [30].

3.2 Model Interpretability for Health Care

In clinical practice, mostly small and simple models based on 1-5 features are deployed [31–34]. While easy to inspect, their accuracy is limited. With rich data now digitized in EHRs [35], more features and model architectures are being developed with improved accuracy [36]. These complex models raise questions of interpretability: How can we understand these models [37–41]? How can we group patients [38, 42, 43]? How can we explain predictions? To explain predictions on EHR data (the focus of this paper), a variety of approaches have been explored starting with generalized additive models [12] over discretized features, or feature crosses fitted using gradient boosted decision trees. Subsequent work has extracted more complex discretized features that also incorporate temporal aspects using the maximum information gain criteria [44]. To explain a patient's risk, statistics of these discretized features can be used such as the odds ratio or the Rothman index [13]. Other lines of work have studied latent Dirichlet allocation [45], convolutional neural networks with feature ablation [46], RNNs with an attention mechanism [47, 48], and co-distillation [49, 50].

Our paper differs from previous work by explaining an increase in predicted risk rather than explaining a static risk prediction. Our methods build up on prior work and lift them to the dynamical setting.

4 PROPOSED DYNAMICAL GRADIENT-BASED METHODS

In this section we describe how we can lift several existing static attributions to the dynamical setting. Some adaptions are rather simplistic, while others provide a new perspective on the dynamics. Our methods rely on a model that is trained to predict the risk p_t given inputs $x_0, \ldots, x_t$ and take various derivatives of the predictions p_t.

4.1 Input Gradients

A natural method [7–9] to explain the risk p_{t_1} considers the derivative of the risk score with respect to an input x_t at a time step $t \leq t_1$:

$$a_t^{t_1} = \frac{\partial p_{t_1}}{\partial x_t}.$$

This captures how a small change in the input affects the prediction. To explain a change in predictions we can apply a time restriction to the time window of interest. The underlying assumption is that good explanations of a change in risk from p_{t_0} to p_{t_1} should contain recent events in $(t_0, t_1]$.

Definition 4.1. Given a sequence of attribution weights $a_0, a_1, \ldots,$ we define the *time-restricted explanations* of the change in prediction between t_0 and t_1, $a_t^{t_0 \to t_1}$, at time t as a_t for $t_0 < t \leq t_1$ and 0 otherwise.

There are several popular variants that improve on the basic input gradient method [10, 24–26]. In the following, we will devise variants with a greater focus on the dynamical aspect of the problem.

4.2 Temporal Integrated Gradients

We extend the *integrated gradient* method [10, 27] that averages out the gradients along the line segment between two chosen input sequences, the target sequence $\hat{x} = (\hat{x}_0, \ldots)$ and a baseline sequence $b = (b_0, \ldots)$, typically set to all zeros. Formally, the path-integrated gradient of a prediction, p_{t_1}, is given by the integral

$$S(b, \hat{x}) = (\hat{x} - b) \int_0^1 \left. \frac{\partial p_{t_1}}{\partial x} \right|_{x = \alpha b + (1-\alpha)\hat{x}} d\alpha \, .$$

We lift path-integrated gradients to the dynamical setting by purposefully constructing a suitable baseline. This allows us to explore how intermediate values between the old and the new ones would have affected the prediction as motivated by the following example.

Example. If a patient currently has a temperature of 102F changing it a bit may not change the risk very much. However, if we compare it with the value the patient had the last time the doctor came around, say 98F, and compute the gradient of interpolations, the predicted risk may be sensitive to values close to the fever threshold of 100.4F.

Specifically, we construct a baseline $b = (b_0, \ldots, b_{t_1})$ given the actual input sequence $\hat{x} = (\hat{x}_0, \ldots, \hat{x}_{t_1})$ as follows. For early time steps $t \leq t_0$, we set $b_t = \hat{x}_t$. Afterwards, for $t > t_0$, the baseline pretends that the results have not changed since t_0, that is we substitute recently measured values of a feature after t_0 with the latest measurement of that feature until time t_0. To address the challenge of irregular measurement patterns, we consider an input $\hat{x}$ in which not all features are measured each time. We construct b by keeping the exact same measurement patterns of $\hat{x}$, replacing all results after t_0 with the most recent measurement up until t_0.

In our example, for temperature measurements of 99F at noon, 100.1F at 1pm, 100.2F at 2pm, 100.9F at 3pm, the baseline to explain the change in risk between 1:30 and 3pm is 99, 100.1, 100.1, 100.1 copying forward the last value before 1:30pm. With other intermittent measurements the sequence could look like

$$99, \perp, \perp, 100.1, \perp, 100.2, 100.9$$

with a baseline of

$$99, \perp, \perp, 100.1, \perp, 100.1, 100.1.$$

4.3 Time Derivatives

Another natural approach considers the derivative of the risk score with respect to time, $\frac{\partial p_t}{\partial t}$. The naive discrete time approximation of this derivative is $p_t - p_{t-1}$. This allows us to assign a weight to each time-step. We setup our input so that there is exactly one feature i_t, $0 \leq i_t < d$, present in the input at step x_t. Then we can set $a_t = (p_t - p_{t-1}) \cdot e_{i_t}$ with absent features receiving a 0 weight. The motivation is similar to that of time restrictions: A good explanation of a change in predictions from p_{t_0} to p_{t_1} should contain events between t_0 and t_1. At the limit, this means that the change $p_t - p_{t-1}$ should be attributed to the event at time t.

Properties: The weights of events are consistent in time. That is even as future events come in, the weights of past events do not change. Moreover, the sum of the weights in $(t_0, t_1]$, $\sum_{t_0 < t \leq t_1} a_t$ is equal to $p_{t_1} - p_{t_0}$, as it is a telescoping sum.

Challenge: The discrete time derivatives perform poorly in the presence of noise and large sampling intervals, as is well known and has motivated much work in signal processing, see, e.g., [51].

To reduce noise, we suggest to alter the training objective of the model. Instead of focusing solely on minimizing the loss of the prediction p_t compared to the actual risk, we additional minimize the changes of the prediction over time by adding a smoothing loss of the form $\eta \sum_t (p_t - p_{t-1})^2$ for some scalar $\eta > 0$. We refer to this refined method as *smoothed discrete-time derivatives*. The loss term promotes a smoother sequence reducing noise. Fig. 4 illustrates the same patient's predicted risk time series with and without smoothing. With a smoothing loss term and $\eta = 0.005$ much of the noise is removed while the shape is retained. Notably, the absolute risk is lower with smoothing which may not affect the explanations since they rely on the differences in risk.

5 ANALYSIS IN A LINEAR DYNAMICAL SYSTEM

Although our methods apply to general differentiable time-series models, we ground our discussion here in a simple illustrative
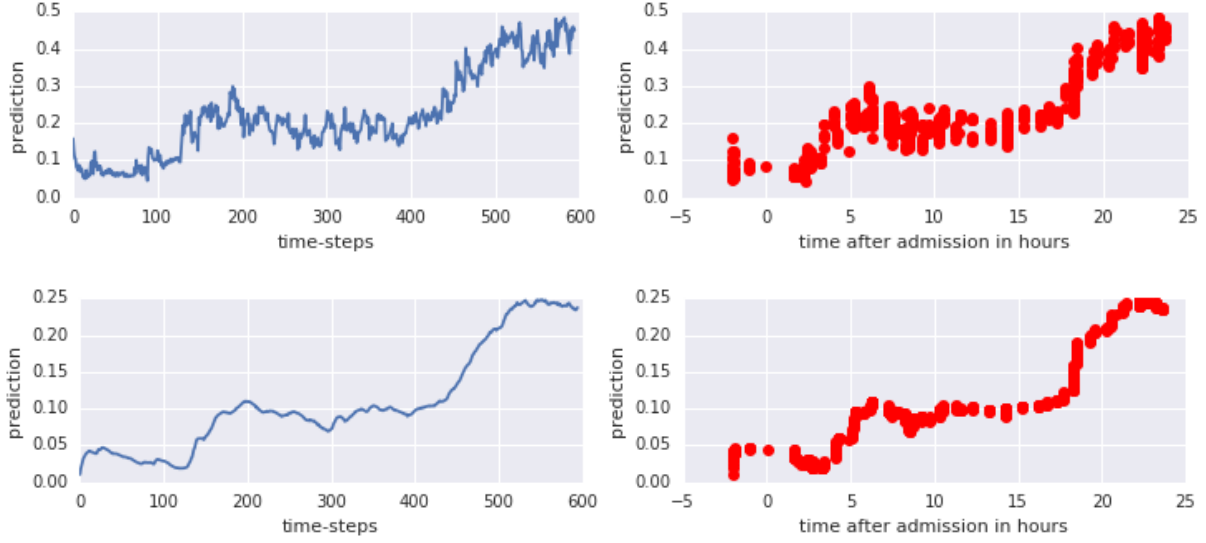
Figure 4: A patient's risk time series predicted from RNN models on the top trained without a smoothing loss term on the bottom and on the right trained with smoothing. The blue graphs in the first column use time steps as x-axis, while red graphs in the second column use the time relative to admission illustrating clustered activity.

example of a linear dynamical system. Recall at each time step t, an input $x_t \in \mathbb{R}^d$ influences the evolution of the hidden state h_t of a stateful model and yields a new risk score p_t. In our linear dynamical system the risk score is defined recursively as

$$p_t = 0.5\|h_t\|^2 \text{ and}$$
$$h_t = Ah_{t-1} + Bx_t.$$

In words, the risk score at time t is the squared Euclidean norm of a hidden state $h_t \in \mathbb{R}^n$ that evolves from a known hidden state h_0 according to a linear dynamical system specified by two linear transformations $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times d}$. For the purpose of our example, we could replace $\|h_t\|^2$ with any quadratic form $h_t^\top Q h_t$.

5.1 Time derivatives

In a continuous version of our linear dynamical system with

$$\frac{\partial h_t}{\partial t} = Ah_{t-1} + Bx_t$$

the continuous time derivatives are (by definition and the chain rule)

$$\frac{\partial p_t}{\partial t} = h_t^\top A h_{t-1} + h_t^\top B x_t$$

It is very common to approximate continuous time derivatives through discrete approximations such as $p_t - p_{t-1}$.

5.2 Input gradients

Applying the multivariate chain rule reveals that in our linear system

$$\frac{\partial p_{t_1}}{\partial x_t} = h_{t_1}^\top A^{t_1 - t} B$$

The index t of the input only influences the power of the matrix A. For fixed t, the gradient can grow or shrink exponentially

as t_1 increases. These phenomena known as *exploding/vanishing gradients* lead to the gradient method emphasizing inputs far in the past or very recently. The issues of vanishing and exploding gradients can be mitigated through careful design of the model, e.g. the use of LSTMs [52], and by the time-restriction to recent inputs.

5.3 Integrated gradients

We analyze the integrated gradient in our linear dynamical system. Denoting by $h_{t_1}[x]$ the hidden state that the system attains on an input sequence $x = (x_1, \ldots, x_{t_1})$, the above expression simplifies using the linearity of the integral operator and the system's update rule to

$$S_t(b, x) = \left((h_{t_1}[b] + h_{t_1}[x])/2\right)^\top A^{t_1 - t} B$$

What we see is that the path-integrated gradient simplifies to the arithmetic mean of the baseline gradient and the target gradient. Although not too different from the basic gradient method in this example, path-integrated gradients with carefully constructed baseline can alleviate some of the shortcomings of the basic gradients.

6 EVALUATION

6.1 Dataset

We use data from the Medical Information Mart for Intensive Care (MIMIC-III) [53] restricted to patients hospitalized for greater than 24 hours. MIMIC-III is widely accessible to researchers under a data use agreement. The data has been deidentified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting.

We included more than 50,000 hospitalizations, with patients randomly split into training (80%), validation (10%), and test (10%) sets. We consider the commonly studied prediction task of in-hospital mortality [17, 46]. We predict at various time points throughout

Michaela Hardt, Alvin Rajkomar, Gerardo Flores, Andrew Dai, Michael Howell, Greg Corrado, Claire Cui, and Moritz Hardt

the hospitalization how likely a patient is to die within this hospital stay. The in-hospital mortality rate was around 8%. In Section 6.6, we study a different form of deterioration, namely acute-kidney injury (AKI).

Input Data for Explanations. For the models generating explanations we create a whitelist of eligible lab tests, because not all lab results are easily interpretable. For instance, the red blood cell distribution width is often picked up by predictive models, but not well understood [54]. Also among highly correlated lab tests, clinicians prefer one over the other (e.g. hemoglobin over hematocrit). Specifically, we restricted the lab tests to the following list of roughly 40 target harmonized features: blood pressure, pulse, respiratory rate, oxygen saturation, blood pressure mean, temperature, urine output, glucose, oxygen source, hematocrit, hemoglobin, potassium, creatinine, bun, sodium, chloride, platelet, wbc count, mcv, magnesium, calcium, phos, inr, pt, ptt, carbon dioxide, tbili, ast, alt, alkphos, lactate, ionized calcium, albumin, troponin, egfr mdrd, tsh, dbili, hgba1c, total cholesterol, HDL, LDL. We used any codes that would map into one of these harmonized features.

We limit the lab values to the past three days.

6.2 Alert Generation

We chose a model for alerting that is separate from the various models used to generate explanations. In particular, we use a sequence model similar to [55]. That way we can decouple the quality of the alerts from the quality of the explanations. However, this means that our explanations may not be faithful to the alerts. Further research is needed to understand the interplay between the quality of model, alert logic, and explanations. We assume the clinician finished their rounds at 12 hours after admission at which point in time they have assessed the patient and developed a treatment plan. The risk predicted at this point is compared to the risk recomputed every 2 hours up until 24 hours after admission when the evening sign-out takes place and another team takes over. We trigger an alert for a patient if their risk increases by at least a factor of 1.5 to at least a risk of 0.2 to largely eliminate false-positives. The cost of human evaluation limits us to one specific setup. We randomly selected patients that are being alerted on for the first time that have at least 40 new events. For patients with fewer new events a doctor could feasibly look over all of them. We excluded subsequent alerts as the clinician may have snoozed the alert. This way of generating alerts hopefully resembles a realistic deployment scenario. The alerts are then simply used to pick the patients and times for which we generate and compare explanations.

6.3 Methods

We compare our gradient-based explanation methods from Sec. 4, the standard gradients, the temporal integrated gradients with a carefully designed baseline, and the time derivatives, to a few methods from the literature. We released our code at https://github.com/google-research/google-research/tree/master/explaining_risk_increase. An *attention mechanism* [11] for recurrent neural networks derives weights for each input event to create a prediction. These weights can serve directly as explanations. We apply time-restriction in our dynamical setting. We further include two statistics commonly used to associate risk with discretized features: The *odds-ratio*, e.g.

Model	Mortality Risk Precision @ 3
Random guessing	0.24 [0.16, 0.33]
Time-restricted gradients	0.31 [0.23, 0.39]
Time-restricted attention	0.33 [0.23, 0.43]
Smoothed discrete-time derivatives	0.40 [0.30, 0.50]
Time-diffed odds ratio	0.52 [0.39, 0.63]
Time-diffed Rothman index	0.52 [0.42, 0.63]
Time-restricted odds ratio	0.52 [0.42, 0.62]
Temporal integrated gradients	**0.57** [0.47, 0.68]

Table 1: For mortality risk, precision of the top-3 highest weighted lab results from distinct lab tests with a 95% bootstrap confidence interval based on expert judgment.

in [12], and the *Rothman index* [13]. To explain a change in prediction between time t_0 and t_1, we consider time-restriction (Def. 4.1) and *time-diffing*. Time-diffing requires not just that the lab result falls in the time window $(t_0, t_1]$ of interest, but also that something about this test has changed since t_0. For example, a continuously high heart rate, although associated with a high risk, does not explain a change in risk. More formally, to compute the time-diffed weight of a lab result in $(t_0, t_1]$, we subtract the odds-ratio/Rothman index from results of the same test on or before t_0 from the current odds-ratio/Rothman index.

Dynamical Attention. Attention-mechanisms provide a way to make recurrent neural networks more interpretable [11]. Attention derives weights for each intermediate state in the RNN by combining it with the final state. The prediction is then produced from the normalized weights sum across the states. Those weight can serve directly as explanations and have been used in the clinical context to identify important diagnosis codes [47, 48].

Odds Ratio and Rothman index. Two statistics commonly used to associate risk with discretizied features, are the odds-ratio, e.g. in [12], and the Rothman index [13].

In particular, we compute the odds of the outcome for one of the ranges of a feature as the ratio of the number of values of the feature that fall into the range for examples with the outcome present vs. absent. We also compute the same odds considering values of this feature outside the particular range and take the ratio of the two odds.

The *Rothman index* is defined as the ratio of the empirical risk associated with a particular range of a feature over the empirical risk of the average feature value.

To lift them to the dynamical setting, in addition to requiring the events to fall into the window of interest $(t_0, t_1]$, we also require those values to have changed since t_0. For example, when we try to explain a risk increase in the past 2 hours, it is not helpful to only point to events that happened more than 2 hours ago or values that have not changed since. This second requirement is important in the case of stateless models generally, since the model does not have the ability to keep track of changes in the input.

Rather than comparing the values of a feature directly which is sensitive to variances across examples and within an example across time, we simply compare their associated weights.

Definition 6.1. Given a sequence of attribution weights $a = a_0, \ldots, a_{t_1}$, we define the *time-diffed explanations* for the change from p_{t_0} to p_{t_1} at time t (with $t_0 < t \leq t_1$) as

$$a_t^{t_0 \to t_1} = a_t - \max_{t' \leq t_0} a_{t'}.$$

Variants of this idea could replace the max over time with the weight of the most recent value until t_0. With time-diffing, explanations of a risk increase from p_{t_0} to p_{t_1} are focused on features whose odds ratio or Rothman index increased between t_0 and t_1.

Properties: Users can understand how the weights are being computed from simple counts over the dataset.

The challenge of confounded irregular measurements is being addressed through normalization by considering the risk of the average lab result or the odds for other results from the same test.

Random strawman. Additionally, a strawman, the *random guessing* method randomly selects three recent inputs conditional on having three different features.

Gradient-based Methods. We implemented an RNN using TensorFlow over normalized lab results (from which outliers were removed) with a single new lab result per step as illustrated in Fig. 3. We take the time-series of normalized observation values, together with an indicator of which lab test is present and an encoding of the time since the last step as input to the RNN as done previously in [17–19]. A notable difference is that no bagging is applied and for each step in the RNN only a single lab result is processed. We predict the label at each time step. Although this does not affect overall accuracy, it allows us to attribute changes in predictions to individual lab values for the time derivatives. We use an LSTM [52] of size 64 with input (0.03), output (0.02) and recurrent (0.01) naive dropout [56]. For optimization we used the Adam optimizer [57] and a learning rate of 0.002 over batches of size 16 and clipped the gradients to 6. These hyperparameters were tuned with a proprietary Bayesian optimization framework. Training was performed using Tensorflow with Tesla P100 GPU. An open-source release of our code accompanies this manuscript.

6.4 Metrics of the Expert Evaluation

We obtain attribution weights for the techniques described above and select the top lab results from 3 distinct lab tests. Our raters included one ICU doctor and two medical students who jointly rated a total of 40 alerts labeling more than 800 lab results. We gave them access to the patient's chart up until 12 hours after admission (around 5000 lab results). We asked to which extent the new lab results indicate that the patient's condition deteriorated since then. The new lab results that were selected as explanations (3 from each method) were presented in a random order to avoid position bias. We average the precision across patients.

The expert raters were asked to give their response on a 5-point Likert scale from "Extremely Likely" to "Extremely Unlikely". We consider explanations marked as "Extremely likely" or "Very likely" as correct. The specific instructions can be found in the Appendix and were accompanied by an example.

Model	AKI Detection Precision @ 1
Random guessing	0.1
Time-restricted gradients	0.52
Time-restricted attention	0.3
Smoothed discrete-time derivatives	0.47
Time-diffed odds ratio	0.4
Time-diffed Rothman index	0.44
Time-restricted odds ratio	0.53
Temporal integrated gradients	**0.55**

Table 2: Precision of the highest weighted lab results for the detection of indicators of acute kidney injury.

6.5 Results for Mortality Risk Prediction

6.5.1 Quantitative Results of the Expert Evaluation. We compare the precision of the three highest weighted lab results from distinct lab tests in Table 1. Overall we see that the temporal integrated gradients and the statistical explanations perform best with an average precision of the top-3 highest-weighted lab results above 0.5. The attention, gradients of the inputs, and smoothed discrete-time derivatives do not perform as well. Their average precision is better than random, however the confidence intervals overlap.

6.5.2 A Qualitative Case Study. A qualitative account by an attending physician highlights strengths and weaknesses of the different methods in a case study of a particular patient.

In one patient presumed sepsis, the risk increased from 62.3% to 86.2%. The time-diffed odds-ratio selected a high temperature of 102.8 degrees Fahrenheit, low urine output, and an oxygen level of 100%. We suspect that such a high oxygen level is more commonly seen in invasively ventilated patient (e.g. with a breathing tube), so it indicates risk indirectly. The human rater selected the first two as convincing evidence of a risk increase. However, we note that in each case, the prior values were all abnormal: the prior temperature was 100.8, oxygen saturation was 100%, and the urine output had been previously low. In the same case, the temporal integrated gradient technique selected low diastolic (55 mmHg) blood pressure, low mean (74 mmHg) blood pressure, and low urine output. These blood pressures, in absolute terms, are somewhat low but not below clinical thresholds that would necessarily require emergent action on their own. However, in the context of this patient, they are relatively low, which does require urgent evaluation to determine the source of clinical change. The prior values were normal at at 72 and 95 mmHg, respectively and had been at those levels for multiple hours. Human raters selected all three values as convincing evidence.

This case study highlights that the odds-ratio technique is excellent at selecting very abnormal values as evidence of increased risk. If the odds-ratio, however, is not adjusted, then we did see likely confounding affecting the thresholds: a high oxygen level indicated high risk because it was confounded by ventilation across the entire dataset. The integrated technique seems, on qualitative inspection, to better select changes in vital signs and lab values with respect to the patient's personal baseline.

6.6 Evaluation of AKI Explanations

To assess the robustness of our methods, we study a second task of AKI detection on the MIMIC-III dataset. The label is defined following a subset of the KDIGO guidelines [58]: If the serum creatinine increases by at least 0.3 mg/dl within the past 48 hours or if the urine volume is less than 0.5 ml/kg/h (25ml/h, assuming 50kg weight) for 6 hours then we say the patient has a positive label. We train models making predictions every 3 hours during the hospital stay, excluding cases when a patient previously had a positive prediction within their encounter. Manual hyper-parameter tuning yielded best results for a learning rate of 0.0003 and a batch size of 64. Other hyper parameters remained the same and were not re-tuned. Note, this is a simple detection task and not a prediction of the future. Therefore, ground-truth explanations are available following the definition of the label. In particular, we consider an explanation correct when it selects any recent urine or creatinine value. Table 2 lists the top-1 precision for 100 randomly chosen positive examples. The temporal integrated gradients remain the best method overall. Other methods differ in effectiveness compared to the mortality prediction task. In particular, all gradient-based methods perform well.

7 CONCLUSIONS AND FUTURE WORK

While much work on model interpretability has focused on explaining risk in a static setting, we introduced a new problem of explaining changes in predicted risk. Our new methods lift static gradient-based attribution techniques [7–10] to this dynamical setting.

We applied our methods to explain clinical alerts of increased mortality risk by identifying three important recent lab values and compare them to attention [11], odds ratio [12] and Rothman index [13]. In our experiments we found that temporal integrated gradients and lifting statistical methods had the highest precision of above 0.5 meaning that among the top 3 explanations physicians would roughly find half of them indicative of the patient's condition worsening. However, the quality of explanations from the odds ratio over discretized lab values was almost as good. Apart from the simple computation of odds ratios, another reason why odds ratios may be preferable in practice is that physicians understand their meaning and are used to considering them in their decision making. A drawback could be that they are not necessarily truthful to the model used to trigger alerts.

Whether a precision of 0.5 will translate to useful explanations for clinical decision making has to be determined through a user study [14] in a clinical setting. Before a deployment, work is needed to ensure that the explanations are reliable [59] and neither misleading [60] nor creating unjustified trust [61]. Informing clinicians about the limitations of the techniques is a starting point. Our study is limited to providing explanations for an increase in risk. Understanding the interaction of model and explanation quality, e.g. whether explanations help identifying false alerts, is left for future work. This work is a first step towards explaining changes in predicted risk and we hope it sparks further ideas, improvements and applications.

REFERENCES

[1] David W Bates and Eyal Zimlichman. Finding patients before they crash: the next major opportunity to improve patient safety. *BMJ Qual. Saf.*, 24(1):1–3, January 2015.

[2] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff.*, 33(7):1123–1131, July 2014.

[3] Vincent Liu, Yan S. Kim, Benjamin J. Turk, Arona Ragins, Brian A. Dummett, Carnen L. Adams, Elizabeth A. Scruth, and Patricia Kipnis. *Early Detection of Impending Deterioration Outside The ICU: A Difference-in-Differences (DiD) Study*, pages A7614–A7614. 2016.

[4] Barbara J Drew, Patricia Harris, Jessica K Zègre-Hemsey, Tina Mammone, Daniel Schindler, Rebeca Salas-Boni, Yong Bai, Adelita Tinoco, Quan Ding, and Xiao Hu. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PLoS One*, 9(10):e110274, October 2014.

[5] Shannon M Fernando, Alexandre Tran, Monica Taljaard, Wei Cheng, Bram Rochwerg, Andrew J E Seely, and Jeffrey J Perry. Prognostic accuracy of the quick sequential organ failure assessment for mortality in patients with suspected infection: A systematic review and meta-analysis. *Ann. Intern. Med.*, February 2018.

[6] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *MLHC*, 2019.

[7] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, August 2010.

[8] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, June 2009. Also presented at the ICML 2009 Workshop on Learning Feature Hierarchies, Montréal, Canada.

[9] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

[10] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. September 2014.

[12] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA, 2015. ACM.

[13] Michael J Rothman, Steven I Rothman, and Joseph Beals, 4th. Development and validation of a continuous measure of patient condition using the electronic medical record. *J. Biomed. Inform.*, 46(5):837–848, October 2013.

[14] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv*, 2017.

[15] Lauren Block, Robert Habicht, Albert W Wu, Sanjay V Desai, Kevin Wang, Kathryn Novello Silva, Timothy Niessen, Nora Oliver, and Leonard Feldman. In the wake of the 2003 and 2011 duty hours regulations, how do internal medicine interns spend their time? *J. Gen. Intern. Med.*, 28(8):1042–1047, August 2013.

[16] Lena Mamykina, David K Vawdrey, and George Hripcsak. How do residents spend their shift time? a time and motion study with a particular focus on the use of computers. *Acad. Med.*, 91(6):827–832, June 2016.

[17] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *CoRR*, abs/1606.01865, 2016.

[18] Zachary C. Lipton, David C. Kale, and Randall C. Wetzel. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In *MLHC*, 2016.

[19] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. May 2017.

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

[21] Robert A Verheij, Vasa Curcin, Brendan C Delaney, and Mark M McGilchrist. Possible sources of bias in primary care electronic health record data use and reuse. *J Med Internet Res*, 20(5):e185, May 2018.

[22] Matti Suistomaa, Aarno Kari, Esko Ruokonen, and Jukka Takala. Sampling rate causes bias in apache ii and saps ii scores. *Intensive Care Medicine*, 26(12):1773–1778, Dec 2000.

[23] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Julier Simon, Raghuveer M. Rao, Troy D. Kelley, David Braines, Murat Sensoy, Christopher J. Willis, and Prudhvi Gurram. Interpretability of deep learning models: a survey of results. In *IEEE Smart World Congress 2017 Workshop: DAIS 2017 - Workshop on Distributed Analytics InfraStructure and Algorithms for Multi-Organization Federations*, 2017.

[24] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. SmoothGrad: removing noise by adding noise. *ICML workshop on visualization for deep learning*, June 2017.

[25] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[26] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014.

[27] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1896–1906, 2018.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[29] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[30] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 883–892, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[31] Lionel A Mandell, Richard G Wunderink, Antonio Anzueto, John G Bartlett, G Douglas Campbell, Nathan C Dean, Scott F Dowell, Thomas M File, Jr, Daniel M Musher, Michael S Niederman, Antonio Torres, Cynthia G Whitney, Infectious Diseases Society of America, and American Thoracic Society. Infectious diseases society of America/American thoracic society consensus guidelines on the management of community-acquired pneumonia in adults. *Clin. Infect. Dis.*, 44 Suppl 2:S27–72, March 2007.

[32] W S Lim, D L Smith, M P Wise, and S A Welham. British thoracic society community acquired pneumonia guideline and the NICE pneumonia guideline: how they fit together. *BMJ Open Respir Res*, 2(1):e000091, May 2015.

[33] Matthew M Churpek and Dana P Edelson. Moving beyond Single-Parameter early warning scores for rapid response system activation. *Crit. Care Med.*, 44(12):2283–2285, December 2016.

[34] Michael D Howell, Long Ngo, Patricia Folcarelli, Julius Yang, Lawrence Mottley, Edward R Marcantonio, Kenneth E Sands, Donald Moorman, and Mark D Aronson. Sustained effectiveness of a primary-team-based rapid response system. *Crit. Care Med.*, 40(9):2562–2568, September 2012.

[35] Julia Adler-Milstein, Catherine M DesRoches, Peter Kralovec, Gregory Foster, Chantal Worzala, Dustin Charles, Talisha Searcy, and Ashish K Jha. Electronic health record adoption in US hospitals: Progress continues, but challenges persist. *Health Aff.*, 34(12):2174–2180, December 2015.

[36] Benjamin Shickel, Patrick Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–1, 2017.

[37] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 322–337, Boston, Massachusetts, 18–19 Aug 2017. PMLR.

[38] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: A convolutional net for medical records. July 2016.

[39] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pages 301–318. jmlr.org, 2016.

[40] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages

507–516, New York, NY, USA, 2015. ACM.

[41] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS One*, 8(6), 2013.

[42] Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm). *J. of Biomedical Informatics*, 54(C):96–105, April 2015.

[43] Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics*, 133(1):e54–e63, 2014.

[44] Mathieu Guillame-Bert, Artur Dubrawski, Donghan Wang, Marilyn Hravnak, Gilles Clermont, and Michael R. Pinsky. Learning temporal rules to forecast instability in continuously monitored patients. *JAMIA*, 24(1):47–53, 2017.

[45] Thomas McCoy Roy Perlis Finale Doshi-Velez Michael C. Hughes, Huseyin Melih Elibol. Supervised topic models for clinical interpretability. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, 2016.

[46] Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H Shah. Improving palliative care with deep learning. November 2017.

[47] Ying Sha and May D. Wang. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, pages 233–240, New York, NY, USA, 2017. ACM.

[48] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3504–3512. Curran Associates, Inc., 2016.

[49] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. 2016:371–380, 02 2017.

[50] Zhengping Che, Sanjay Purushotham, Robinder G. Khemani, and Yan Liu. Distilling knowledge from deep networks with applications to healthcare domain. In *NIPS Workshop on Machine Learning for Healthcare*, 2015.

[51] Hany Farid and Eero Simoncelli. Differentiation of discrete multidimensional signals. *IEEE Transactions on Image Processing*, 13(4):496–508, 4 2004.

[52] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[53] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, 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*, 2016.

[54] G. Michael Felker, Larry A. Allen, Stuart J. Pocock, Linda K. Shaw, John J.V. McMurray, Marc A. Pfeffer, Karl Swedberg, Duolao Wang, Salim Yusuf, Eric L. Michelson, and Christopher B. Granger. Red cell distribution width as a novel prognostic marker in heart failure: Data from the charm program and the duke databank. *Journal of the American College of Cardiology*, 50(1):40 – 47, 2007.

[55] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Peter J Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Gavin E Duggan, Gerardo Flores, Michaela Hardt, Jamie Irvine, Quoc Le, Kurt Litsch, Jake Marcus, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H Shah, Atul J Butte, Michael Howell, Claire Cui, Greg Corrado, and Jeff Dean. Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 1, January 2018.

[56] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1019–1027. Curran Associates, Inc., 2016.

[57] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011.

[58] J. A. Kellum and N. Lameire. Diagnosis, evaluation, and management of acute kidney injury: a KDIGO summary (Part 1). *Crit Care*, 17(1):204, Feb 2013.

[59] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (Un)reliability of saliency methods. *NIPS workshop on Explaining and Visualizing Deep Learning*, 2017.

[60] Zachary Chase Lipton. The mythos of model interpretability. In *ICML Workshop on Human Interpretability*, 2016.

[61] Aaron Springer, Victoria Hollis, and Steve Whittaker. Dice in the black box: User experiences with an inscrutable algorithm, 2017.

9 APPENDIX

Your task is to to judge how much each test and value raise your concern that the patient's health has gotten worse since rounds.

On the right hand side are NEW values that have been collected since you last reviewed the patient's chart. Specifically, for each value in column E-F (test and value), what is the likehood that this information in indicates that the patient's condition is worse since you last looked at the patients chart (i.e. during rounds). You should score your responses in column H from 1-5 with:

5 Extremely likely (i.e. extremely likely to indicate a patient's condition is worse)
4 Very likely
3 Unlikely (does not help me assess more or less likely)
2 Very unlikely
1 Extremely unlikely (i.e. this information likely indicates the patient's condition is better)

Notes. The information in columns A-C is generally sorted by type (there are a few values out of order occasionally due to coding issues) to help you find the data easily. Please bias AWAY from picking 3 (explained below). Choosing 5 does not necessarily mean you think the patient's risk has gone up a significant amount - only that the signal is clear that the risk has gone up unambiguously. In the example above, the respiratory rate of 32 is scored at 4 because respiratory rates are known to be inconsistently recorded, so measurement error might account for that value. However, if a patient's lactate increased from 1.4 to 3 then you should score that as a 5 because it is clearer that the risk has actually increased. Ignore all data not in columns A though H.