HHS Public Access

Author manuscript

Nat Biomed Eng. Author manuscript; available in PMC 2019 April 16.

Published in final edited form as:

Nat Biomed Eng. 2018 October ; 2(10): 749–760. doi:10.1038/s41551-018-0304-0.

Explainable machine-learning predictions for the prevention of hypoxaemia during surgery

Scott M. Lundberg[1], **Bala Nair**[2,5,6], **Monica S. Vavilala**[2,5,6], **Mayumi Horibe**[4], **Michael J. Eisses**[2,3], **Trevor Adams**[2,3], **David E. Liston**[2,3], **Daniel King-Wai Low**[2,3], **Shu-Fang Newman**[2,5], **Jerry Kim**[2,3], and **Su-In Lee**[1,*]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA.

[2]Department of Anesthesiology and Pain Medicine, University of Washington, Seattle, WA, USA.

[3]Seattle Children's Hospital, Seattle, WA, USA.

[4]Veterans Affairs Puget Sound Health Care System, Seattle, WA, USA.

[5]Center for Perioperative and Pain initiatives in Quality Safety Outcome, University of Washington, Seattle, WA, USA.

[6]Harborview Injury Prevention and Research Center, University of Washington, Seattle, WA, USA Seattle, WA, USA

Abstract

Although anaesthesiologists strive to avoid hypoxemia during surgery, reliably predicting future intraoperative hypoxemia is not currently possible. Here, we report the development and testing of

*To whom correspondence should be addressed: suinlee@cs.washington.edu.

Author contributions

S.-I.L., S.M.L., B.N., and J.K. initiated the study. S.-I.L. and S.M.L. developed the Prescience algorithms and designed data analyses and experiments. S.M.L. performed data analyses, experiments, and data preprocessing. B.N., and S.-F. N. provided the electronic medical record data. J.K. recruited anaesthesiologists and helped design the anaesthesiologist test and survey. M.H., M.J.E., T.A., D.E.L., D.K.-W.L. performed the web-based anaesthesiologist experiments and provided survey data. M.H. provided manuscript feedback. M.S.V. provided clinical assessment, interpretation of feature importances, and connections with anaesthesiologists' workflow. S.-I L., S.M.L. wrote the paper in conjunction with B.N., J.K., and M.S.V. who wrote sections on clinical interpretation and integration with current practices.

Reporting Summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article

Code availability

The model-explanation code originally used for Prescience is available open-source (and since improved) at https://github.com/slundberg/shap. Modeling code, processing code, and web-interface code specific to Prescience is available for reference purposes at https://gitlab.cs.washington.edu/prescience.

Data availability

Owing to patient-privacy considerations, the operating-room datasets from participating hospitals are not publicly available. The raw data from the anaesthesiologist comparisons in Fig. 3 is available in Supplementary Tables 6 and 7, and data from Fig. 5 is available in Supplementary Tables 8 and 9.

Competing interests

Bala Nair is an advisor for Perimatics LLC and holds equity in the company. Daniel Low is a Chief Medical Officer for MDmetrix, Inc. The other authors declare no competing interests.

a machine-learning-based system that, in real time during general anaesthesia, predicts the risk of hypoxemia and provides explanations of the risk factors. The system, which was trained on minute-by-minute data from the electronic medical records of over fifty thousand surgeries, improved the performance of anaesthesiologists when providing interpretable hypoxemia risks and contributing factors. The explanations for the predictions are broadly consistent with the literature and with prior knowledge from anaesthesiologists. Our results suggest that if anaesthesiologists currently anticipate 15% of hypoxemia events, with this system's assistance they would anticipate 30% of them, a large portion of which may benefit from early intervention because they are associated with modifiable factors. The system can help improve the clinical understanding of hypoxemia risk during anaesthesia care by providing general insights into the exact changes in risk induced by certain patient or procedure characteristics.

Over 300 million surgeries are performed worldwide every year *(1)*. Although an integral part of healthcare, surgery and anaesthesia pose considerable risk of complications and death. Studies have shown a perioperative mortality rate of 0.4 to 0.8% and a complication rate of 3 to 17%, just in industrialized countries *(2, 3)*. Fortunately, half of these complications are preventable *(2, 3)*. With increasing adoption of electronic medical record systems, high fidelity heterogeneous data are being captured during surgery and anaesthesia care, yet the utilization of this data to improve patient safety and quality of care remains poor *(4)*. There is untapped potential for data science to utilize perioperative data to positively impact surgical and anaesthesia care *(5)*. To address this unmet need we leverage recent advances in perioperative informatics and present new machine learning methods to predict harmful physiological events and to inform anaesthesiologists.

Hypoxemia or low arterial blood oxygen tension is an unwanted physiological condition known to cause serious patient harm during general anaesthesia and surgery *(6)*. Hypoxemia is associated with cardiac arrest, cardiac arrhythmias, postoperative infections and wound healing impairments, decreased cognitive function and delirium, and cerebral ischemia through a number of metabolic pathways *(7)*. Despite the advent and use of pulse oximetry to continuously monitor blood oxygen saturation (SpO_2) during general and regional anaesthesia, hypoxemia can neither be reliably predicted nor prevented at future time points *(8)*. Real-time blood oxygen monitoring through pulse oximetry only allows anaesthesiologists to take reactive actions to minimize the duration of hypoxemic episodes after occurrence. Decision support systems that process electronic medical record data have been shown to help increase adherence to guidelines, but remain primarily reactive rather than predictive in nature *(9, 10)*; see *(4)* for a full review. If hypoxemia can be predicted or anticipated before it occurs, then actions can be taken by anaesthesiologists to proactively prevent hypoxemia and minimize patient harm.

Machine learning (ML) techniques use statistical methods to infer relationships between patient attributes and outcomes in large datasets, and have been successfully applied to predict adverse events in health care settings, such as sepsis, or patient deterioration in the intensive care unit *(11–15)*. Yet ML techniques to predict adverse events such as hypoxemia in a considerably more complex setting such as the operating room are currently lacking. Moreover, though previous complex ML approaches provide good prediction accuracy, their

application in an actual clinical setting is limited because their predictions are difficult to interpret, and hence not actionable. Interpretable methods explain *why* a certain prediction was made for a patient, i.e., specific patient characteristics that led to the prediction. This lack of interpretability has thus far limited the use of powerful methods such as deep learning and ensemble models in medical decision support.

We present an ensemble model based machine learning method, *Prescience*, that predicts the near-term risk of hypoxemia during anaesthesia care *and* explains the patient and surgery specific factors that led to that risk (Figure 1). We believe this is an important step forward for machine learning in medicine because while machine learning models have significantly improved the ability to predict a patient's future condition *(*16, 17*)*, the inability to explain the predictions from accurate, complex models is a serious limitation. Understanding what drives a prediction is important for determining targeted interventions in a clinical setting. For this reason, machine learning methods employed in clinical applications avoid using complex, yet more accurate, models and retreat to simpler interpretable (e.g., linear) models at the expense of lower accuracy. To address this problem, some approaches have achieved interpretability by carefully limiting the complexity of the machine learning model *(*15*)*. In contrast, we demonstrate how to retain interpretability, even when complex models such as nonparametric methods or deep learning are used, by developing a method to provide theoretically justified explanations of model predictions that build on recent advances in model-agnostic prediction explanation methods *(*18–21*)*. This allows these accurate, but traditionally hard to interpret, models to be used while still providing intuitive explanations of what led to a patient's predicted risk. The ability to provide simple explanations of predictions from arbitrarily complex models helps eliminate the typical accuracy vs. interpretability tradeoff, thus allowing broader applicability of machine learning to medicine.

Prescience was trained to use standard operating room sensors to predict hypoxemic events in the near future and explain why an event is, or is not, likely to occur. It departs from the relatively few previous approaches to this problem in two important ways:

First, unlike previous approaches that used a linear autoregressive support vector machine on arterial oxygen saturation times series *(*11*)* and that used Parzen windows to find outliers from five input patient measurement types *(*22*)*, Prescience integrates a comprehensive dataset from a hospital's Anaesthesia Information Management System (AIMS) (see Methods for details). While some operating room forecasting approaches have relied on simulated physiology *(*23*)*, the AIMS data consists of high fidelity *real-time data* – such as time series data from patient monitors and anaesthesia machines, bolus and infusion medications, input and output fluid totals, laboratory results, templated and free text descriptions of anaesthesia techniques and management, and *static data* – such as American Society for Anesthesiology (ASA) physical status, surgical procedure and diagnoses codes *(*24*)*, as well as patient demographic information such as age, sex, smoking status, height and weight. Continuously integrating a broad set of patient and procedure *features* extracted from the AIMS data, Prescience surpasses human-level accuracy while maintaining consistent performance during every minute of a surgery.

Second, Prescience explains why a prediction was made, regardless of the complexity of the machine learning model used to make the prediction. Significant progress has been made recently integrating predictive machine learning solutions into medical care (11–14). However, accurately and intuitively conveying to doctors *why* a prediction was made remains a key challenge. For example, a numeric representation of risk is useful (e.g., the 2.4 odds ratio in Figure 1). However, a more detailed presentation that shows the risk is due to the patient's BMI, current tidal volume, and pulse rate is more clinically meaningful since some factors may be modifiable and result in clinical changes mitigating that risk (Figure 1). Typically, understanding why a prediction was made requires limiting the complexity of the model (15), but Prescience enables explanations for models of arbitrary complexity. The feature impact values computed by Prescience essentially represent the change in the model's predicted risk when we observe a feature (such as a patient's weight) vs. when we do not observe the feature (such as not knowing a patient's weight). This change in a model's output prediction when a feature is observed indicates its importance for the prediction. Feature importances do not imply a causal relationship, and so do not represent a complete diagnosis of hypoxemia in a patient. However, they do enable an anaesthesiologist to better formulate a diagnosis by knowing which attributes of the patient and procedure contributed to the current risk predicted by the machine learning model.

Results

To demonstrate the value of Prescience's explained predictions and gain insight into factors affecting intraoperative hypoxemia, we present the following results: 1) a comparison of Prescience hypoxemia predictions against anaesthesiologists' predictions with and without the aid of Prescience, 2) an example of how Prescience explains hypoxemia risk at a specific time-point during a surgical procedure, 3) a comparative summary of relevant AIMS data features for hypoxemia prediction chosen by Prescience and by anaesthesiologists, and 4) a detailed presentation of key risk factors for hypoxemia identified by Prescience.

Prescience overview – data preparation, model learning and feature importance estimation

Based on World Health Organization recommendations and for the purposes of prediction, we defined hypoxemia as the decrease in SpO_2, i.e., arterial blood oxygen saturation as measured by pulse oximetry, to a threshold value of 92% or lower (see Methods; Supplementary Figure 1). From the AIMS data, we extracted 3,797 *static extracted features* for each patient from 20+ original static sources and an expanded superset of 3,905 *real-time* and *static extracted features* for each time point during anaesthesia care from the 20+ original static sources as well as 45 different real-time data sources (see Methods; Supplementary Table 1). Features such as words from text data get directly mapped to the display in Figure 1, while sets of features from a single time series (such as tidal volume) are combined in Figure 1. We excluded select cases (heart transplant, lung transplant, tracheostomy, and coronary artery bypass surgeries) in which SpO_2 and other hemodynamic parameters can be significantly affected by non-physiological measurements such as during cardiopulmonary bypass. All the experiments were performed after appropriate Institutional Review Board (IRB) approval (see Methods), with clinical data summarized in Figure 2.

We trained a gradient boosting machine model *(26)* to solve the following two types of prediction problems: A) *initial prediction*: predicting at the start of a procedure the risk of hypoxemia anytime during a procedure based on the static extracted features, and B) *real-time prediction*: predicting hypoxemia in the next 5 minutes at various points of the operative period based on real-time and static extracted features collected up to that time point. We chose this 5 minute window to be long enough that an anaesthesiologist could have time to intervene, while also keeping the window short enough that it represents near term risks which would benefit from immediate attention. For task A) we used 42,420 procedures (each a single surgical case) as *training samples* to train the gradient boosting machine, 5,649 procedures as *validation samples* to choose the tuning parameters for the gradient boosting machine (and other prediction models for comparison), and 5,057 as *test samples* for comparing across different prediction models (Supplementary Figure 2). For task B), we used 8,087,476 per-minute time points as training samples, 1,053,629 as validation samples, and 963,674 as test samples, where all time points from the same procedure were included in the same sample set and no missing data imputation was performed (Supplementary Figure 3). Dividing the time points by procedure is important since samples from the same procedure are not independently and identically distributed but have some time dependence. To ensure that there was no bias towards the final test set, the test data was initially compressed and left compressed until method development was completed.

As shown in Supplementary Figures 2 and 3, the gradient boosting machine outperforms alternative prediction models previously used for similar problems, particularly for the primary task of real-time prediction.

For tasks A) and B), we use 198 and 523 test samples respectively for evaluating anaesthesiologists' performance for initial and real-time prediction tasks, respectively (see below). Prescience outputs the risk prediction and its explanations (Figure 1; Figure 4A) which show a set of features that increased (purple-colored) and decreased (green-colored) the risk.

We developed an efficient, theoretically justified machine learning technique to estimate the importance of each feature on a prediction made for a single patient, which drives real-time explanations (Figure 4) for the Prescience model. We verified the quality of the explanations given to the anaesthesiologists (in the experiments described below) by comparing the explanations with the change in model output when a feature is perturbed (Supplementary Figure 4). We also developed effective visualizations of these explanations that encodes them in a compact visual form for anaesthesiologists (Figure 1; Supplementary Figures 5-7), and a more detailed visualization which highlights the relevant contributing features (Figure 4) (see Methods for details).

Prescience improves anaesthesiologist's ability to predict hypoxemia

To test the potential of Prescience to aid hypoxemia prediction we replayed prerecorded intraoperative data from test sample procedures in a web-based visualization to five practicing anaesthesiologists (Supplementary Figures 5-7). Each anaesthesiologist was given two types of prediction tasks: A) *initial prediction* (198 tasks), and B) *real-time prediction*

(523 tasks). For each prediction task, anaesthesiologists were asked to provide a *relative risk* of hypoxemia as compared to a normal acceptable risk, for example, 0.01 for 1/100th the normal risk or 3.4 for 3.4 times the normal risk. These relative risks were then used to calculate standard receiver operating characteristic (ROC) curves averaged over five anaesthesiologists as shown in Figure 3, which plots the true positive rate (i.e., % of desaturations correctly predicted) in the y-axis against the false positive rate (i.e., % of non-desaturations incorrectly predicted) in the x-axis. Note that ROC curves only depend on the order of the relative risk values among predictions from a single anaesthesiologist. This eliminates the need to choose a threshold and the need to separately calibrate risk scores between anaesthesiologists.

Figure 3A-B shows that for both types of prediction tasks, predictions made by Prescience (purple) are considerably more accurate than anaesthesiologists' predictions (green). The prediction accuracy of anaesthesiologists (green) markedly improved when the anaesthesiologists were given Prescience's risk prediction and its explanations in addition to the original procedure data (blue) (Supplementary Figures 5-7). A clear separation between the performance of anaesthesiologists with (blue) and without (green) the aid of Prescience is observed for both initial prediction (Figure 3A, P-value < 0.0001) and real-time prediction (Figure 3B, P-value < 0.0001). This suggests that Prescience can enhance anaesthesiologists' assessment of future risk and their ability to proactively anticipate hypoxemia events. Interestingly, the prediction performance of anaesthesiologists with Prescience explanations (blue) was slightly lower than direct predictions from Prescience (purple). This means that when the anaesthesiologists adjust their risk estimate for a patient away from what Prescience originally predicted they are more likely to be wrong than right.

To avoid the scenario in which an anaesthesiologist is tested on the same prediction task twice – one with and the other without Prescience, we created replicate test sets by dividing the prediction tasks into two groups of similar size: (100, 98) tasks for initial prediction and (260, 263) tasks for real-time prediction. Each of the five recruited anaesthesiologists was assigned to receive Prescience's assistance in one of these two replicate test sets (Methods). The procedures shown to anaesthesiologists were chosen such that ~50% showed at least one incident of hypoxemia (for preoperative prediction), and time points were chosen such that ~33% had hypoxemia in the next 5 minutes (for intraoperative prediction). The anaesthesiologist test time points for hypoxemia were chosen to be drops of SpO_2 (Supplementary Figure 1) with a preceding period of stable and normal SpO2. However, the entire dataset also includes easier to predict hypoxemic events that follow prior SpO2 drops and decreasing SpO2 trends. For the entire dataset, Prescience achieves an even higher area under the ROC curve of 0.90 (Supplementary Figure 3), and when using a larger training dataset achieves an area under the ROC of 0.92 (see below; Supplementary Figure 9).

If we extrapolate the real-time results to the 30 million annual surgeries in the US under the assumption that doctors anticipate 15% of hypoxemic events while SpO_2 is still ≥ 95, then with Prescience assistance they may be able to anticipate 30% of these events, or approximately 2.4 million additional episodes of hypoxemia annually (defined here as SpO_2 ≤ 92). Since 20% percent of the Prescience risk prediction is based on drugs and settings under the control of the anaesthesiologist (Supplementary Table 5;

STable5_RealtimeFeatures.csv), a large portion of these predicted events may benefit from early intervention. Note that these estimates are based on retrospective data from an AIMS system, the addition of non-AIMS data available in the operating room, such as waveforms, may improve the performance of both anaesthesiologists and Prescience.

The anaesthesiologists consulted had experience after residency ranging from 3 to 26 years (median 7 years), and were all actively practicing at University of Washington Medical Center, VA Puget Sound Health System or Seattle Children's hospital. The extensive experience level of the anaesthesiologists who participated in our study may not represent the typical experience level of anaesthesia providers, especially when nurse anaesthetists and residents in training provide anaesthesia care.

When using Prescience predictions to generate early warning alarms in the operating room, it is important to minimize false alarm rates. This can be accomplished by adjusting the tradeoff between precision (the positive predictive value) and recall (the sensitivity). High precision means a low false alarm rate (1 - precision), however, it comes at the cost of low recall. Supplementary Figure 10 plots the precision and recall tradeoff for Prescience on the full set of test time points. Since the performance of the complex model in Prescience improves with larger datasets, we also included results from a model trained on an expanded 175 thousand procedure dataset to measure the benefit of using more data to train Prescience. The larger dataset resulted in notably better performance and could capture 9% of all minutes with upcoming hypoxemia at 70% precision, (or 44% of all minutes with upcoming hypoxemia at a precision of 30% if the threshold for precision-recall tradeoff is selected for higher recall). These are strikingly higher precisions than those we project anaesthesiologists would achieve on the full test data set (Supplementary Figure 10). We also note that the predictive capacity can be further improved by shortening the predictive window to less than 5 minutes (Supplementary Figure 9).

Anaesthesiologists must not only decide when to act to prevent hypoxemia, but also when not to act. To assist in this, Prescience can predict not just when hypoxemia will occur, but also when it will not occur. Prescience can predict when hypoxemia will not occur for 60% of all time points while maintaining a precision of 99.9% (Supplementary Figure 11).

Explained risks reveal both procedure and time specific effects

An explanation from Prescience represents the effects of interpretable groups of extracted patient features (see Figure 1 and Figure 4A), where each group corresponds to the set of extracted patient features from a single input feature in the AIMS dataset, such as the SpO2 monitor time series. These effects explain why the model predicted a specific risk, and thus allow an anaesthesiologist to plan appropriate interventions. In Figure 1 only the most significant features contributing to hypoxemia risk are shown for quick reference, however in Figure 4 the relative contributions of all patient and case features (i.e. attributes) towards hypoxemia risk can be seen at every sample time point during a procedure (Figure 4B). Without a meaningful explanation, the sudden increase in risk shown at the time point marked 'Now' might be hard to interpret; however, by representing the predicted risk as a cumulative effect of contributing patient and procedure features, the reason for the increase becomes clear (Figure 4A).

The increase in the risk of hypoxemia in the next 5 minutes shown in Figure 4 is driven by a set of features capturing both static attributes, such as patient height and weight, and dynamic parametric values, such as tidal volume (i.e., volume of gas exhaled per breath) and administration of drugs. The risk explanation bar in Figure 4A has purple-colored features that push the risk higher (to the right) and green-colored features that push the risk lower (to the left). Each group of features is sorted by the magnitude of their impact, and the largest impact features are labeled. Through this representation we can see that many of the 3,905 real-time extracted features have only a small impact, and the risk for this time point is predominantly driven by a few features. The choice of features provided to the model was driven by the data recorded in the AIMS system and hence was available for training. Rather than only provide the model with features we believed important, we let the model use any feature it chose. This means that it may find features we would not initially expect are predictive of hypoxemia. For some of these features it is helpful before final deployment in an operating room to tag them with indicators of how they relate to hypoxemia risk. This can help anaesthesiologists quickly see non-obvious connections with patient physiology, such as how the muscle relaxant succinylcholine in Figure 4 does not represent a direct causal impact on hypoxemia, but rather is a proxy that captures the risk from a potentially difficult airway or full stomach (in the hospital system we considered, succinylcholine is given to patients with a high risk for a difficult airway during intubation). Figure 4B shows the trend in the Prescience risk predictions over the course of the procedure. The plot in Figure 4B is equivalent to rotating the feature explanation in Figure 4A by 90 degrees and then stacking the explanations for each time point horizontally. We can see from the risk trend in Figure 4B that the large increase in risk at the current time was driven by ‘Tidal volume’, meaning a drop in the patient’s tidal volume. The future SpO_2 (blood oxygen concentration) measurements confirm that the patient did indeed progress to hypoxemia (i.e., $SpO_2 \leq 92$). Not only does Prescience alert anaesthesiologists when a patient’s risk for hypoxemia is high, but also provides information on the factors and their relative contributions driving the risk. This informed risk prediction enables anaesthesiologists to plan an appropriate course of action to avoid hypoxemia.

Averaged feature importance estimates broadly align with a survey of prior expectations

To gain an understanding of the general impact of features across all procedures, we computed the average importance of each feature in the Prescience model. In contrast to the explanations shown in Figure 1 and Figure 4A which are specific to a single prediction at a particular time point, these average feature importance estimates are over many procedures and time points *(27)*. These averaged feature importance estimates are shown for both initial prediction (Figure 5A) and real-time prediction (Figure 5B).

To estimate which clinical features anaesthesiologists use to estimate hypoxemia risk, we first performed a survey before using Prescience, which asked four anaesthesiologists to list the most important factors they consider when assessing the risk of hypoxemia, both before (for initial prediction) and during a procedure (for real-time prediction). Their responses were then aggregated into a single ranked list of features (Supplementary Tables 2 and 3). Figure 5 shows the rankings chosen by anaesthesiologists next to the feature importance

estimates derived by Prescience for (A) initial and (B) real-time predictions. The ranking of features by anaesthesiologists appears to correspond well with the ranking by Prescience.

As another way to measure which features anaesthesiologists' think contribute to hypoxemia, we learned from anaesthesiologists' behavior by training a separate gradient boosting machine model based on their predictions. This allows a direct comparison between the anaesthesiologists and Prescience on the same set of features. We fit this model to all the anaesthesiologist relative risk predictions using 10-fold cross validation. We then computed the feature importance estimates for this model that was trained to mimic the behavior of anaesthesiologists. Given the relatively smaller set of training examples used to train the model (198 initial predictions, and 523 real-time predictions), we used bootstrapping to estimate the variability of the feature importance estimates (Figure 5 right).

In general, there is reasonable agreement between the Prescience feature importance estimates and those identified by the anaesthesiologists. However, there are important differences that may stem from the comprehensive nature of the Prescience analysis, while anaesthesiologists necessarily focus on what they consider the most likely causes for hypoxemia concern. One striking difference is the reduced role of current SpO_2 levels in anaesthesiologists' predictions. While anaesthesiologists are clearly influenced by the recent patterns of patient SpO_2 levels, Prescience strongly depends on these patterns, while anaesthesiologists appear to be equally influenced by other factors, such as end tidal CO_2 and peak ventilation pressure. The second and fourth ranked features by anaesthesiologists for initial prediction were lung disease and asthma respectively and did not show up as important features for Prescience. This is potentially because they must be extracted from preoperative text notes and only about 1% of the procedures recorded the term "COPD" for example, and only 3% of case notes mention "asthma".

Our study used data from two hospitals and initial hypoxemia predictions were driven by a bias between the two hospitals. This is perhaps unsurprising since one hospital is a level 1 trauma center and a significant proportion of its surgical cases involve trauma patients who are more susceptible to hypoxemia. However, it is interesting to note that the importance of hospital as a risk factor became insignificant for the intraoperative real-time predictions, presumably because the risk differences in each hospital were captured by the real-time features.

Among the static features, BMI (body mass index) and age were significant risk factors. These features are well understood in the medical literature as risk factors that can increase the chances of hypoxemia *(*28, 29*)*. The American Society of Anesthesiology (ASA) physical status feature represents the severity of a patient's medical condition and a higher ASA number indicates a higher comorbidity. Prescience determined that higher ASA physical status values predisposes a patient to higher hypoxemia risk. While this finding may be clinically intuitive, anaesthesiologists can now use this information in their preoperative evaluation as a pre-specified risk factor for intraoperative hypoxemia. Eye procedures were informative to the model and carried a reduced risk of hypoxemia, while surgeries for fractures had a slightly higher risk. These patterns may reflect the composite risk of hypoxemia to patients undergoing these particular procedures. In the case of eye

surgeries, the risk was lower even though many are elderly and have accompanying co-morbid conditions. Together these findings provide new data on the relative "risk" of these procedures which has implications for anaesthesia staffing, need for equipment, and preparation for the ability to rescue patients from hypoxemia. Eye procedures and surgeries for fractures are two examples of text-based features extracted from diagnosis and preoperative procedure notes. They demonstrate that unstructured text notes can be combined with structured patient data to improve patient risk prediction. Although many of the risk factors identified by Prescience reconfirmed those expected by the anaesthesiologists, it is informative that Prescience independently identified these features with no prior knowledge.

Among real-time (intraoperative) features SpO_2 (arterial oxygen saturation) is, as expected, the strongest predictor of future potential decreases in SpO_2. End tidal CO_2 (amount of carbon dioxide exhaled by the patient) was also a significant intraoperative feature identified by Prescience as predictive of hypoxemia. Lower values may indicate inadequate ventilation or airway obstruction which can in-turn increase the risk of hypoxemia. Prescience also determined that hypotension (systolic blood pressure below 80mmHg) increases the risk of hypoxemia. On the other hand, higher FiO_2 (inspired O_2 concentration) and positive pressure ventilation can reduce the risk of hypoxemia, as expected by anaesthesiologists.

Prescience's estimated importance of individual features on hypoxemia risk highlight important clinical relationships

Three important features each for beginning of anaesthesia care (initial) and during anaesthesia care (real-time) predictions were chosen to illustrate how the Prescience model modifies hypoxemia risk based on changes to feature characteristics (Figure 5). While many such relationships are present for the various features, Figure 6 shows a representative selection demonstrating informative risk relationships that are captured in the Prescience model.

Among static features we find that patient BMI has a clear effect on the risk of hypoxemia. When the BMI is over 26, the risk of hypoxemia increases linearly until it has more than doubled when BMI is over 50. Though a qualitative association between hypoxemia and body weight is well established in the field of anaesthesia *(*28, 29*)*, Prescience quantifies this relative risk.

Prescience shows that patients with higher ASA physical status codes have higher risk of intraoperative hypoxemia. This is not surprising since higher ASA codes represent increased severity of a patent's physical condition such as preexisting pulmonary and cardiac conditions that can predispose a patient to develop hypoxemia. Prescience data support clinical observations that the ASA status' effect on risk of hypoxemia more than doubles when the ASA status increases from I to V. Advancing age also predicted intraoperative hypoxemia, likely representing the presence of comorbidities *(*28*)*. These data show that BMI > 30, which meets the clinical definition of obesity *(*30*)*, is associated with intraoperative hypoxemia, suggesting impaired pulmonary mechanics. While we agree that these findings confirm clinical observations and suspicions of the relationship between these

patient factors and adverse anaesthesiology outcomes, Prescience quantifies this association and the risks, giving a more clinically useful interpretation to anaesthesiologists.

For real-time prediction, measurements from each time series are represented by a set of multiple features. For simplicity, we focus here only on the effect of the shortest time lag exponentially weighted moving average, which essentially represents the most recent reported value in the time series (see Methods for details).

Tidal volume represents the amount of gas exhaled per breath when the patient is either breathing spontaneously or mechanically ventilated during general anaesthesia. As the tidal volume drops below 0.6 liters (keeping all other features the same), Prescience risk for hypoxemia increases. This increase could be due to hypoventilation, in which case anaesthesiologists take preventative steps to avoid inadequate ventilation.

End tidal CO_2 represents the amount of carbon dioxide exhaled gas. Figure 6 shows the relationship between end tidal CO_2 and risk of hypoxemia under general anaesthesia. End tidal CO_2 below 35 mmHg is associated with an increasing risk of intraoperative hypoxemia. While we cannot definitively attribute hypocapnia with intraoperative hypoxemia, these associations may represent underlying patient conditions such as chronic obstructive pulmonary disease that affect both physiological conditions. Alternately, the low end tidal CO_2 and may result from either intentional or unwanted hyperventilation during anaesthesia care.

Examining FiO_2 is important because anaesthesiologists can control the amount of oxygen delivered to patients. Current practice not to provide all patients with 100% FiO_2 because not all patients need it, prolonged ventilation with 100% FiO_2 is associated with pulmonary atelectasis, and delivering oxygen when it is not needed is costly and wasteful. These data show that FiO_2 below 40% is independently associated with intraoperative hypoxemia irrespective of other features. These findings provide important information regarding safe practice of FiO_2 in patients during general anaesthesia. It is possible that the routine practice of maintaining FiO_2 30% or close to room air may be harmful to patients and not desirable. While these effects are adjusted for all other available features, it is important to note that as with any observational study some residual confounding with patient risk may still exist. This could explain the increase in hypoxemia risk we observed for high O_2 levels.

These representative features illustrate the ability of our machine learning-based prediction method to not only provide explained risk predictions for a complex model, but also quantitative insights into the exact change in risk induced by certain patient or procedure characteristics.

Discussion

Prescience is designed to comprehensively integrate high fidelity oerating room data to predict intraoperative hypoxemia events before they occur. Based on a comparison against practicing anaesthesiologists and existing computational methods applied to other clinical problems, Prescience achieves superior performance when predicting hypoxemia risk from electronically recorded intraoperative data.

Prescience combines high accuracy complex models with interpretable explanations. This combination of accuracy and interpretability allows physicians to receive the best possible predictions while also gaining insight into why those predictions were made. To test how Prescience predictions with explanations would impact an anaesthesiologist's ability to estimate hypoxemia risk we compared anaesthesiologist predictions with and without Prescience assistance. We observed a clear increase in prediction accuracy when doctors were assisted by Prescience, demonstrating that anaesthesiologists may make more accurate hypoxemia risk assessments in the operating room if they had access to Prescience. Prescience based technology may also be an important tool to account for the variation in knowledge and or practice among providers.

Empirically derived black box algorithms such as the bispectral index have been used to track brain states of patients undergoing general anaesthesia by processing real-time EEG *(*31, 32*)*. These algorithms have been criticized because they do not utilize physiological models, do not identify factors associated with risk of events, and produce empirically derived metrics to represent neurophysiology of how the anaesthetics affect the brain. The black box nature of the EEG algorithms has made it difficult to interpret their output and understand how physiological mechanisms and anaesthetic states determine the algorithm output. A similar danger exists with the application of complex black-box machine learning models in the operating room, where predictions are difficult to interpret, and hence less actionable. Prescience demonstrates a solution that promises to avoid the obscurity traditionally associated with black-box models, and instead maintain interpretability even as increasingly complex machine learning models are applied to operating room decision support.

It should be clarified that our exercise at developing machine learning methods to predict intraoperative hypoxemia, though promising, should still be considered an initial attempt. In this first attempt, we did not categorize procedures to assess hypoxemia predictions in specific types of procedures. For this reason, clinical interpretation of the results had to be somewhat generic. For enhanced interpretation of risks, future attempts can focus on specific categories of cases and phases of anaesthesia. Another future enhancement would be integrating additional preoperative data such as a patient's detailed medical history into the prediction models. Higher fidelity intraoperative data such as patient monitor waveform data could enrich machine learning, thus potentially leading to more accurate predictions. Prospective trials of Prescience during live procedures are also needed before deployment to verify the improvements in anaesthesiologist's performance we retrospectively observed in prerecorded procedures *(*33*)*.

This paper focuses on hypoxemia risk during intraoperative anaesthesia care. However, the importance of coupling accurate predictions from complex models with interpretable explanations of why a prediction was made, has broad applicability throughout medicine. To support this we have open-sourced the explanation tools used in Prescience, and are continuing to improve and extend them (http://github.com/slundberg/shap). Because Prescience effectively decouples the interpretable explanation from the prediction model, we are also free to continue to refine the core prediction model without changing the user experience for anaesthesiologists.

The global risk profiles learned by Prescience (Figures 5–6) are clinically relevant for a number of reasons. First, they show that in the health system examined, trauma hospital patients may be more critically ill as they have more intraoperative hypoxemia. In current times when harmonization of care and standardization are considered to reduce unwanted clinical variation, these data suggest that resources may need to be differentially deployed to address differential rates of adverse events. Second, anaesthesiologists can now quantify risks of intraoperative hypoxemia adjusted for other factors to the very elderly, those who are overweight, and those with more comorbid conditions. The exact relationships described in Figure 6 clearly show the patterns and threshold points for the risk. Whereas low tidal volume is often suggested for patients with acute lung injury *(*34*)*, these data suggest that overall, low lung tidal volumes are, in-fact, associated with intraoperative hypoxemia. The relationship between low end-tidal CO_2 levels and intraoperative hypoxemia may reflect underlying critical illness. Despite our inability to fully exclude residual confounding, these data shed new light on physiological relationships as well as provide a mechanism to facilitate provision of anaesthesia care that can mitigate intraoperative hypoxemia.

As a limitation we acknowledge that there are several clinical diagnoses that are associated with hypoxemia, but not directly observable in Prescience. The main clinical diagnoses include mainstem intubation, mucus plug, low FiO2, low tidal volume, tracheal tube balloon leak, patient factors like COPD from smoking, and pulmonary embolus. Among these only low FiO2 and low tidal volume are directly observable in Prescience since the other data elements are not fully captured in the clinical databases. In these cases, secondary risk indicators will show up in Prescience. The differential diagnosis of hypoxemia could also have been categorized using ACLS strategies. However, as opposed to a study where factors are *a priori* identified, Prescience considers all available factors, and renders an output with associated relative risks. Clinicians must then evaluate the feature relevance based on context and clinical relevance.

The field of medicine is full of data science challenges that have the potential to fundamentally impact the way medicine is practiced. More and more data driven predictions of patient outcomes are being proposed and used. However, black-box prediction models which provide simply predictions, without explanation, are difficult for physicians to trust and provide little insight about how they should respond. The interpretable explanations used by Prescience represent a technique that can transform any current prediction method from one that provides *what* the prediction is, into one that also explains *why*.

Methods

IRB statement

The electronic data for this study was retrieved from institutional electronic medical record and data warehouse systems after receiving approval from the Institutional Review Board (University of Washington Human Subjects Division, Approval #46889). Protected health information was excluded from the data set that was used for machine learning methods.

Data sources

Our hospital system has installed an Anaesthesia Information Management System (AIMS) (Merge AIM, Merge Inc, Hartland, WI) that automatically captures minute by minute hemodynamic and ventilation parameters from the patient monitor and the anaesthesia machine. The system also integrates with other hospital electronic medical record (EMR) systems to automatically acquire laboratory and patient registration information. The automatic capture of data is supplemented by manual documentation of medications and anaesthesia interventions to complete the anaesthesia record during a surgical episode. For the current project, we extracted the high-fidelity anaesthesia data from the AIMS database for the period May 2012 through June 2014. Additionally, for each patient, medical history data were extracted from our EMR data warehouse (Caradigm, Bellevue, WA). The high-fidelity anaesthesia record data and the corresponding medical history data from the hospital EMR formed the underlying data for machine learning. The various data elements used for machine learning are outlined in Supplementary Table 1.

SpO_2 desaturation labels

We considered $SpO_2 \leq 92\%$ as hypoxemia, which falls between the World Health Organization's recommended intervention level (< 94%) and emergency level (< 90%) *(35)*. Predictions of hypoxemia were made for a window 5 minutes into the future. If the SpO_2 was ≤ 92% at any point during those 5 minutes then it was considered a positive label, otherwise it was negative. The machine learning algorithm was trained using these *training labels* on all time points where SpO_2 was not already ≤ 92% at that time point.

When evaluating the machine learning algorithm's performance by comparing with anaesthesiologists (Figure 3) we deliberately chose to use hypoxemia events encountered after a period of stable and normal SpO2 (Supplementary Figure 1). This was done to maximize the separation observed between different prediction approaches and so minimize the number of time points anaesthesiologists needed to label. For a more generalized prediction of all low SpO_2 values the performance reported on the full test set using training labels should be used (Supplementary Figure 3). The more stringent testing definition used for Figure 3 excludes some time points, leading to a smaller set of *anaesthesiologist testing labels.* Anaesthesiologist testing labels were positive only if SpO_2 was ≥ 95% for the past 10 minutes and then fell below 92% in the next five minutes (Supplementary Figure 1; left). Anaesthesiologist testing labels were negative only if SpO_2 remained ≥ 95% for the past ten minutes and the next ten minutes (Supplementary Figure 1; right). All the other cases do not have anaesthesiologist testing labels. This more restrictive labeling scheme ensures that positive testing labels are clear drops in SpO_2 levels that would be hard to predict in advance, while negative testing labels are clearly not drops in SpO_2 (Supplementary Figure 1).

An important point to consider when building labels for health outcome prediction is that anaesthesiologist interventions can affect outcomes. It has been noted that models can learn when a doctor is likely to intervene and hence lower the risk of an otherwise high-risk patient *(36)*.

This means that patients with low risk (from the model) may still need treatment. To address this, they proposed removing examples from the training set where doctors have intervened. This allows one to learn a model which predicts patient outcome without intervention. In our case, it is not possible to fully identify when or how an anaesthesiologist is intervening (and if that intervention prevented hypoxemia), so we sought to address this issue in two ways:

1. It must be recognized that the model predicts hypoxemia when following standard procedures, *not* the occurrence of hypoxemia if the anaesthesiologist takes no action to influence hypoxemia. This is a natural assumption in the operating room where interventions that may affect SpO_2 levels are performed frequently.
2. By focusing on clear explanations of why a certain risk was predicted we enable anaesthesiologists to identify when the algorithm may be basing its risk on their actions vs. when the risk is based on other factors.

Extracted time series features

To make a prediction at an arbitrary point in time, a consistent set of extracted features should be computed that capture the information present in all previous time points. All the data provided about a procedure is associated with a specific date and time. Text data has the time it was provided, minute-by-minute data from the patient monitor has the time at which each measurement was taken, and single point measurements have the times they were recorded.

We summarized these unevenly sampled time registered data into a fixed length feature vector at any point in time using several complementary methods:

- Patient data, procedure information, and pre-operative notes are represented by a “last value” extracted feature, which is zero before any data is recorded and the data’s value afterwards.
- Time series data are captured using exponentially decaying weighted average and variance estimates using multiple decay rates. These decay rates specify how much impact each past time point has on the computed mean or variance for the time series. We used 6 second, 1 minute, and 5-minute half-life times to capture both high and low frequency components of the signal in each time series (Supplementary Figure 12).
- Drug dose data are captured using both an exponentially decaying sum, and a time since the last measurement. Decay rates with half-lives of 5 minutes and 60 minutes were used to capture both near term and longer average drug dosing effects.

To ensure that there was enough training data for each extracted feature we removed extracted features that had less than 100 recorded data values for the real-time model, and less than 50 for the initial model. For a full list of the 3,797 extracted features used by Prescience for initial predictions see Supplementary Table 4 (STable4_InitialFeatures.csv). For the 3,905 extracted features used in intraoperative predictions see Supplementary Table

5 (STable5_RealtimeFeatures.csv). Note that over 2,000 of the initial and intraoperative features represent words from text data sources.

Gradient boosting machines for prediction

The extracted features we compute from real-time operating room data have a variety of complex nonlinear interactions. Capturing these requires a model with significant flexibility, and we chose a non-parametric approach called *gradient boosting machines (*26*)*.

We compared the performance of gradient boosting against three baseline methods: Lasso penalized linear logistic regression; a linear SVM autoregressive model previously proposed for predicting hypoxemia based only on the SpO_2 data stream *(*11*)*; and an unsupervised Parzen window method used previously to predict patient deterioration *(*22*)*. Gradient boosting machines significantly outperformed all baseline methods for our primary endpoint, real-time hypoxemia prediction (Supplementary Figure 3). For our secondary task of initial prediction gradient boosting machines were only slightly superior (Supplementary Figure 2). The large performance gain of gradient boosting for intraoperative prediction (Supplementary Figure 3) is likely because there are 8 million training samples, while for preoperative predictions (Supplementary Figure 2) there are only 42,000 samples and no time series data. Note that for initial prediction the autoregressive SVM and Parzen window methods were not applicable and hence not evaluated.

Gradient boosting machines are non-parametric models that draw a parallel between boosting and gradient descent in function space. They additively build up simpler models, like boosting, and these models are fit to the gradient of the loss at every data point. The most common type of basic model used is a regression tree because it is both robust to outliers and flexible. Taking some small fraction, η, of many trees fit to the gradient results in many small gradient descent steps in function space.

Fitting the trees is computationally challenging on large datasets so we used XGBoost, a high performance implementation of gradient boosting machines *(*27*)*. For the real-time model we used $\eta = 0.2$ and 1,242 trees, while for the initial model we chose $\eta = 0.1$ and 4,000 trees. Using a smaller η value means more trees are required for fitting, which requires more time to run, but results in a smoother (and generally better) model. For both initial and real-time models we used bagging, where trees were trained on a random 50% subsample of the training data. For the preoperative model the max tree depth was 4 and the minimum child weight of any branch in the trees was 1. For the real-time model the max tree depth was 6 and the minimum child weight of any branch in the trees was 10.

All method parameters were tuned (and methods were chosen) using a validation set of operating room procedures separate from the final test set used for all final performance results. To ensure that there was no bias towards the final test set, the test data were initially compressed and left compressed until after method development was completed.

Computing feature importance estimates

Understanding why a statistical model has made a specific prediction is a key challenge in machine learning. It engenders appropriate trust in predictions and provides insight into how

a model may be improved. However, many complex models with excellent accuracy, such as gradient boosting, make predictions even experts struggle to interpret. This forces a tradeoff between accuracy and interpretability. In response to this we chose to use a model agnostic representation of feature importance, where the impact of each feature on the model is represented using *Shapley values* (18, 37), which have been shown to be the only way to assign feature importance while maintaining two important properties *local accuracy*, and *consistency* (defined below) (20). The application of these values in Prescience uses fast estimation methods we have developed to compute the Shapley values (i.e., the estimated importance of features for a particular prediction) in a real-time manner (20, 21).

Shapley values are from the game theory literature and provide a theoretically justified method for allocation of a coalition's output among the members of the coalition (see Equation (1)). In Prescience the coalition is a set of interpretable model input feature values, and the coalition's output is the value of the prediction made by the model when given those input feature values. Feature impact is defined as the change in the expected value of the model's output when a feature is observed vs. unknown. Some feature values have a large impact on the prediction, while others have a small impact. The Shapley values $\phi_i(f,x)$, explaining a prediction $f(x)$, are an allocation of credit among the various features in x (such as age, weight, time series features, etc.), and are the only such allocation that obeys a set of desirable properties. Note that $\phi_i(f,x)$ is a single numerical value representing the impact of feature i, on the prediction of the model f when given the input x. For Prescience f is a gradient boosting model, and x is the set of all input features from a time point. We provide a brief summary of these properties below, and refer the reader to (20) for a full discussion, and for connections with several other recent methods in complex model interpretability. In the properties below $f_x(S) = E[f(x)|x_S]$, where x_S is a subset of the input vector with only the features in the set S present.

Local Accuracy.

$$f(x) = \phi_0(f,x) + \sum_{i=1}^{M} \phi_i(f,x)$$

where $\phi_0(f,x) = E[f(x)]$ the expected value of the model over the training data set), and M is the number of 'interpretable' inputs, which each correspond to a group of original input features (such as those shown in Figure 4). The local accuracy assumption forces the attribution values to correctly capture the difference between the expected model output and the output for the current prediction. For Prescience the input feature groups are the sets of extracted features associated with each time series. For instance, the 6 second, 1 minute, and 5 minute moving average extracted features, and the 5-minute moving variance extracted feature from the SpO_2 time series are all considered as a single group. This manual grouping process is not strictly necessary but can help improve the interpretation of partially redundant features.

Consistency. For any two models f and f', if

$$f'_x(S \cup \{i\}) - f'_x(S) \geq f_x(S) - f_x(S \cup \{i\}),$$

for all $S \in Z \backslash \{i\}$ where Z is the set of all M input features, then $\phi_i(f', x) \geq \phi_i(f, x)$. This states that if a feature is more important in one model than another, no matter what other features are also present, then the importance attributed to that feature should be also be higher.

Only one allocation of credit satisfies these two properties (and also trivial assumptions about unused model inputs) and that allocation is the one given by the Shapley values *(20)*.

Given a specific prediction $f(x)$ we can compute the Shapley values using a weighted sum that represents the impact of each feature being added to the model averaged over all possible orders of features being introduced:

$$\begin{aligned} \phi_i(f, x) &= \sum_{S \subseteq S_{all} \backslash \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\} - f_x(S) \right] \qquad (1) \\ &= \sum_{S \subseteq S_{all} \backslash \{i\}} \frac{1}{(M choose |S|)(M - |S|)} \left[f_x(S \cup \{i\}) - f_x(S) \right]. \end{aligned}$$

In practice, there are far too many terms to evaluate this sum completely, so we can instead approximate it by a sampling procedure *(18, 20)*. We have released an open implementation of this explanation approach which also includes additional upcoming improvements for tree models at: http://github.com/slundberg/shap

To compute the Shapley values of each prediction we need to estimate the predictions of the model when specific input features are missing (those not in the set S). Since the model was not trained to support missing values we approximate what the model would predict (if retrained on that subset of input features) by sampling from the training data set and replacing the missing features with the values they would have had in that sample. By averaging many such samples, we can estimate the expected value of $f_x(S)$ only using evaluations of $f_x(S_{all})$ where no features are missing.

The approach above requires nested sampling, once to estimate the Shapley value and then from each sample we again sample to estimate $f_x(S)$ and $f_x(S \cup \{i\})$. To reduce the number of samples in the inner step, we used k-medians to generate 20 medians of the entire dataset, and then performed a weighted evaluation for only these 20 summary inputs as an approximation for the entire dataset. This removes the need for nested sampling.

In Prescience we also used a non-linear link function h such that:

$$h(f(x)) = \sum_{i=0}^{M} \phi_i(f, x).$$

Since Prescience uses logistic regression the use of a $h = logit$ link function transforms the output space from probabilities to log odds. Assuming the importance of features is additive in the log-odds space is much more natural than assuming they are additive in the space of probabilities (which must fall between 0 and 1). The same reasoning also drives the use of the logit link function during standard logistic regression.

We were able to get stable feature importance estimates for thousands of features in less than 5 seconds on our server (in large part because these inputs typically had less than 100 non-zero entries). We compared these theoretically grounded explanations with a simple estimate of feature importance to verify they showed reasonable agreement. The simple method we chose was to replace a single feature group with random values from other samples in the data set and determine the average model output over different possible samplings. We then subtracted this mean value from the original model prediction to get a difference from a prediction with a typical value of that feature vs. the current value. This simple method is not very scalable and does not account for interactions with other features yet is useful to compare with the Prescience explanations to ensure the Prescience estimates of feature effects are consistent with an intuition of how much a feature's change from its typical value effects the current risk of hypoxemia (Supplementary Figure 4).

Physician evaluation

The potential benefit Prescience provides to physicians was evaluated using previously recorded procedures. Both before a procedure begins, and at several time points during the operation all the available electronically recorded data were shown to the anaesthesiologist and they were asked to predict if a desaturation (as defined above) will occur in the next 5 minutes (Supplementary Figures 5, 6, and 7). For half of the procedures anaesthesiologists are given Prescience explained risks (Supplementary Figures 5 and 6), and for the other half they are given the same data, but without any Prescience assistance (Supplementary Figure 7). In both cases anaesthesiologists are asked to provide a fold change in the risk that desaturation will occur.

The test procedures were divided into two equal sized groups, replicate 1 and replicate 2. Anaesthesiologists were also divided into two groups, A and B. Group A was given Prescience assistance on replicate 1 but not on replicate 2, while group B was given Prescience assistance on replicate 2 but not replicate 1. After randomly assigning anaesthesiologists to groups, three anaesthesiologists from group A completed the evaluation and two anaesthesiologists from group B. We pooled the results within each group and between groups, and the results of this evaluation are shown in Figure 3. The order in which anaesthesiologists were presented with cases was random across both replicate sets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Gabriel Erion, Marco Tulio Ribeiro, Jacob Schreiber, and members of the Lee lab for feedback and suggestions that improved the manuscript and experiments. This work was supported by a National Science

Foundation (NSF) DBI-135589 and DBI-1552309, National Institutes of Health (NIH) 1R35GM128638, NSF Graduate Research Fellowship DGE-1256082, and a UW eScience/ITHS seed grant *Machine Learning in Operating Rooms*.

References:

1. Weiser TG, Haynes AB, Molina G, Lipsitz SR, Esquivel MM, Uribe-Leitz T, Fu R, Azad T, Chao TE, Berry WR, Gawande AA, Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes, Lancet 385, S11 (2015).
2. Gawande AA, Thomas EJ, Zinner MJ, Brennan TA, The incidence and nature of surgical adverse events in Colorado and Utah in 1992, Surgery 126, 66–75 (1999). [PubMed: 10418594]
3. Kable AK, Gibberd RW, Spigelman AD, Adverse events in surgical patients in Australia., Int. J. Qual. Heal. care J. Int. Soc. Qual. Heal. Care 14, 269–76 (2002).
4. Nair BG, Gabel E, Hofer I, Schwid HA, Cannesson M, Intraoperative Clinical Decision Support for Anesthesia, Anesth. Analg 124, 603–617 (2017). [PubMed: 28099325]
5. G. D. H. & P. J. Maier-Hein Lena, Swaroop S Stefanie Speidel Vedula, Navab Nassir, Kikinis Ron, Park Adrian, Eisenmann Matthias, Feussner Hubertus, Forestier Germain, Giannarou Stamatia, Hashizume Makoto, Katic Darko, Kenngott Hannes, Kranzfelder Michael, Malpani Anand, Surgical data science for next-generation interventions, Nat. Biomed. Eng 1, 691–696 (2017).
6. Dunham CM, Hileman BM, Hutchinson AE, Chance EA, Huang GS, Perioperative hypoxemia is common with horizontal positioning during general anesthesia and is associated with major adverse outcomes: a retrospective study of consecutive patients, *BMC Anesthesiol.* 14, 43 (2014). [PubMed: 24940115]
7. Strachan L, Noble DW, Hypoxia and surgical patients--prevention and treatment of an unnecessary cause of morbidity and mortality., *J. R. Coll. Surg. Edinb.* 46, 297–302 (2001). [PubMed: 11697699]
8. Ehrenfeld JM, Funk LM, Van Schalkwyk J, Merry AF, Sandberg WS, Gawande A, The incidence of hypoxemia during surgery: evidence from two institutions, Can. J. Anesth. Can. d'anesthésie 57, 888–897 (2010).
9. Kooij FO, Klok T, Hollmann MW, Kal JE, Decision support increases guideline adherence for prescribing postoperative nausea and vomiting prophylaxis, *Anesth. Analg.* 106, 893–898 (2008). [PubMed: 18292437]
10. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB, Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review., JAMA 293, 1223–38 (2005). [PubMed: 15755945]
11. ElMoaqet H, Tilbury DM, Ramachandran SK, Multi-Step Ahead Predictions for Critical Levels in Physiological Time Series, IEEE Trans. Cybern 46, 1704–1714 (2016). [PubMed: 27244754]
12. Lipton ZC, Kale DC, Wetzell RC, Phenotyping of Clinical Time Series with LSTM Recurrent Neural Networks, (2015) (available at http://arxiv.org/abs/1510.07641).
13. Henry KE, Hager DN, Pronovost PJ, Saria S, A targeted real-time early warning score (TREWScore) for septic shock, Sci. Transl. Med 7 (2015).
14. Saria S, Rajani AK, Gould J, Koller D, Penn AA, Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants, Sci. Transl. Med 2, 48ra65–48ra65 (2010).
15. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N, Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission, Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '15, 1721–1730 (2015).
16. Deo RC, Machine Learning in Medicine, Circulation 132 (2015).
17. Memarian N, Kim S, Dewar S, Engel J, Staba RJ, Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy, Comput. Biol. Med 64, 67–78 (2015). [PubMed: 26149291]
18. Štrumbelj E, Kononenko I, Explaining prediction models and individual predictions with feature contributions, Knowl. Inf. Syst 41, 647–665 (2014).
19. Ribeiro MT, Singh S, Guestrin C, Why Should I Trust You? Explaining the Predictions of Any Classifier, doi:10.1145/2939672.2939778.

20. Lundberg S, Lee S-I, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst, 1–10 (2017).
21. Lundberg SM, Erion GG, Lee S-I, Consistent Individualized Feature Attribution for Tree Ensembles, (2018) (available at http://arxiv.org/abs/1802.03888).
22. Tarassenko L, Hann A, Young D, Integrated monitoring and analysis for early warning of patient deterioration, Br. J. Anaesth 97, 64–68 (2006). [PubMed: 16707529]
23. Summers RL, Pipke M, Wegerich S, Conkright G, Isom KC, Functionality of empirical model-based predictive analytics for the early detection of hemodynamic instabilty., Biomed. Sci. Instrum 50, 219–224 (2014). [PubMed: 25405427]
24. American Medical Association, Current procedural terminology: CPT (2007).
25. National Center for Health Statistics, Health, United States, 2016: With Chartbook on Long-term Trends in Health, Cent. Dis. Control, 314–317 (2017).
26. Friedman JH, Greedy function approximation: A gradient boosting machine., *Ann. Stat.* 29, 1189–1232 (2001).
27. Chen T, Guestrin C, XGBoost: A Scalable Tree Boosting System, doi:10.1145/2939672.2939785.
28. Lumachi F, Marzano B, Fanti G, Basso SMM, Mazza F, Chiara GB, Relationship between body mass index, age and hypoxemia in patients with extremely severe obesity undergoing bariatric surgery., In Vivo 24, 775–7. [PubMed: 20952748]
29. Kendale SM, Blitz JD, Increasing body mass index and the incidence of intraoperative hypoxemia, J. Clin. Anesth 33, 97–104 (2016). [PubMed: 27555141]
30. Defining Adult Overweight and Obesity | Overweight & Obesity | CDC (available at https://www.cdc.gov/obesity/adult/defining.html).
31. Myles PS, Leslie K, McNeil J, Forbes A, Chan MTV, Bispectral index monitoring to prevent awareness during anaesthesia: The B-Aware randomised controlled trial, Lancet 363, 1757–1763 (2004). [PubMed: 15172773]
32. Avidan MS, Zhang L, Burnside BA, Finkel KJ, Searleman AC, Selvidge JA, Saager L, Turner MS, Rao S, Bottros M, Hantler C, Jacobsohn E, Evers AS, Anesthesia Awareness and the Bispectral Index, N. Engl. J. Med 358, 1097–1108 (2008). [PubMed: 18337600]
33. Epstein RH, Dexter F, Patel N, Influencing Anesthesia Provider Behavior Using Anesthesia Information Management System Data for Near Real-Time Alerts and Post Hoc ReportsAnesth. Analg 121, 678–692 (2015).
34. Guay J, Ochroch EA, in Cochrane Database of Systematic Reviews, Guay J, Ed. (John Wiley & Sons, Ltd, Chichester, UK, 2015), p. CD011151.
35. World Health Organization, Pulse Oximetry Training Manual (2011).
36. Dyagilev K, Saria S, Learning (predictive) risk scores in the presence of censoring due to interventions, Mach. Learn 102, 323–348 (2016).
37. Roth AE, The Shapley Value - Cambridge University PressCambridge Univ. Press (1988) (available at http://www.cambridge.org/catalogue/catalogue.asp?isbn=0511829728).

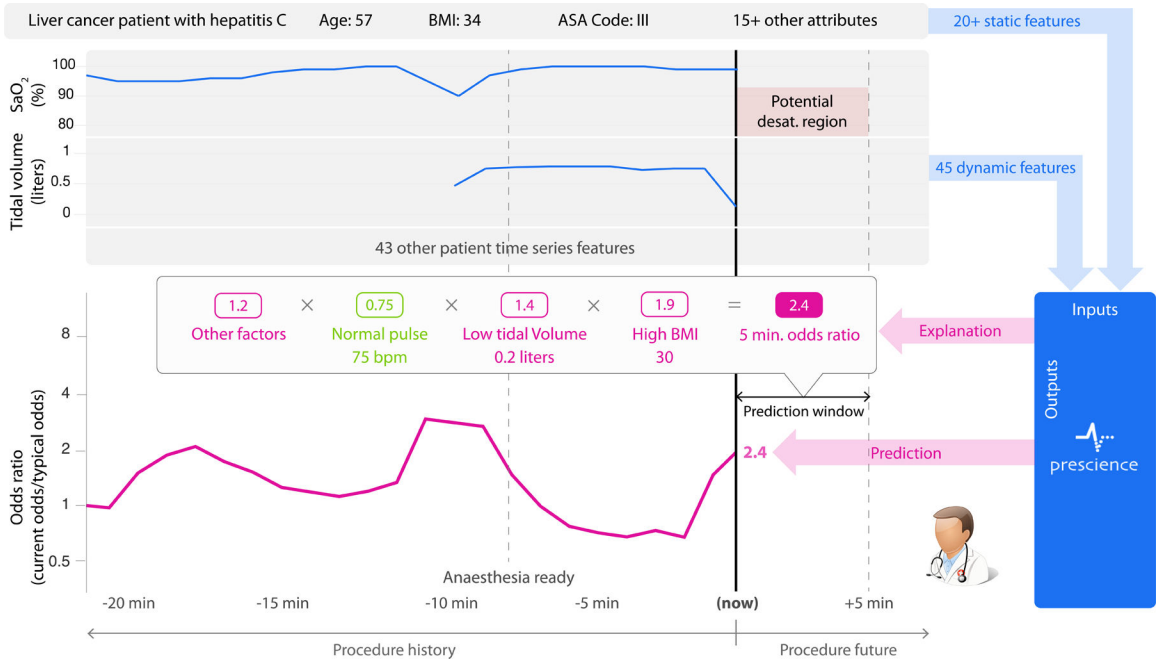

Fig 1. Prescience integrates many data sources into a single risk, which is explained through a succinct visual summary.

A wide variety of data sources were used to build a predictive model of hypoxemia events. An explanation (shown above) is then built for each prediction. Purple features have values that increased risk, while green features decreased hypoxemia risk. The combination of impacts of all features is the predicted Prescience risk; in this case the odds are 2.4 times higher than normal. Each feature impact value represents the change in risk when that feature's value is known vs. unknown. Qualitative terms such as "low" or "high" are based on a feature value's distribution in our dataset.

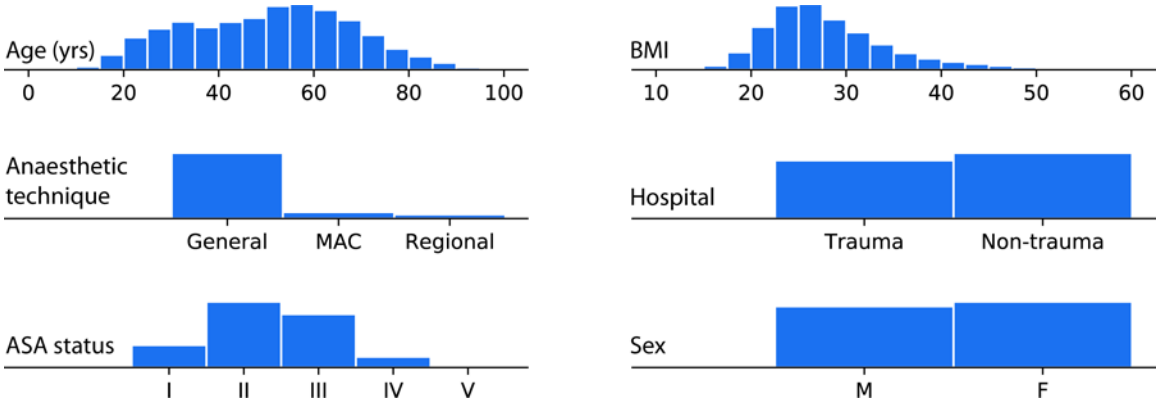

Fig 2. Patient and procedure characteristics.

Histograms summarizing basic properties of the anaesthesia procedures used for training. Prescience was trained and evaluated using data from 53,126 procedures recorded at two hospitals over two years (representing unique patients). MAC = monitored anaesthesia care; M = Male and F = Female, ASA = American Society of Anesthesiologists; BMI = Body Mass Index. In our dataset 37.4% of adults aged 20 or over have a BMI of 30 or more, which is a close match to the U.S. obesity rate of 37.9% *(25)*.

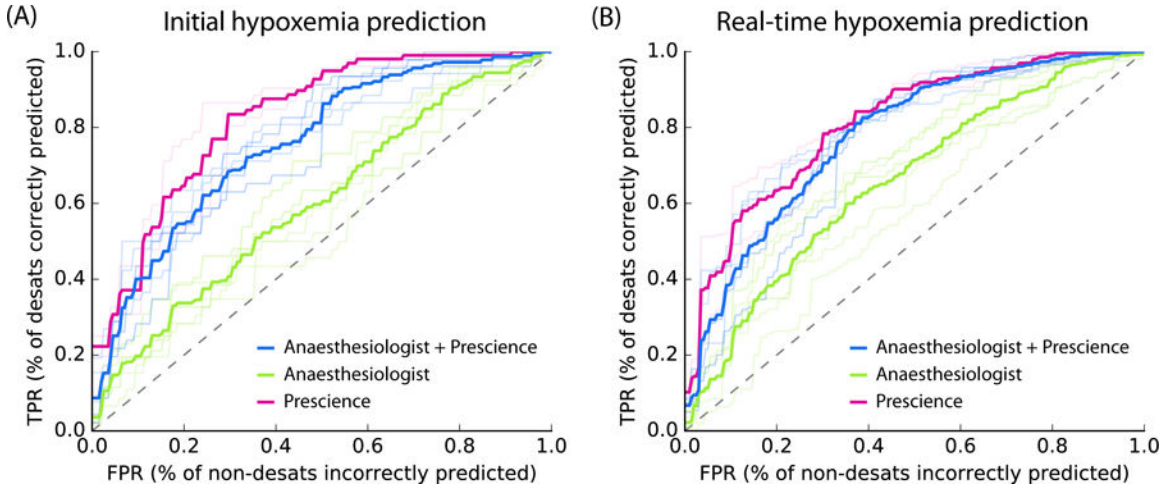

Fig 3. Pooled comparison of five anaesthesiologists' prediction performance with and without assistance by Prescience.

Receiver Operating Characteristic (ROC) plots comparing five anaesthesiologists' predictions from recorded data with and without Prescience assistance. Light colored lines represent individual anaesthesiologist's performances; dark lines represent their average performance. (**A**) For initial risk prediction, anaesthesiologists (green, AUC = 0.60) performed significantly better with Prescience assistance (blue, AUC = 0.76; P-value < 0.0001) than without Prescience assistance, and Prescience performed better in a direct comparison with anaesthesiologists (purple, AUC = 0.83; P-value < 0.0001). (**B**) For intraoperative real-time (next 5 minute) risk prediction anaesthesiologists (green, AUC = 0.66) again performed better with Prescience assistance (blue, AUC = 0.78; P-value < 0.0001), and Prescience alone outperformed anaesthesiologists predictions (purple, AUC = 0.81; P-value < 0.0001). Note that the False Positive Rate (FPR) (x-axis) measures how many points without upcoming hypoxemia were incorrectly predicted to have upcoming hypoxemia. The True Positive Rate (TPR) (y-axis) measures what fraction of hypoxemic events were correctly predicted. P-values were computed using bootstrap resampling over the tested time points while measuring the difference in area between the curves. If we instead resample over anaesthesiologists we observe bootstrap P-values of 0, and t-test P-values < 0.001 for Prescience improvements. See Supplementary Figure 8 for plots of the statistical separation between the mean ROC curves across all false positive rates.

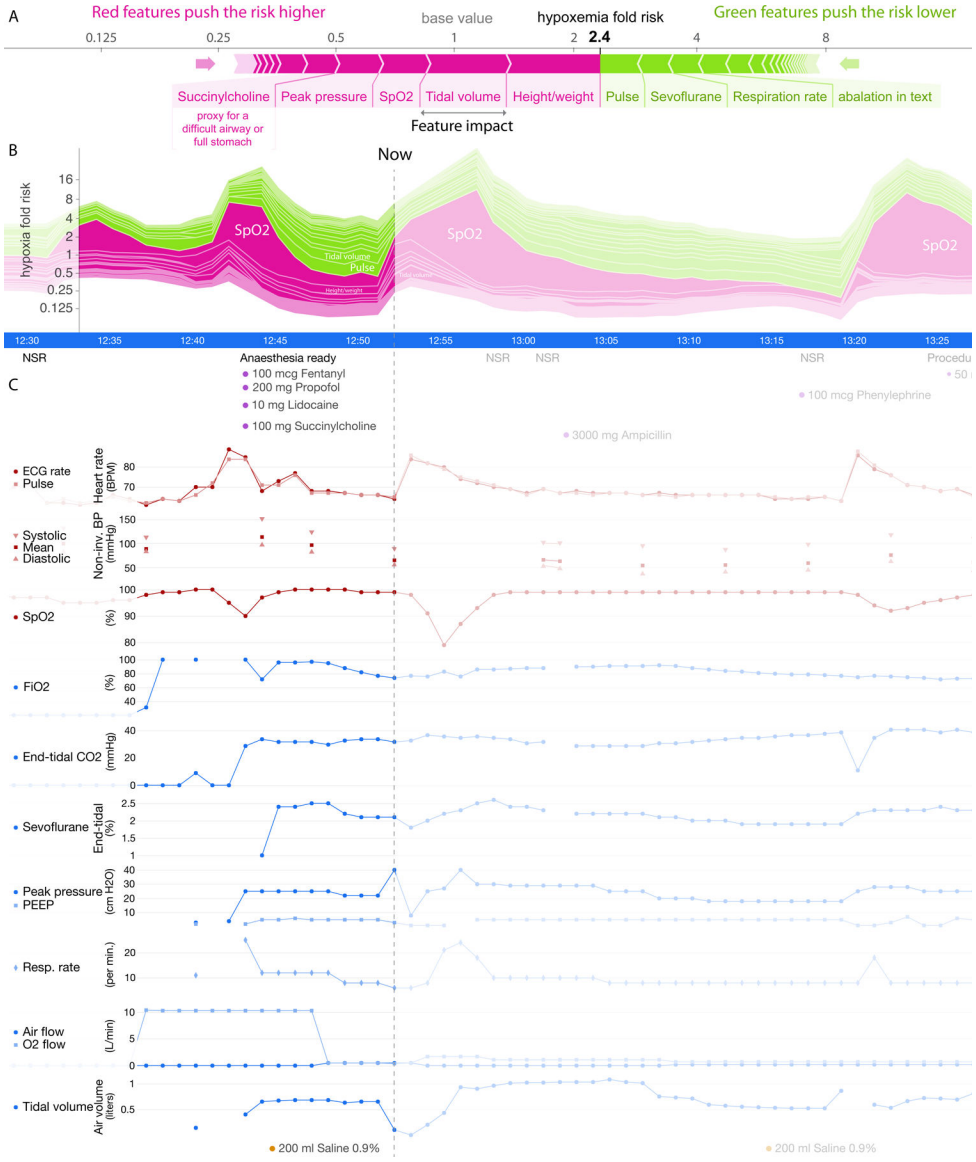

Fig 4. Sample real-time prediction during a procedure.

One hour of data is shown from a procedure. (**A**) Explained risk of hypoxemia in the next five minutes. (**B**) Plot of the explained risks evolving over time. This plot is equivalent to rotating (A) 90 degrees and stacking the risk explanations for every time point horizontally. (**C**) A subset of the patient data for this procedure, plotted both before and after the current time point.

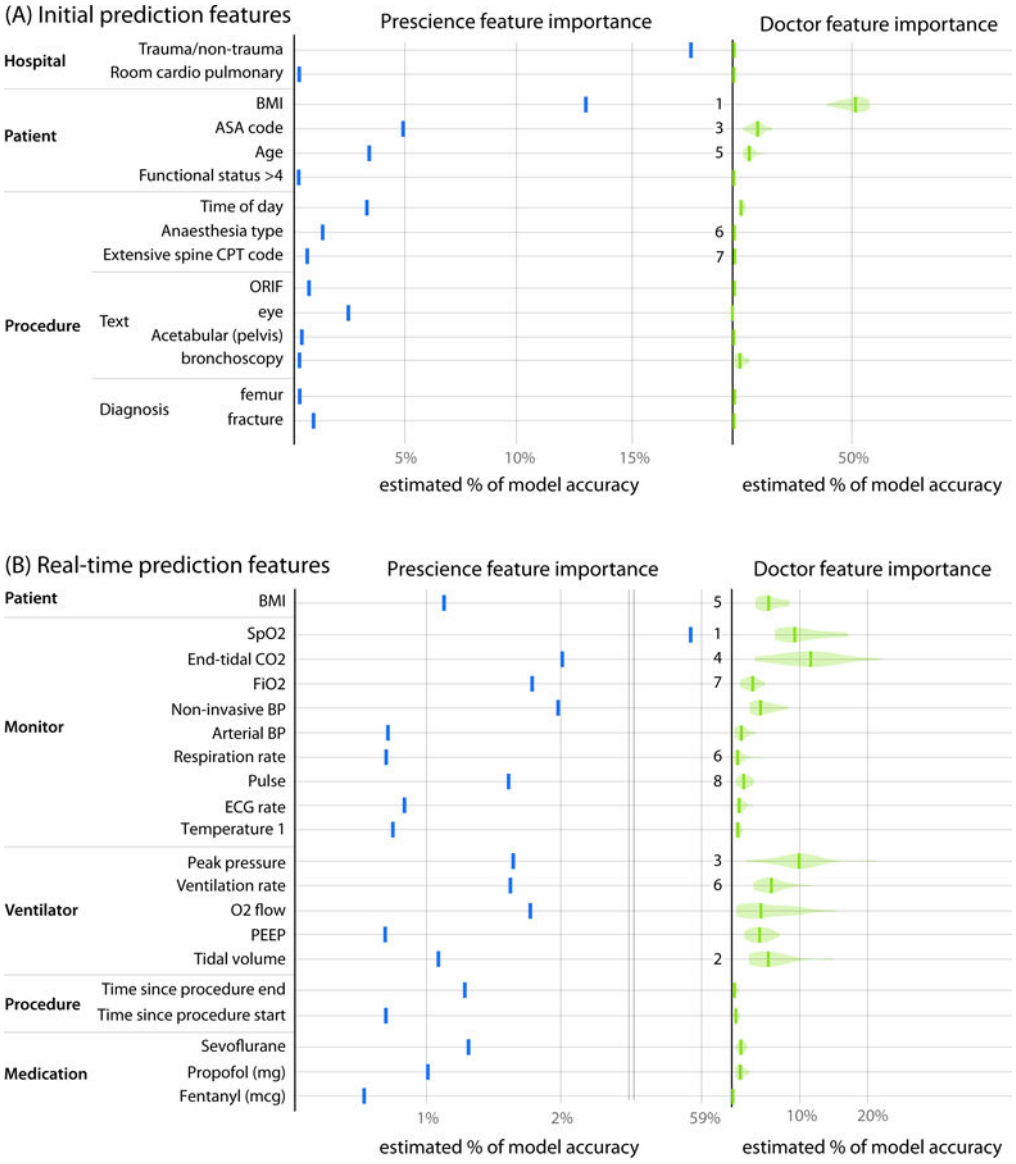

Fig 5. Comparison of averaged feature importance estimates between Prescience and anaesthesiologists for both initial and real-time prediction.

Importance estimates assigned by the Prescience model (blue) and anaesthesiologists (green) to the top features in both (**A**) initial and (**B**) real-time prediction. The importance of features is measured as the estimated percent of the model's prediction accuracy that is due to that feature. The numbers presented to the left of the imputed anaesthesiologist importance estimates are feature rankings from a consensus of anaesthesiologist responses about which features they believed would be important (Supplementary Tables 2 and 3). Note that the 2nd (asthma) and 4th (lung disease) features as ranked by anaesthesiologists for initial prediction were not in the top Prescience features. The 6th ranked feature by anaesthesiologists for real-time prediction corresponds to two Prescience data sources. The quantitative anaesthesiologist feature importance estimates were estimated using 20 bootstrapped models trained to mimic the anaesthesiologist's predictions when unassisted by Prescience.

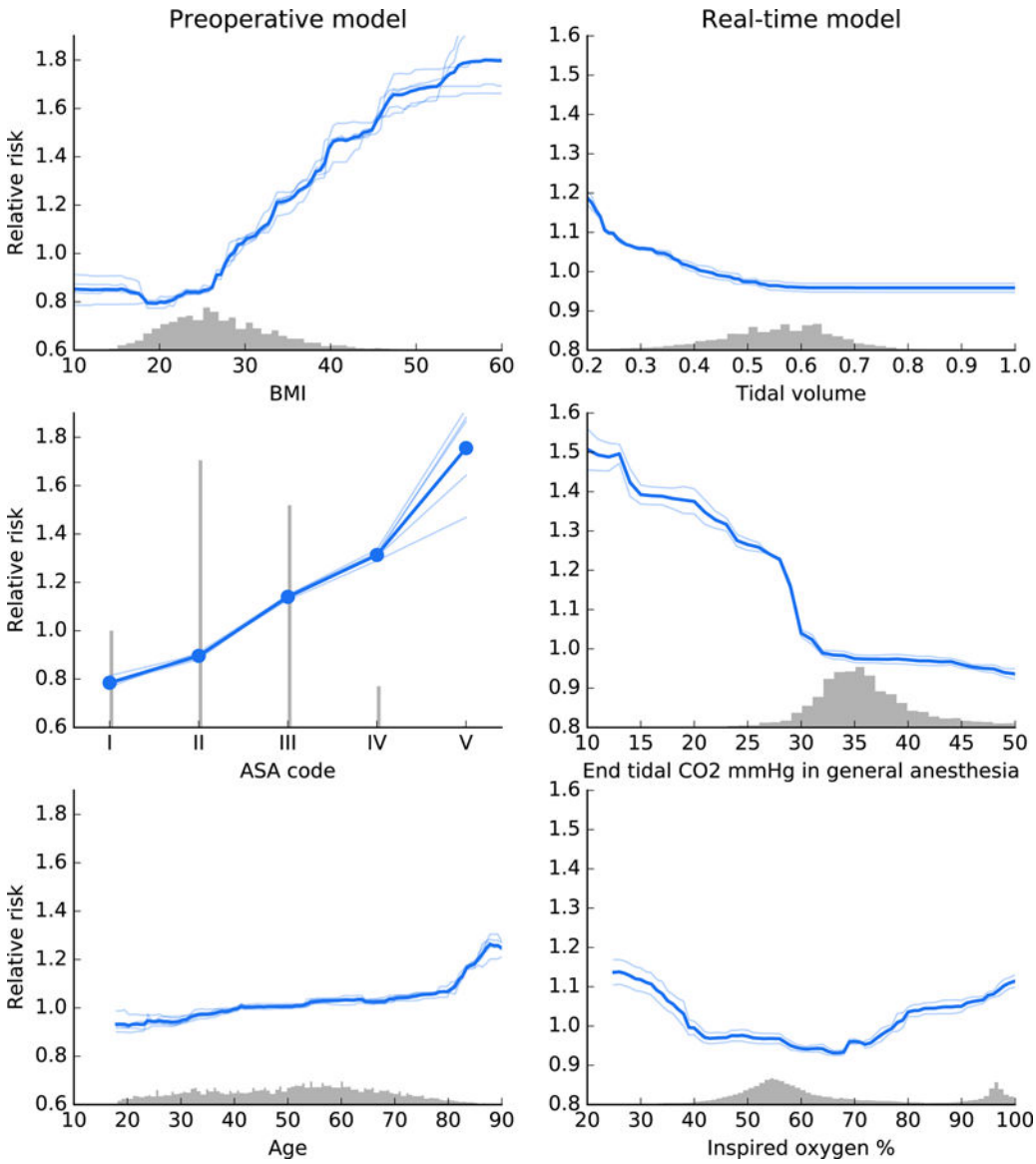

Fig 6. Effect of varying individual feature values for both initial features (left) and real-time features (right).

These partial dependence plots show the change in hypoxemia risk for all values of a given feature. The gray histograms on each plot show the distribution of values for that feature in the validation dataset. Light colored lines represent model variability from bootstrap resampling of the training data.