nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-018-0304-0

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

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

Although anaesthesiologists strive to avoid hypoxaemia during surgery, reliably predicting future intraoperative hypoxaemia is not possible at present. Here, we report the development and testing of a machine-learning-based system that predicts the risk of hypoxaemia and provides explanations of the risk factors in real time during general anaesthesia. The system, which was trained on minute-by-minute data from the electronic medical records of over 50,000 surgeries, improved the performance of anaesthesiologists by providing interpretable hypoxaemia 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 hypoxaemia events, with the assistance of this system they could anticipate 30%, 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 hypoxaemia risk during anaesthesia care by providing general insights into the exact changes in risk induced by certain characteristics of the patient or procedure.

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 0.4–0.8% and a complication rate of 3–17%, just in industrialized countries[2,3]. Fortunately, half of these complications are preventable[2,3]. With an increase in the adoption of electronic medical record systems, high-fidelity heterogeneous data are being captured during surgery and anaesthesia care, but these data are rarely used to improve patient safety and quality of care[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.

Hypoxaemia, or low arterial blood oxygen tension, is an unwanted physiological condition known to cause serious patient harm during general anaesthesia and surgery[6]. Hypoxaemia is associated with cardiac arrest, cardiac arrhythmias, post-operative infections and wound healing impairments, decreased cognitive function and delirium, and cerebral ischaemia 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, hypoxaemia 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 their 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 ref. [4] for a full review. If hypoxaemia can be predicted or anticipated before it occurs, then actions can be taken by anaesthesiologists to proactively prevent hypoxaemia and minimize patient harm.

Machine-learning 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 machine-learning techniques to predict adverse events such as hypoxaemia in a considerably more complex setting such as the operating room are lacking at present. Moreover, although previous complex machine-learning 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, that is, the 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 hypoxaemia during anaesthesia care and explains the patient- and surgery-specific factors that led to that risk (Fig. 1). We believe this is an important step forwards for machine learning in medicine because although machine-learning models have significantly improved the ability to predict the future condition of a patient[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 (for example, linear) models at the expense of accuracy. To address this problem, some approaches have achieved interpretability by carefully limiting the complexity of the

[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]Center for Perioperative and Pain initiatives in Quality Safety Outcome, University of Washington, Seattle, WA, USA. [4]Harborview Injury Prevention and Research Center, University of Washington, Seattle, WA, USA. [5]Veterans Affairs Puget Sound Health Care System, Seattle, WA, USA. [6]Seattle Children's Hospital, Seattle, WA, USA. *e-mail: suinlee@cs.washington.edu

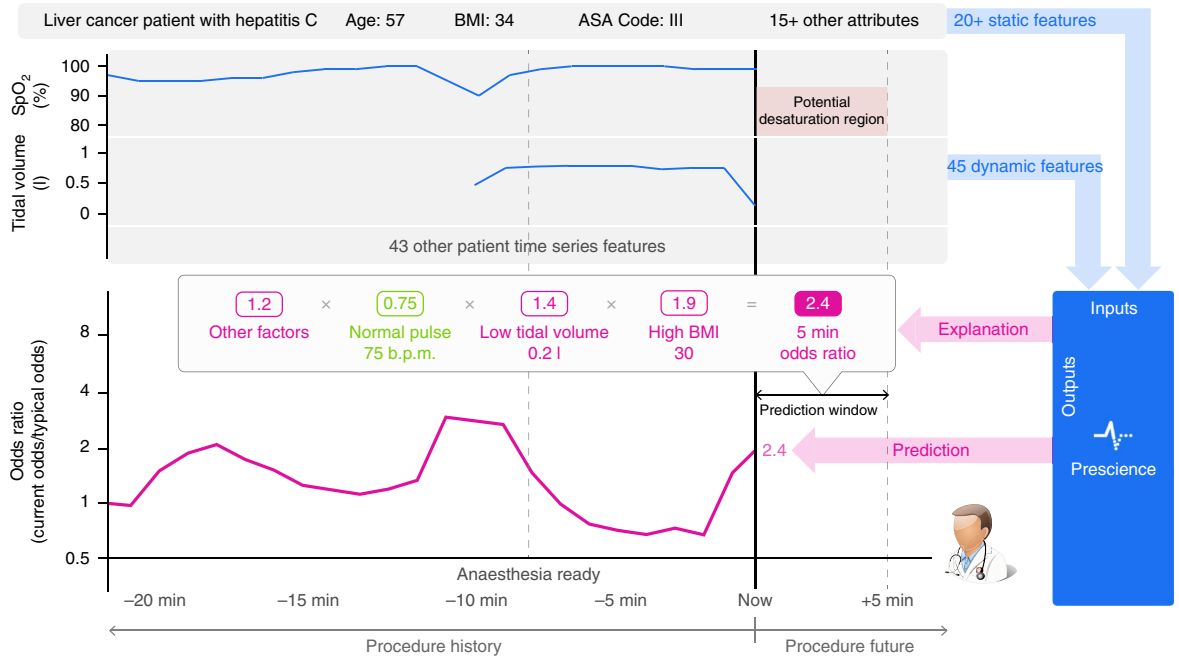

Fig. 1 | Prescience integrates many data sources into a single risk, which is explained through a succinct visual summary. A wide variety of data sources were used to build a predictive model of hypoxaemia events. An explanation (overlaid) is then built for each prediction. Pink features have values that increased risk, whereas green features decreased hypoxaemia risk. The combination of the impacts of all features is the predicted Prescience risk; in this case, the odds are 2.4× higher than normal. Each feature impact value represents the change in risk when the value of that feature is known versus unknown. Qualitative terms such as 'low' or 'high' are based on the distribution of a feature value in our dataset.

machine-learning model[15]. In contrast, we demonstrate how to retain interpretability, even when complex models such as non-parametric methods or deep learning are used, by developing a method to provide theoretically justified explanations of model predictions that builds 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 versus interpretability trade-off, 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 time 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). Whereas some operating room forecasting approaches have relied on simulated physiology[23], the AIMS data consist 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 in 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 (for example, the 2.4 odds ratio in Fig. 1). However, a more detailed presentation that shows the risk is due to the body mass index (BMI), current tidal volume and pulse rate of a patient is more clinically meaningful as some factors may be modifiable and result in clinical changes mitigating that risk (Fig. 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 predicted risk of the model when we observe a feature (such as the weight of a patient) versus when we do not observe the feature (such as not knowing the weight of a patient). This change in the output prediction of a model 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 hypoxaemia 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 the explained predictions made by Prescience and gain insight into factors that affect intraoperative hypoxaemia, we present the following results: (1) a comparison of Prescience hypoxaemia predictions against anaesthesiologists' predictions with and without the aid of Prescience; (2) an example of how Prescience explains hypoxaemia risk at a specific time-point during a surgical procedure; (3) a comparative summary of relevant AIMS data features for hypoxaemia prediction chosen by Prescience

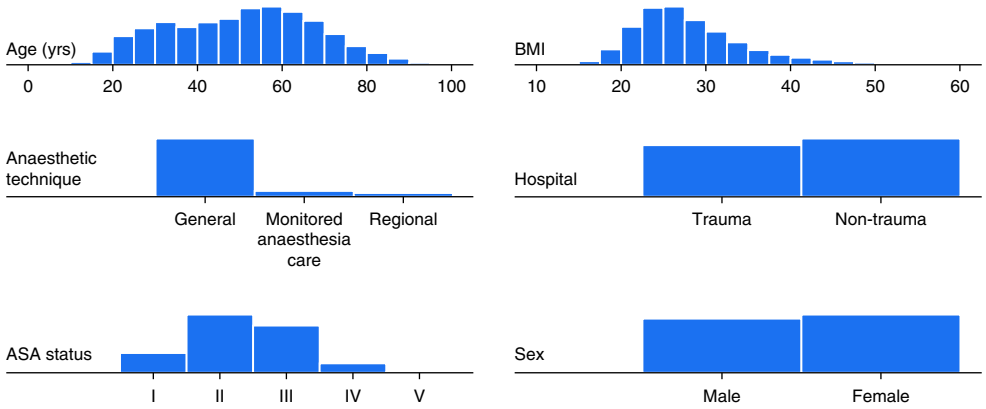

Fig. 2 | Patient and procedure characteristics. Histograms summarizing the basic properties of the anaesthesia procedures used for training (*y* axes represent normalized procedure counts). Prescience was trained and evaluated using data from 53,126 procedures recorded at two hospitals over two years (representing 36,232 unique patients). 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 US obesity rate of 37.9%[37]

and by anaesthesiologists and (4) a detailed presentation of key risk factors for hypoxaemia identified by Prescience.

Prescience overview. Based on the World Health Organization recommendations and for the purposes of prediction, we defined hypoxaemia as the decrease in SpO_2, that is, arterial blood oxygen saturation as measured by pulse oximetry, to a threshold value of 92% or lower (see Methods; Supplementary Fig. 1). From the AIMS data, we extracted 3,797 static extracted features for each patient from more than 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 more than 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 Fig. 1, whereas sets of features from a single time series (such as tidal volume) are combined in Fig. 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, for example, during cardiopulmonary bypass. All the experiments were performed after appropriate Institutional Review Board approval (see Methods), with clinical data summarized in Fig. 2.

We trained a gradient boosting machine model[25] to solve the following two types of prediction problems. The first is the initial prediction, at the start of a procedure the risk of hypoxaemia at any time during a procedure is predicted on the basis of the static extracted features. The second is the real-time prediction, predicting hypoxaemia in the next 5 min 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 min window as it would be long enough to allow an anaesthesiologist enough time to intervene but also short enough so that it represents near-term risks that would benefit from immediate attention. For initial prediction 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 Fig. 2). For real-time prediction we used 8,087,476 per-minute time points as training samples, 1,053,629 as validation samples and 963,674 as test samples, for which all time points from the same procedure were included in the same sample set and no missing data imputation was performed (Supplementary Fig. 3). Dividing the time points by procedure is important as samples from the same procedure are not independently and identically distributed but have some time dependency. 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 Figs. 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.

We used 198 and 523 test samples to evaluate the performance of anaesthesiologists for initial and real-time prediction tasks, respectively (see below; Fig. 3). Prescience outputs the risk prediction and its explanations (Figs. 1 and 4a), which show a set of features that increased (pink) and decreased (green) the risk.

We developed an efficient, theoretically justified machine-learning technique based on recent advances for interpreting models to estimate the importance of each feature in a prediction made for a single patient. This drives real-time explanations (Fig. 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 Fig. 4). We also developed effective visualizations of these explanations that encode them in a compact visual form for anaesthesiologists (Fig. 1; Supplementary Figs. 5–7) and a more detailed visualization that highlights the relevant contributing features (Fig. 4; see Methods for details).

Prescience improves the ability of an anaesthesiologist to predict hypoxaemia. To test the potential of Prescience to aid hypoxaemia prediction we replayed pre-recorded intraoperative data from test sample procedures in a web-based visualization to five practising anaesthesiologists (Supplementary Figs. 5–7). Each anaesthesiologist was given both types of prediction tasks, initial prediction (198) and real-time prediction (523). For each prediction task, anaesthesiologists were asked to provide a relative risk of hypoxaemia compared to a normal acceptable risk, for example, 0.01 for one-hundredth of 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 Fig. 3, which plots the true positive rate (that is, the percentage of correctly predicted desaturations) on the *y* axis against the false positive rate (that is, the percentage of incorrectly predicted non-desaturations) on the *x* axis. Note that ROC curves depend only on the order of the relative risk values among

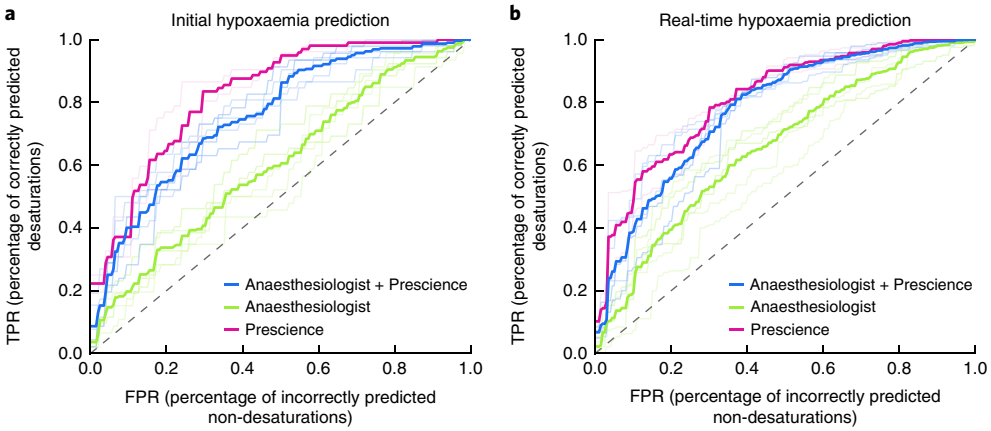

Fig. 3 | Pooled comparison of the prediction performance of five anaesthesiologists with and without the assistance of Prescience. Receiver operating characteristic plots comparing the predictions made by five anaesthesiologists from recorded data, results with and without Prescience assistance are shown. The lighter coloured lines represent the performances of the individual anaesthesiologists; the darker lines represent their average performance. **a**, For initial risk prediction, anaesthesiologists performed significantly better with Prescience (area under curve (AUC) = 0.76; $P < 0.0001$)) than without Prescience (AUC = 0.60) and Prescience performed better in a direct comparison with anaesthesiologists (AUC = 0.83; $P < 0.0001$). **b**, For intraoperative real-time (next 5 min) risk prediction, anaesthesiologists (AUC = 0.66) again performed better with Prescience (AUC = 0.78; $P < 0.0001$) and Prescience alone outperformed anaesthesiologists predictions (AUC = 0.81; $P < 0.0001$). Note that the false positive rate (FPR; *x* axis) measures how many points without upcoming hypoxaemia were incorrectly predicted to have upcoming hypoxaemia. The true positive rate (TPR; *y* axis) measures the fraction of hypoxaemic events that 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 Student's *t*-test $P < 0.001$ for Prescience improvements. See Supplementary Fig. 8 for plots of the statistical separation between the mean ROC curves across all false positive rates.

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, the predictions made by Prescience (pink) are considerably more accurate than the predictions made by anaesthesiologists (green). The prediction accuracy of anaesthesiologists markedly improved when the anaesthesiologists were given Prescience's risk prediction and its explanations in addition to the original procedure data (Supplementary Figs. 5–7). A clear separation between the performance of anaesthesiologists with and without the aid of Prescience is observed for both initial prediction (Fig. 3a, $P < 0.0001$) and real-time prediction (Fig. 3b, $P < 0.0001$). This suggests that Prescience can enhance the assessment of future risk made by anaesthesiologists and their ability to proactively anticipate hypoxaemia events. Interestingly, the prediction performance of anaesthesiologists with Prescience explanations was slightly lower than direct predictions from Prescience. 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 twice on the same prediction task—one with and the other without Prescience, we created replicate test sets by dividing the prediction tasks into two groups of similar size: tasks for initial prediction (100 and 98 tasks) and tasks for real-time prediction (260 and 263 tasks). Each of the five recruited anaesthesiologists was assigned to receive Prescience's assistance in one of these two replicate test sets (see Methods). The procedures shown to anaesthesiologists were chosen such that about 50% showed at least one incident of hypoxaemia (for preoperative prediction) and time points were chosen such that about 33% had hypoxaemia in the next 5 min (for intraoperative prediction). The anaesthesiologist test time points for hypoxaemia were chosen to be drops in SpO_2 (Supplementary Fig. 1) with a preceding period of stable and normal SpO_2. However, the entire dataset also included easier-to-predict hypoxemic events that follow previous SpO_2 drops and decreasing SpO_2 trends. For the entire dataset, Prescience achieves a higher real-time prediction area under the ROC curve of 0.90 (as opposed to 0.81 in Fig. 3b; see Supplementary Fig. 3), and when a larger training dataset was used, the area under the real-time prediction ROC curve rose to 0.92 (Supplementary Fig. 9).

If we extrapolate the real-time results to the 30 million annual surgeries in the United States under the assumption that doctors anticipate 15% of hypoxemic events while SpO_2 is still ≥95%, then with the assistance of Prescience they may be able to anticipate 30% of these events, or approximately 2.4 million additional episodes of hypoxaemia annually. Given that 20% of the risk predicted by Prescience is based on drugs and settings under the control of the anaesthesiologist (Supplementary Table 5), 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 yr) and were all actively practising 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 the false alarm rates. This can be accomplished by adjusting the trade-off between precision (the positive predictive value) and recall (the sensitivity). High precision means a low false alarm rate (which is 1 – precision), however, it comes at the cost of low recall. Supplementary Figure 10 plots the precision and recall trade-off for Prescience on the full set of test time points. Given that the performance of the complex model in Prescience improves with larger datasets, we also included results from a model trained on

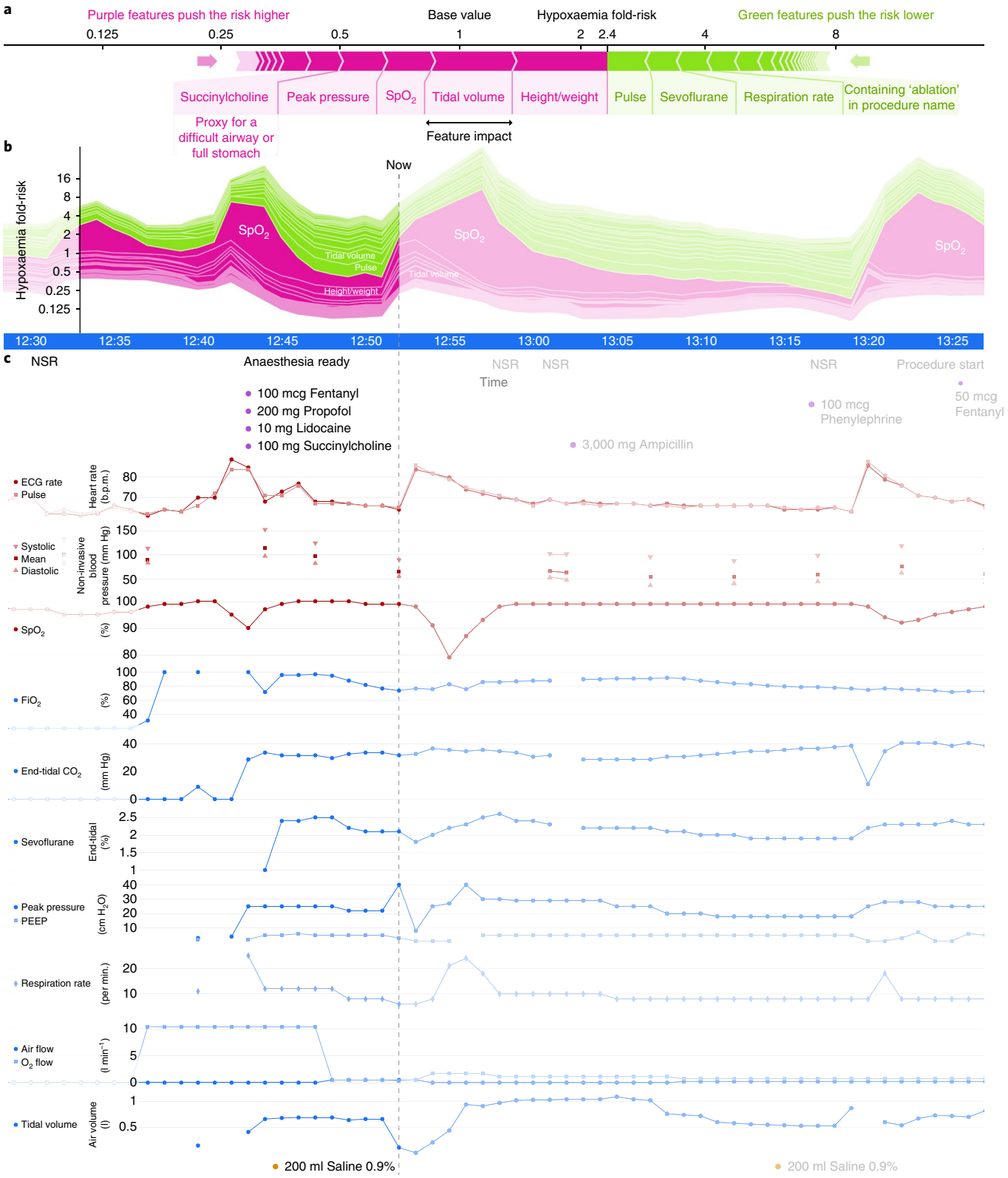

Fig. 4 | Sample real-time prediction during a procedure. An hour of data from a procedure is shown. **a**, Explained risk of hypoxaemia in the next five minutes. **b**, Plot of the explained risks evolving over time. This plot is equivalent to rotating **a** by 90° and stacking the risk explanations for every time point horizontally. **c**, Subset of the patient data for this procedure, plotted both before and after the current time point. Procedure events and drug doses are shown at the top above the time-series data, static patient data are not shown for the sake of patient privacy. NSR, normal sinus rhythm; PEEP, positive end-expiratory pressure.

an expanded dataset of 175,000 procedures 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 hypoxaemia at 70% precision (or 44% of all minutes with upcoming hypoxaemia at a precision of 30%, if the threshold for precision-recall trade-off is selected for higher recall). These

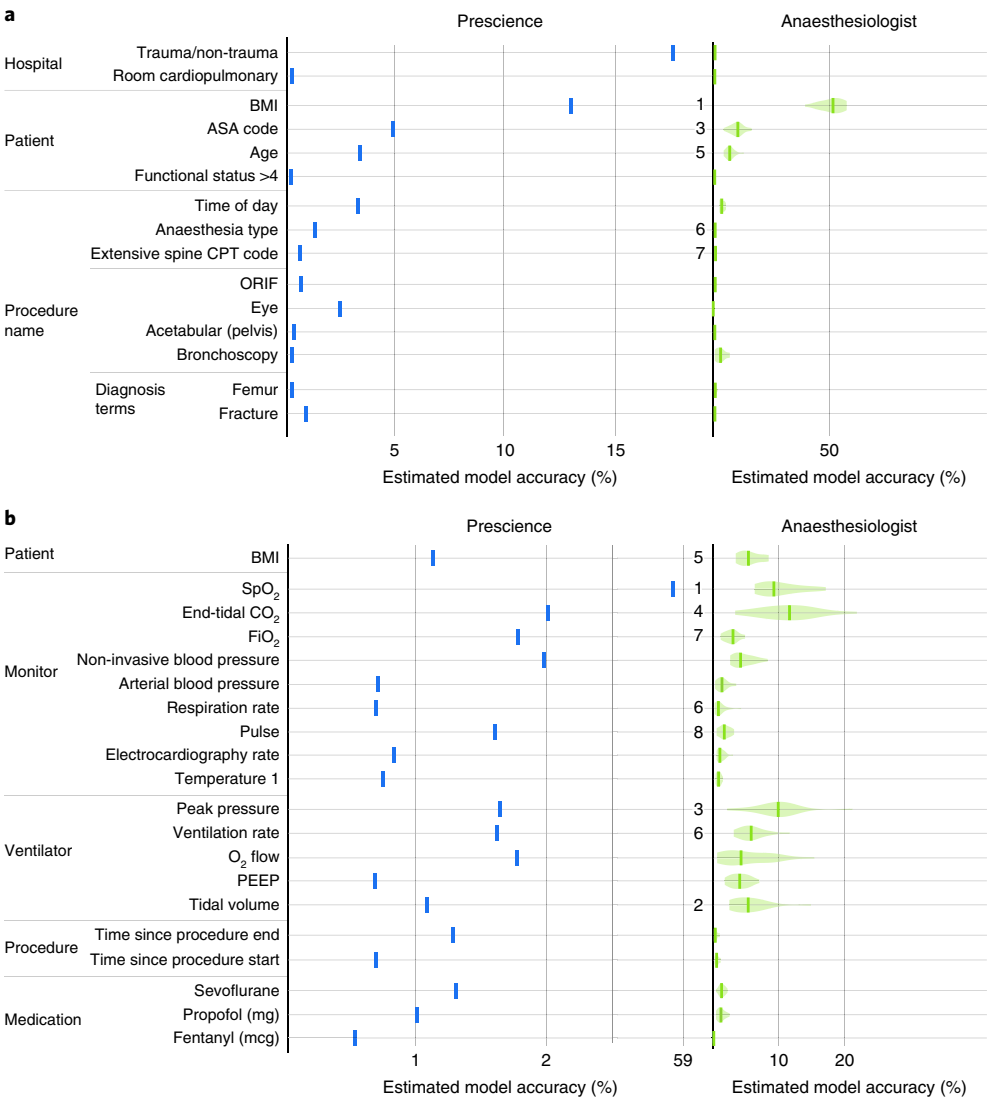

Fig. 5 | Comparison of averaged feature-importance estimates between Prescience and anaesthesiologists for both initial and real-time prediction. **a**,**b**, Importance estimates assigned by the Prescience model (blue) and anaesthesiologists (green) to the top features in both the initial (**a**) and real-time (**b**) prediction. The importance of features is measured as the estimated percentage 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 second (lung disease) and fourth (asthma) features, as ranked by anaesthesiologists for initial prediction (Supplementary Table 2), were not in the top Prescience features. The sixth 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 predictions of an anaesthesiologist when unassisted by Prescience. ORIF, open reduction internal fixation.

precisions are strikingly higher than those we project that anaesthesiologists would achieve on the full test dataset (Supplementary Fig. 10). We also note that the predictive accuracy can be further improved by shortening the predictive window to less than five minutes (Supplementary Fig. 9).

Anaesthesiologists must not only decide when to act to prevent hypoxaemia, but also when not to act. To assist in this, Prescience can predict not only when hypoxaemia will occur, but also when it will not occur. Prescience can predict when hypoxaemia will not occur for 60% of all time points while maintaining a precision of 99.9% (Supplementary Fig. 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 Figs. 1 and 4a), where each group corresponds to the set of extracted patient features from a single input feature in the AIMS dataset, such as the SpO_2 monitor time series. These effects explain why the model predicted a specific risk and thus allow an anaesthesiologist to plan appropriate interventions. In Fig. 1 only the most significant features contributing to hypoxaemia risk are shown for quick reference. However, in Fig. 4 the relative contributions of all patient and case features (that is, attributes) towards hypoxaemia risk can be seen at every sample time point during a procedure (Fig. 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 (Fig. 4a).

The increase in the risk of hypoxaemia in the next five minutes shown in Fig. 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 (that is, the volume of gas exhaled per breath) and the administration of drugs. The risk explanation bar in Fig. 4a has pink features that push the risk higher (to the right) and green features that push the risk lower (to the left). Each group of features is sorted by the magnitude of their impact and the features with the greatest impact are labelled. 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 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 to be predictive of hypoxaemia. For some of these features it is helpful to tag them with indicators of how they relate to hypoxaemia risk before final deployment in an operating room. This can help anaesthesiologists to quickly see non-obvious connections with patient physiology, such as how the muscle relaxant succinylcholine in Fig. 4 does not represent a direct causal impact on hypoxaemia, 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 of a difficult airway during intubation). Figure 4b shows the trend in the Prescience risk predictions over the course of the procedure. The plot in Fig. 4b is equivalent to rotating the feature explanation in Fig. 4a by 90° and then stacking the explanations for each time point horizontally. We can see from the risk trend in Fig. 4b that the large increase in risk at the current time was driven by tidal volume (there was a recent drop in this patient's tidal volume). The future SpO_2 measurements confirm that the patient did indeed progress to hypoxaemia. Not only does Prescience alert anaesthesiologists when a patient's risk of hypoxaemia is high, but it also provides information on the factors driving the risk and their relative contributions. This informed risk prediction enables anaesthesiologists to plan an appropriate course of action to avoid hypoxaemia.

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 Figs. 1 and 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[26]. Estimates of average feature importance are shown in Fig. 5 for both initial prediction and real-time prediction.

To estimate which clinical features anaesthesiologists use to estimate hypoxaemia risk, we first performed a survey before using Prescience, asking four anaesthesiologists to list the most important factors they consider when assessing the risk of hypoxaemia, both before (for initial prediction) and during (for real-time prediction) a procedure. 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 the initial (Fig. 5a) and real-time predictions (Fig. 5b). 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 hypoxaemia, we learned from their behaviour by training a separate gradient-boosting-machine model on the basis of their predictions. This allows a direct comparison between the anaesthesiologists and Prescience on the same set of features. We fit this model to all of the anaesthesiologist relative-risk predictions using ten-fold cross-validation. We then computed the feature-importance estimates for this model that was trained to mimic the behaviour of anaesthesiologists. Given the 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 (Fig. 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, whereas anaesthesiologists necessarily focus on what they consider the most likely causes for hypoxaemia. One striking difference is the reduced role of current SpO_2 levels in the predictions of anaesthesiologists. Although anaesthesiologists are clearly influenced by the recent patterns of patient SpO_2 levels, Prescience strongly depends on these patterns, whereas anaesthesiologists appear to be equally influenced by other factors, such as end-tidal CO_2 (the amount of CO_2 exhaled by the patient) and peak ventilation pressure. The second and fourth ranked features by anaesthesiologists for initial prediction were lung disease and asthma, respectively, which 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 chronic obstructive pulmonary disease, for example, and only 3% of case notes mention asthma.

Our study used data from two hospitals and the initial hypoxaemia predictions were driven by a bias between the two hospitals. This is perhaps unsurprising as one hospital is a Level-1 trauma centre and a significant proportion of its surgical cases involve trauma patients who are more susceptible to hypoxaemia. 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 and age were significant risk factors. These features are well understood in the medical literature as risk factors that can increase the chances of hypoxaemia[27,28]. The American Society of Anesthesiology (ASA) physical status feature represents the severity of the medical condition of a patient and a higher ASA number indicates a higher co-morbidity. Prescience determined that higher ASA physical status values predisposes a patient to a higher hypoxaemia risk. Although this finding may be clinically intuitive, anaesthesiologists can now use this information in their preoperative evaluation as a pre-specified risk factor for intraoperative hypoxaemia. Eye procedures were informative to the model and carried a reduced risk of hypoxaemia, whereas surgeries for fractures had a slightly higher risk. These patterns may reflect the composite risk of hypoxaemia 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, the need for equipment and preparation for the ability to rescue patients from hypoxaemia. 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 the real-time (intraoperative) features, SpO_2 is as expected, the strongest predictor of future potential decreases in SpO_2. End-tidal CO_2 was also a significant intraoperative feature identified by Prescience as predictive of hypoxaemia. Lower values may indicate inadequate ventilation or airway obstruction, which can in turn increase the risk of hypoxaemia. Prescience

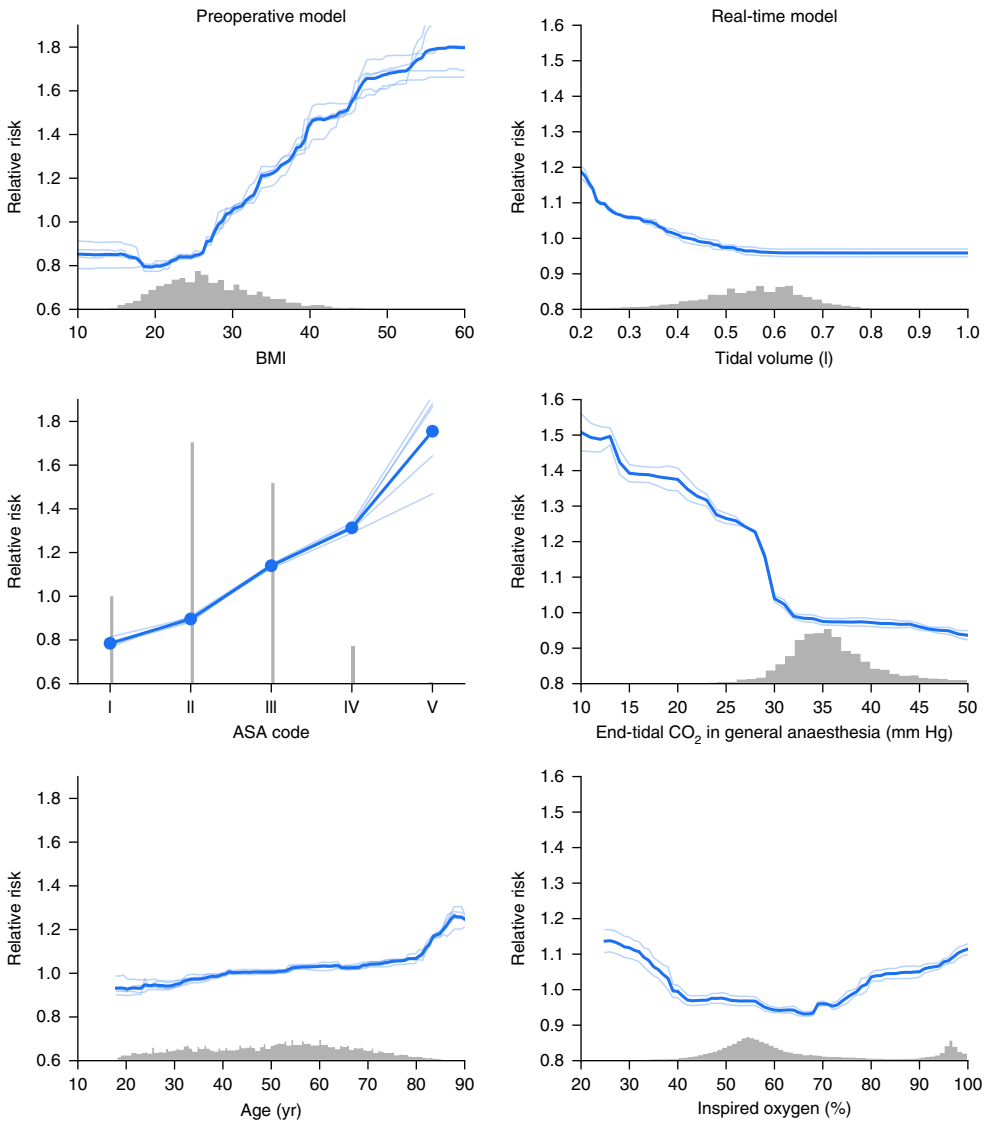

Fig. 6 | Effect of varying individual feature values for both preoperative and real-time features. These partial dependence plots show the change in hypoxaemia risk for all values of a given feature. The grey histograms on each plot show the distribution of values for that feature in the validation dataset. The lighter coloured lines represent model variability from bootstrap resampling of the training data, the dark lines represent the average of the bootstrap runs.

also determined that hypotension (systolic blood pressure below 80 mm Hg) increases the risk of hypoxaemia. On the other hand, moderately higher FiO_2 (inspired O_2 concentration) and positive pressure ventilation can reduce the risk of hypoxaemia, as expected by anaesthesiologists.

Prescience's estimated importance of individual features for hypoxaemia risk highlight important clinical relationships. Three important features for both initial and real-time predictions were chosen to illustrate how the Prescience model modifies hypoxaemia risk on the basis of changes to feature characteristics (Fig. 6). Although many such relationships are present for the various features, Fig. 6 shows a representative selection that demonstrates informative risk relationships that are captured in the Prescience model.

Among the static features, we find that patient BMI has a clear effect on the risk of hypoxaemia. When the BMI is greater than 26, the risk of hypoxaemia increases linearly until it has more than doubled when the BMI is greater than 50. Although a qualitative association between hypoxaemia and body weight is well established in the field of anaesthesia[27,28], Prescience quantifies this relative risk.

Prescience shows that patients with higher ASA physical-status codes have a higher risk of intraoperative hypoxaemia. This is not surprising as higher ASA codes represent increased severity of the physical condition of a patient, such as pre-existing pulmonary and cardiac conditions that can predispose a patient to develop hypoxaemia. Prescience data support clinical observations that the effect of ASA status on the risk of hypoxaemia more than doubles when the ASA status increases from I to V. Advancing age also predicted intraoperative hypoxaemia, probably representing the presence of co-morbidities[27]. These data show that BMI > 30, which meets the clinical definition of obesity[29], is associated with intraoperative hypoxaemia, suggesting impaired pulmonary mechanics. These findings confirm clinical observations and suspicions of the relationship between these patient factors and adverse anaesthesiology outcomes. They also quantify the risk associations, 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 l (keeping all other features the same), Prescience risk for hypoxaemia increases. This increase could be due to hypoventilation, in which case anaesthesiologists can take preventative steps to avoid inadequate ventilation.

Figure 6 shows the relationship between end-tidal CO_2 and risk of hypoxaemia under general anaesthesia. End-tidal CO_2 below 35 mm Hg is associated with an increasing risk of intraoperative hypoxaemia. Although we cannot definitively attribute intraoperative hypoxaemia to hypocapnia, this association may represent underlying patient conditions, such as chronic obstructive pulmonary disease, that affect both physiological conditions. Alternately, the low end-tidal CO_2 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. The current practice is to not 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 hypoxaemia, irrespective of other features. These findings provide important information regarding the safe provision of FiO_2 in patients during general anaesthesia. It is possible that the routine practice of maintaining FiO_2 at 30% or close to room air may be harmful to patients and not desirable. Although 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 hypoxaemia 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 operating room data to predict intraoperative hypoxaemia events before they occur. Based on a comparison with practising anaesthesiologists and existing computational methods applied to other clinical problems, Prescience achieves superior performance when predicting hypoxaemia 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 the ability of an anaesthesiologist to estimate hypoxaemia 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 hypoxaemia risk assessments in the operating room if they had access to Prescience. The augmented-intelligence style approach of Prescience may be particularly helpful for mitigating the unwanted effects of variations in knowledge and/or practice among providers.

Empirically derived black box algorithms such as the bispectral index have been used to track the brain states of patients undergoing general anaesthesia by processing real-time electroencephalograms[30,31]. 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 electroencephalogram 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 developing machine-learning methods to predict intraoperative hypoxaemia, although promising, should still be considered an initial attempt. In this first attempt, we did not categorize procedures to assess hypoxaemia predictions in specific types of procedures. For this reason, the 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 the integration of additional preoperative data, such as the detailed medical history of a patient, 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 the performance of anaesthesiologists that we retrospectively observed in pre-recorded procedures[32].

This paper focuses on hypoxaemia 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 made the explanation tools initially used in Prescience open-source and have continued to improve and extend them (http://github.com/slundberg/shap). Because Prescience effectively decouples the interpretable explanation from the prediction model, we are also able to continue to refine the core prediction model without changing the user experience for anaesthesiologists.

The global risk profiles learned by Prescience (Figs. 5 and 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 hypoxaemia. When considering the standardization of care 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 hypoxaemia adjusted for other factors to the very elderly, those who are overweight and those with more co-morbid conditions. The exact relationships described in Fig. 6 clearly show the patterns and threshold points for the risk. Although low tidal volume is often recommended for patients with acute lung injury[33], these data suggest that low lung tidal volumes are, in fact, associated with intraoperative hypoxaemia. The relationship between low end-tidal CO_2 levels and intraoperative hypoxaemia may reflect underlying critical illness. Despite our inability to fully exclude residual confounding factors, these data shed new light on physiological relationships as well as provide a mechanism to facilitate the provision of anaesthesia care that can mitigate intraoperative hypoxaemia.

As a limitation, we acknowledge that there are several clinical diagnoses that are associated with hypoxaemia, but not directly observable in Prescience. The main clinical diagnoses include mainstem intubation, mucus plug, low FiO_2, low tidal volume, tracheal tube balloon leak, and patient factors such as chronic obstructive pulmonary disease from smoking and pulmonary embolus. Among these, only low FiO_2 and low tidal volume are directly observable in Prescience as 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 hypoxaemia could also have been categorized using ACLS (advanced cardiovascular life support) strategies. However, contrary to a study where factors are a priori identified, Prescience considers all of the available factors and renders an output with associated relative risks. Clinicians must then evaluate the feature relevance on the basis of 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 practised. More and more data-driven predictions of patient outcomes are being proposed and used. However, black-box prediction models, which simply provide predictions without explanation, are difficult for physicians to trust and provide little insight into 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

Institutional review board 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 no. 46889). Protected health information was excluded from the dataset that was used for the machine-learning methods.

Data sources. Our hospital system has installed an AIMS (Merge AIM, Merge Inc.) 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 systems to automatically acquire laboratory and patient registration information. The automatic capture of data is supplemented by the 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 from May 2012 to June 2014. The medical history data of each patient were also extracted from our electronic medical record data warehouse (Caradigm). The high-fidelity anaesthesia record data and the corresponding medical history data from the hospital electronic medical record 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 hypoxaemia, which falls between the intervention level (<94%) and emergency level (<90%) recommended by the World Health Organization[34]. Predictions of hypoxaemia were made for a window of five minutes into the future. If the SpO_2 was ≤92% at any point during those 5 min, 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 performance of the machine-learning algorithm by comparing with anaesthesiologists (Fig. 3), we deliberately chose to use hypoxaemia events that were encountered after a period of stable and normal SpO_2 (Supplementary Fig. 1). This was done to maximize the separation observed between the 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 Fig. 3). The more stringent testing definition used for Fig. 3 excludes some time points, which leads to a smaller set of anaesthesiologist testing labels. Anaesthesiologist testing labels were positive only if SpO_2 was ≥95% for the previous 10 min and then fell below 92% in the following 5 min (Supplementary Fig. 1; left). Anaesthesiologist testing labels were negative only if SpO_2 remained ≥95% for the previous 10 min and the following 10 min (Supplementary Fig. 1; right). All other cases do not have anaesthesiologist testing labels. This more restrictive labelling scheme ensures that positive testing labels are clear drops in SpO_2 levels that would be hard to predict in advance, whereas negative testing labels are clearly not drops in SpO_2 (Supplementary Fig. 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[35].

This means that patients with low risk (from the model) may still need treatment. To address this, the authors of ref. [35] proposed removing examples from the training set where doctors have intervened. This allows one to train a model that predicts patient outcomes without the effect of doctor's interventions. In our case, it is not possible to fully identify when or how an anaesthesiologist is intervening (and if that intervention prevented hypoxaemia), so we sought to address this issue in two ways:

1. It must be recognized that the model predicts hypoxaemia when following standard procedures, not the occurrence of hypoxaemia if the anaesthesiologist takes no action to influence hypoxaemia. 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 versus 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 of the data provided about a procedure are associated with a specific date and time. Text data have the time it was provided, minute-by-minute data from the patient monitor have the time at which each measurement was taken and single point measurements have the times at which 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 value of the data 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 s, 1 min and 5 min half-life times to capture both high and low frequency components of the signal in each time series (Supplementary Fig. 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 min and 1 h were used to capture both near-term and longer-term average drug dosing effects.

To ensure that there was enough training data for each extracted feature, we removed extracted features that had fewer than 100 recorded data values for the real-time model and fewer 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. For the 3,905 extracted features used in intraoperative predictions see Supplementary Table 5. Note that more than 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'[25].

We compared the performance of gradient boosting against three baseline methods: Lasso penalized linear logistic regression; a linear SVM autoregressive model that was proposed for predicting hypoxaemia based only on the SpO_2 data stream in an earlier report[11] and an unsupervised Parzen window method that was used in an earlier report to predict patient deterioration[22]. Gradient-boosting machines significantly outperformed all baseline methods for our primary endpoint, real-time hypoxaemia prediction (Supplementary Fig. 3). For our secondary task of initial prediction, gradient-boosting machines were only slightly superior (Supplementary Fig. 2). The large performance gain of gradient-boosting for intraoperative prediction (Supplementary Fig. 3) is probably because there are 8 million training samples, whereas for preoperative predictions (Supplementary Fig. 2) there are only 42,000 samples and no time-series data. Note that for the initial prediction, the autoregressive support vector machine (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 a 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 for large datasets, so we used XGBoost, a high performance implementation of gradient-boosting machines[26]. For the real-time model we used $\eta = 0.2$ and 1,242 trees, whereas for the initial model we chose $\eta = 0.1$ and 4,000 trees. Using a smaller η value means that more trees are required for fitting, which requires more time to run but results in a smoother (and generally better) model. For both the initial and real-time models we used bagging, where trees were trained on a random 50% sub-sample of the training data. For the preoperative model, the maximum tree depth was four and the minimum child weight of any branch in the trees was one. For the real-time model, the maximum tree depth was six and the minimum child weight of any branch in the trees was ten.

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 that even experts struggle to interpret. This forces a trade-off 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[18,36], 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 that we have developed to compute the Shapley values (that is, the estimated importance of features for a particular prediction) in real-time[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 output of the coalition 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 versus unknown. Some feature values have a large impact on the prediction, whereas 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 and so on) 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 ref. [20] for a full discussion and for connections with several other recent methods in complex model interpretability. In the properties below $f_x(S) = E[f(x) \mid x_S]$, where x_S is a subset of the input vector with only the features in the set S present.

Local accuracy. The local accuracy property (also known as completeness or additivity) is given by the following equality

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

where $\phi_0(f, x) = E[f(x)]$ (the expected value of the model over the training dataset) and M is the number of 'interpretable' inputs, which each correspond to a group of original input features (such as those shown in Fig. 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 s, 1 min and 5 min moving average extracted features and the 5 min 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

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

for all $S \in Z / \{i\}$ 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 also be higher. Note that consistency is known as monotonicity in game theory literature.

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:

$$\begin{aligned}\phi_i(f, x) &= \sum_{S \subseteq S_{\text{all}} / \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \bigcup \{i\}) - f_x(S)] \\ &= \sum_{S \subseteq S_{\text{all}} / \{i\}} \frac{1}{(M \text{ choose } |S|)(M - |S|)} [f_x(S \bigcup \{i\}) - f_x(S)]\end{aligned} \quad (1)$$

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 improvements for tree models (developed after Prescience) 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). Given that 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 dataset 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 $f_x(S_{\text{all}})$ 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 nonlinear link function h such that:

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

As 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 s on our server (in large part because these inputs typically had fewer than 100 non-zero entries). We compared these theoretically grounded explanations with a simple estimate of feature importance to verify that they showed reasonable agreement. The simple method we chose was to replace a single feature group with random values from other samples in the dataset 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 versus the current value. This method is not very scalable and does not account for interactions with other features, yet it is useful to compare with the Prescience explanations to ensure that the Prescience estimates of feature effects are consistent with an intuition of how much a feature's change from its typical value affects the current risk of hypoxaemia (Supplementary Fig. 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 of 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 min (Supplementary Figs. 5–7). For half of the procedures, anaesthesiologists are given Prescience explained risks (Supplementary Figs. 5 and 6) and for the other half they are given the same data but without any Prescience assistance (Supplementary Fig. 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, whereas group B was given Prescience assistance on replicate 2 but not replicate 1. After randomly assigning anaesthesiologists to the groups, three anaesthesiologists from group A and two anaesthesiologists from group B completed the evaluation. We pooled the results within each group and between groups, and the results of this evaluation are shown in Fig. 3. The order in which anaesthesiologists were presented with cases was random across both replicate sets.

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

Code availability. The model-explanation code originally used, and subsequently improved, for Prescience is available open-source at https://github.com/slundberg/

shap. Modelling, processing and web-interface codes specific to Prescience are 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 are available in Supplementary Tables 6 and 7, and data from Fig. 5 are available in Supplementary Tables 8 and 9.

Received: 11 October 2017; Accepted: 31 August 2018;
Published online: 10 October 2018

References

1. Weiser, T. G. et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *Lancet* **385**, S11 (2015).
2. Gawande, A. A., Thomas, E. J., Zinner, M. J. & Brennan, T. A. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery* **126**, 66–75 (1999).
3. Kable, A. K., Gibberd, R. W. & Spigelman, A. D. Adverse events in surgical patients in Australia. *Int. J. Qual. Health Care* **14**, 269–276 (2002).
4. Nair, B. G., Gabel, E., Hofer, I., Schwid, H. A. & Cannesson, M. Intraoperative clinical decision support for anesthesia. *Anesth. Analg.* **124**, 603–617 (2017).
5. Maier-Hein, L. et al. Surgical data science for next-generation interventions. *Nat. Biomed. Eng.* **1**, 691–696 (2017).
6. Dunham, C. M., Hileman, B. M., Hutchinson, A. E., Chance, E. A. & Huang, G. S. 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).
7. Strachan, L. & Noble, D. W. Hypoxia and surgical patients–prevention and treatment of an unnecessary cause of morbidity and mortality. *J. R. Coll. Surg. Edinb.* **46**, 297–302 (2001).
8. Ehrenfeld, J. M. et al. The incidence of hypoxemia during surgery: evidence from two institutions. *Can. J. Anaesth.* **57**, 888–897 (2010).
9. Kooij, F. O., Klok, T., Hollmann, M. W. & Kal, J. E. Decision support increases guideline adherence for prescribing postoperative nausea and vomiting prophylaxis. *Anesth. Analg.* **106**, 893–898 (2008).
10. Garg, A. X. et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* **293**, 1223–1238 (2005).
11. ElMoaqet, H., Tilbury, D. M. & Ramachandran, S. K. Multi-step ahead predictions for critical levels in physiological time series. *IEEE Trans. Cybern.* **46**, 1704–1714 (2016).
12. Lipton, Z. C., Kale, D. C. & Wetzell, R. C. Phenotyping of clinical time series with LSTM recurrent neural networks. Preprint at http://arxiv.org/abs/1510.07641 (2015).
13. Henry, K. E., Hager, D. N., Pronovost, P. J. & Saria, S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* **7**, 299ra122 (2015).
14. Saria, S., Rajani, A. K., Gould, J., Koller, D. & Penn, A. A. Integration of early physiological responses predicts later illness severity in preterm infants. *Sci. Transl. Med.* **2**, 48ra65 (2010).
15. Caruana, R. et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 1721–1730 (ACM, 2015).
16. Deo, R. C. Machine learning in medicine. *Circulation* **132**, 1920–1930 (2015).
17. Memarian, N., Kim, S., Dewar, S., Engel, J. & Staba, R. J. Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. *Comput. Biol. Med.* **64**, 67–78 (2015).
18. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**, 647–665 (2014).
19. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. Know. Disc. Data Min.* 1135–1144 (ACM, 2016).
20. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. In *Adv. Neural Information Processing* 4765–4774 (Curran Associates, 2017).
21. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. Preprint at http://arxiv.org/abs/1802.03888 (2018).
22. Tarassenko, L., Hann, A. & Young, D. Integrated monitoring and analysis for early warning of patient deterioration. *Br. J. Anaesth.* **97**, 64–68 (2006).
23. Summers, R. L., Pipke, M., Wegerich, S., Conkright, G. & Isom, K. C. Functionality of empirical model-based predictive analytics for the early detection of hemodynamic instabilty. *Biomed. Sci. Instrum.* **50**, 219–224 (2014).
24. *Current Procedural Terminology: CPT* (American Medical Association, 2007).
25. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
26. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD Int. Conf. Know. Disc. Data Min.* 785–794 (ACM, 2016).
27. Lumachi, F. et al. Relationship between body mass index, age and hypoxemia in patients with extremely severe obesity undergoing bariatric surgery. *In Vivo* **24**, 775–777 (2010).
28. Kendale, S. M. & Blitz, J. D. Increasing body mass index and the incidence of intraoperative hypoxemia. *J. Clin. Anesth.* **33**, 97–104 (2016).
29. *Defining Adult Overweight and Obesity* (Centers for Disease Control and Prevention, 2016); https://www.cdc.gov/obesity/adult/defining.html
30. Myles, P. S., Leslie, K., McNeil, J., Forbes, A. & Chan, M. T. V. Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial. *Lancet* **363**, 1757–1763 (2004).
31. Avidan, M. S. et al. Anesthesia awareness and the bispectral index. *N. Engl. J. Med.* **358**, 1097–1108 (2008).
32. Epstein, R. H., Dexter, F. & Patel, N. Influencing anesthesia provider behavior using anesthesia information management system data for near real-time alerts and post hoc reports. *Anesth. Analg.* **121**, 678–692 (2015).
33. Guay, J. & Ochroch, E. A. Intraoperative use of low volume ventilation to decrease postoperative mortality, mechanical ventilation, lengths of stay and lung injury in patients without acute lung injury. *Cochrane Datab. Syst. Rev. J.* **2018**, CD011151 (2018).
34. *Pulse Oximetry Training Manual* (World Health Organization, 2011).
35. Dyagilev, K. & Saria, S. Learning (predictive) risk scores in the presence of censoring due to interventions. *Mach. Learn.* **102**, 323–348 (2016).
36. Roth, A. E. (ed.) *The Shapley Value: Essays in Honor of Lloyd S. Shapley* (Cambridge Univ. Press, Cambridge, 1988).
37. *Health, United States, 2016: With Chartbook on Long-term Trends in Health* 314–317 (National Center for Health Statistics, Hyattsville, 2017).

Acknowledgements

We thank G. Erion, M. T. Ribeiro, J. Schreiber and members of the Lee laboratory for feedback and suggestions that improved the manuscript and experiments. This work was supported by National Science Foundation grant nos. DBI-135589 and DBI-1552309, National Institutes of Health grant no. 1R35GM128638, NSF Graduate Research Fellowship grant no. DGE-1256082 and a UW eScience/ITHS seed grant Machine Learning in Operating Rooms.

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. and D.K.-W.L. performed the web-based anaesthesiologist experiments and provided survey data. M.S.V. provided clinical assessment, interpretation of feature importances and connections with anaesthesiologists' workflow. S.-I L. and S.M.L. wrote the paper in conjunction with B.N., J.K. and M.S.V. who wrote the sections on clinical interpretation and integration with current practices. M.H. provided manuscript feedback.

Competing interests

B.N. is an advisor for Perimatics LLC and holds equity in the company. D.K.-W.L. is a Chief Medical Officer for MDmetrix, Inc. The other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41551-018-0304-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.-I.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Su-In Lee

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

For software training we took all surgical records from two university-affiliated hospitals over the course of two years. For software testing we used 5 anaesthesiologists, chosen because they were willing to participate. Our criterium was to include at least 4 anaesthesiologists, to provide replicate results across two groups.

2. Data exclusions

Describe any data exclusions.

We excluded surgical cases with a heart transplant, lung transplant, tracheostomy, and coronary-artery bypass surgeries in which SpO2 and other hemodynamic parameters can be significantly affected by non-physiological measurements such as during cardiopulmonary bypass.

3. Replication

Describe whether the experimental findings were reliably reproduced.

For human test results we observed consistent patterns across anaesthesiologists, as shown in the manuscript. For learning task evaluations we used a held-out test set with thousands of samples from surgery cases that were not used during training, implying strong statistical significance.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Anaesthesiologists were randomized between the two comparison groups. Events used to test anaesthesiologists were randomly selected from the held-out test dataset within criteria described in the manuscript. The training and validation sets were split randomly by surgical procedure.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Anaesthesiologists were blind to their group assignment. Scott Lundberg and Su-In Lee were aware of the group allocation. All investigators were blind to the test set during method development since the test set was randomly divided when the data was first delivered and left in a compressed ZIP file until after algorithm development was frozen.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The algorithms used in this work used custom data-processing methods written in Python and Julia. XGBoost, Scikit-learn and related packages were used during implementation. Code specific to the paper's experiments is available at https://gitlab.cs.washington.edu/prescience (primarily for reference, since the original patient data is not publicly available), and the more broadly applicable explanation methods originally developed for Prescience are available at https://github.com/slundberg/shap

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No cell lines were used.
b. Describe the method of cell line authentication used.	No cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Demographic data is provided in Fig. 2 of the paper.