Development and external validation of temporal fusion transformer models for continuous intraoperative blood pressure forecasting

Lorenz Kapral,[a,b,c,e] Christoph Dibiasi,[a,b,e] Natasa Jeremic,[d] Stefan Bartos,[a,b] Sybille Behrens,[a,b] Aylin Bilir,[a,b] Clemens Heitzinger,[c] and Oliver Kimberger[a,b,*]

[a]Medical University of Vienna, Department of Anaesthesia, Intensive Care Medicine and Pain Medicine, Währinger Gürtel 18-20, Vienna 1090, Austria
[b]Ludwig Boltzmann Institute Digital Health and Patient Safety, Währinger Straße. 104/10, Vienna, 1180 Wien, Austria
[c]Technical University Vienna, Department of Informatics, Research Unit Machine Learning, Favoritenstraße 9/11, Vienna 1040 Wien, Austria
[d]Medical University of Vienna, Department of Ophthalmology and Optometry, Währinger Gürtel 18-20, Vienna 1090 Wien, Austria

eClinicalMedicine
2024;75: 102797
Published Online 30 August 2024
https://doi.org/10.1016/j.eclinm.2024.102797

Summary

Background During surgery, intraoperative hypotension is associated with postoperative morbidity and should therefore be avoided. Predicting the occurrence of hypotension in advance may allow timely interventions to prevent hypotension. Previous prediction models mostly use high-resolution waveform data, which is often not available.

Methods We utilised a novel temporal fusion transformer (TFT) algorithm to predict intraoperative blood pressure trajectories 7 min in advance. We trained the model with low-resolution data (sampled every 15 s) from 73,009 patients who were undergoing general anaesthesia for non-cardiothoracic surgery between January 1, 2017, and December 30, 2020, at the General Hospital of Vienna, Austria. The data set contained information on patient demographics, vital signs, medication, and ventilation. The model was evaluated using an internal (n = 8113) and external test set (n = 5065) obtained from the openly accessible Vital Signs Database.

Findings In the internal test set, the mean absolute error for predicting mean arterial blood pressure was 0.376 standard deviations—or 4 mmHg—and 0.622 standard deviations—or 7 mmHg—in the external test set. We also adapted the TFT model to binarily predict the occurrence of hypotension as defined by mean arterial blood pressure < 65 mmHg in the next one, three, five, and 7 min. Here, model discrimination was excellent, with a mean area under the receiver operating characteristic curve (AUROC) of 0.933 in the internal test set and 0.919 in the external test set.

Interpretation Our TFT model is capable of accurately forecasting intraoperative arterial blood pressure using only low-resolution data showing a low prediction error. When used for binary prediction of hypotension, we obtained excellent performance.

Funding No external funding.

Keywords: Intraoperative hypotension; Continuous prediction; Machine learning; Temporal fusion transformer; Haemodynamic monitoring; Blood pressure forecasting

Introduction

General anaesthesia for surgical interventions routinely involves administrating hypnotics and opioid analgesics to induce a loss of consciousness and tolerance to surgery. Commonly used anaesthetics interfere with the cardiovascular system by reducing cardiac inotropy and systemic vascular resistance, ultimately leading to hypotension.[1] This is further amplified by additional stressors such as hypovolemia, blood loss during surgery or intraoperative positioning (e.g., Trendelenburg position). Intraoperative hypotension, which is commonly defined as mean arterial pressure (MAP)

*Corresponding author. Medical University of Vienna, Department of Anaesthesia, Intensive Care Medicine and Pain Medicine, Währinger Gürtel 18-20, Vienna 1090, Austria.
E-mail address: oliver.kimberger@meduniwien.ac.at (O. Kimberger).
[e]Contributed equally.

Research in context

Evidence before this study

We searched PubMed database, from January 01, 2000, to June 01, 2024, for papers published in English using the terms "blood pressure", "prediction", "hypotension", and "forecasting". Our search yielded 131 results, indicating that intraoperative hypotension is a common occurrence during anaesthesia for non-cardiac surgery that is thought to be associated with postoperative morbidity. Predicting intraoperative hypotension before its occurrence could help anaesthesiologists to initiate prophylactic measures and thereby reduce the incidence of intraoperative hypotension. Existing machine learning algorithms mostly rely on the presence of high-resolution waveform data, which may not be available in many settings.

Added value of this study

We implemented the temporal fusion transformer (TFT) algorithm to predict intraoperative blood pressure trajectories using low-resolution data sampled at 15-s intervals from a large cohort of patients undergoing non-cardiothoracic surgery. We obtained robust predictive performance using low-resolution data, which renders our algorithm potentially more practical in clinical use. In addition to predicting continuous blood pressure values, the TFT model also provides binary predictions of hypotension with excellent discrimination and calibration. In contrast to previous studies, we incorporated data on intraoperative medication.

Implications of all the available evidence

The prediction algorithm developed by us is capable of accurately predicting intraoperative hypotension using low-resolution data. Implementation of our algorithm into clinical practice could help reduce the incidence of intraoperative hypotension, and thereby potentially reduce postoperative morbidity. Future research should prioritise integrating this predictive model into the clinical workflow and evaluating its impact on patient outcomes.

below 65 mmHg,[2] is potentially harmful, being linked to conditions such as myocardial injury,[3] kidney injury,[3,4] delirium[5] and postoperative nausea and vomiting.[6] Therefore, anaesthesiologists monitor patients under general anaesthesia and typically respond to hypotension when it occurs, for example, by administering vasopressors, by giving an intravenous fluid bolus, or by adjusting the depth of anaesthesia, in a *reactive* fashion.[3] However, the actual *prevention* of hypotensive episodes may be advantageous, yet this requires accurate *prediction* of hypotension in advance.

As a result, several tools for predicting intraoperative hypotension in advance have been developed through the use of conventional machine learning methods[7–9] and neural networks.[10–12] These models do not forecast actual MAP values but either make binary predictions (i.e., the patient will be hypotensive or not)[9,13] or provide a dimensionless number indicating the probability of hypotension.[14] In addition, those models are limited in terms of the input variables used for prediction because they mainly employ past vital signs and data on patient demographics. There is also discussion whether existing prediction models are superior to simply extrapolating the MAP trajectory.[15] Finally, most of the existing models require the use of high-quality arterial blood pressure waveform data and cannot be used when invasive arterial blood pressure monitoring is not in use.

There have been recent technical advances in time series data forecasting: The novel temporal fusion transformer (TFT) algorithm is an attention-based model that is designed for advanced multi-horizon forecasting.[16] It employs recurrent layers to effectively process short-term temporal patterns while using interpretable self-attention layers to understand long-term dependencies.[17] Hence, it can appropriately integrate static, time-stamped and time series data. In addition, the TFT algorithm can selectively focus on the relevant data points that are the most important for its forecast while filtering out nonessential elements.[18]

We hypothesised that the TFT algorithm would be well suited to predict intraoperative blood pressure trajectories and that it could be used to predict the occurrence of intraoperative hypotension, even with low resolution vital sign data. Therefore, we trained a TFT model to predict intraoperative MAP using a data set consisting of pre- and intraoperative data collected during routine patient care. To evaluate our model's performance, we assessed discrimination and calibration in both internal and external validation.

Methods

This retrospective observational study was performed after approval of the Ethics Committee of the Medical University of Vienna (reference number 2387/2020, January 19, 2021). Given the retrospective nature of the study, the requirement for informed consent was waived.

We screened all patients who underwent anaesthesia at the General Hospital of Vienna between January 1, 2017, and December 30, 2020, for eligibility. The General Hospital of Vienna is a tertiary academic hospital in Vienna, Austria. Anaesthesia is conducted by resident and consultant anaesthetists from the Department of Anaesthesia, Intensive Care Medicine and Pain Medicine of the Medical University of Vienna.

Patients older than 18 years at the time of surgery who had general anaesthesia performed for a diagnostic or surgical intervention were included. We excluded patients who had cardiac, thoracic and/or vascular surgery and patients who had neuraxial, regional or local anaesthesia without general anaesthesia. We defined general anaesthesia as the administration of sedatives and invasive mechanical ventilation (either via laryngeal mask, endotracheal intubation or tracheostomy).

Preprocessing

We generated the data set from pre-, intra-, and postoperative data recorded for routine patient care in the patient data management system (IntelliSpace Critical Care and Anaesthesia, Philips Austria GmbH, Vienna, Austria). The following variables were static: age, sex, weight, American Society of Anaesthesiologists (ASA) score and surgical urgency (elective/urgent/emergency). The following variables were time series: heart rate (beats per minute), pulse rate (beats per minute), peripheral transcutaneous oxygen saturation (SpO2, %), non-invasive systolic, diastolic, and mean blood pressures (each in mmHg), invasive systolic, diastolic, and mean blood pressures (each mmHg) and end-tidal partial pressure of carbon dioxide (etCO2; mmHg). Anaesthetic agents, ventilation parameters and perfusion parameters were time-stamped but processed as time series; Supplemental Table S1 lists all the input variables.

The vital parameters heart rate, pulse rate, and SpO2 were available at a 15-s resolution. Invasive blood pressure was also available at a 15-s resolution while non-invasive blood pressure was available at a 3-min interval. We sampled all other time series variables including non-invasive blood pressure up to a 15-s resolution.

We grouped input features by type, differentiating between categorical and numerical variables as well as time-dependent and static variables. We checked the values of the input features for plausibility by analysing the maximum, minimum, and frequency distribution. Using the 'forward fill' method,[19] we replaced implausible and missing values, as detailed Supplemental Table S2, which lists their frequency of missingness. We scaled numerical variables to a standard deviation of 1 and a mean value of 0. Categorical variables underwent a one-hot encoding process, transforming each categorical variable into a dichotomous variable.

We split the complete data set into training set (70%), validation set (20%) and holdout test set (10%). This was done by randomly assigning patient IDs to each set. To prevent any potential leakage of data between different patients, we grouped each patient's data independently.

Model development

Google DeepMind's GitHub repository served as the foundational framework for the development of this TFT model.[20] We modified the model to handle data sets lacking future-known time points. To enhance the model's performance evaluation, we incorporated the metrics discussed in model evaluation. We integrated TensorBoard—a tool to visualise metrics—to track the training process.

We configured the TFT model to use the previous 32 values, corresponding to an input time interval of 8 min, for each variable to predict the subsequent 28 MAP values spanning 7 min. If the surgery duration was shorter than the combined duration of the input and output time intervals, the patients were excluded from training. When less than 8 min of history were available, we padded the oldest data point to form a complete input window for prediction.

We trained the model on the training set and evaluated its performance on the validation set every 10 epochs. To prevent overfitting, we stopped the training early if the error in the validation set did not reach a new optimal value for three consecutive iterations.

The TFT model was optimised using a 'RandomSearch' algorithm, focusing on the optimisation of several parameters, including batch size, learning rate, number of attention heads, number of hidden neurons, dropout rate and length of the input sequence; the final hyperparameters can be found in Supplemental Information S1.

Internal and external validation

We evaluated both MAP predictions themselves as well as binary predictions of whether hypotension will occur (defined by MAP < 65 mmHg). We used the holdout test set for internal validation and generated an external test set using the open public database 'Vital Signs DataBase' (VitalDB),[21] which contains high-resolution intraoperative data from 6388 patients. We transformed VitalDB data to match the format of our training data set.

Continuous MAP prediction

We evaluated continuous MAP predictions using two different metrics: mean squared error (MSE) and mean absolute error (MAE). MSE is the average of the squared differences between predicted and actual values, and MAE is the average of the absolute differences between predicted and actual values. MSE emphasises large errors, whereas MAE treats all errors equally, is easy to interpret and can be directly translated into units such as mmHg. We calculated the cumulative average of these metrics across all patients in the holdout test sets. This involved calculating the mean of all errors from the 28 predicted values for each data point of each patient in the test set.

Binary prediction of hypotension

To generate binary predictions of hypotension, we extracted the continuous MAP predictions at one, three, five, and 7 min (Fig. 1). We used these values to

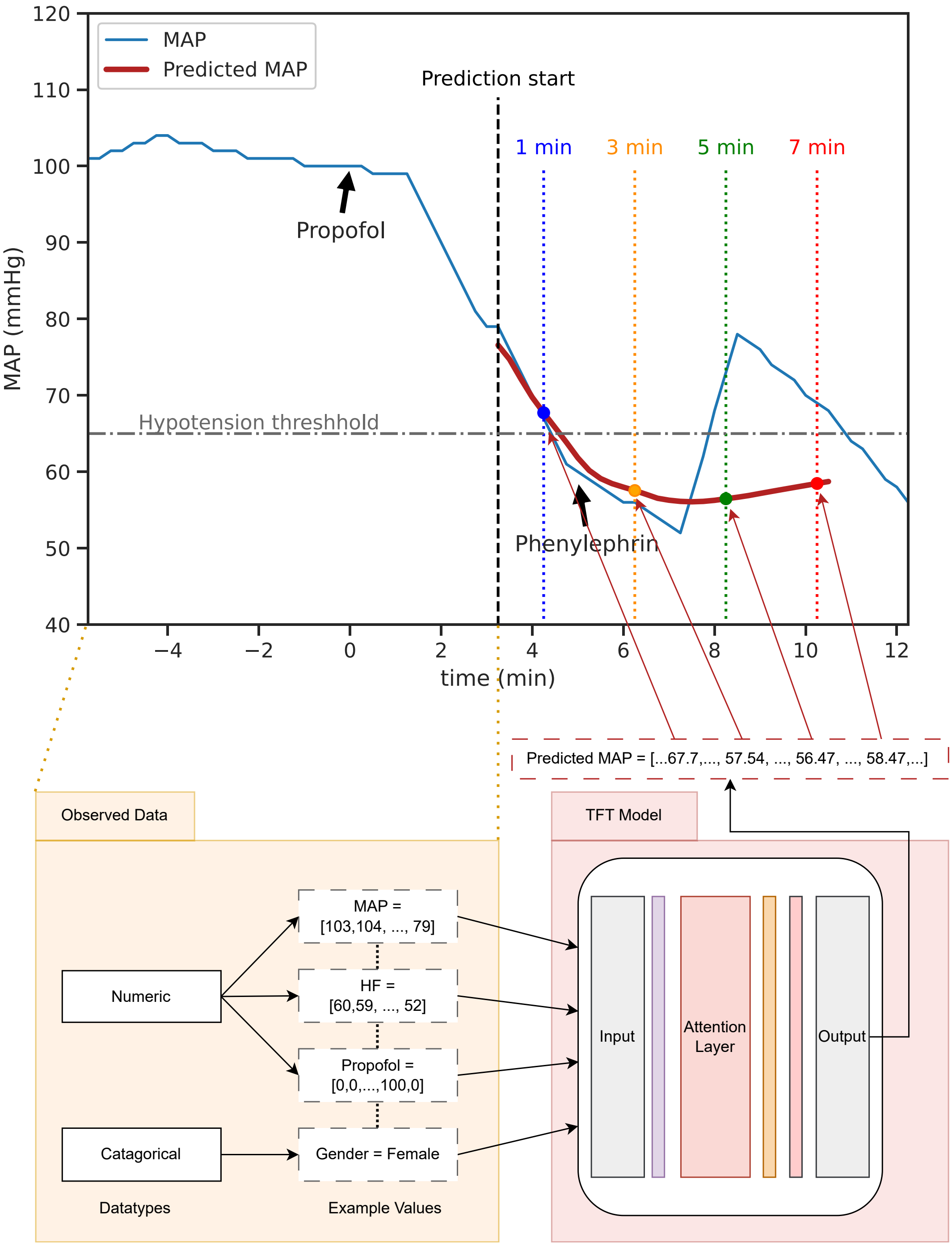

Fig. 1: Prediction of mean arterial pressure. A graphical representation of the temporal fusion transformer (TFT) model prediction process for mean arterial pressure (MAP). The top graph shows the observed MAP over time, the model predicted values and expected future MAP. The lower left section details the data input structure, separating real values and categorical data, with example values given. The bottom right shows a simplified architecture of the TFT model, highlighting the input, attention layer and output. The blue, orange, green and red lines indicate the specific time points used to assess hypotension, corresponding to predictions made 1, 3, 5, and 7 min into the future, respectively. The hypotension threshold was set at 65 mmHg. Propofol leads to arterial hypotension which is counteracted by the alpha-adrenergic agent phenylephrine. As the administration of phenylephrine occurs after the prediction start, it cannot be taken into account for forecasting MAP.

construct a binary prediction model that could estimate the likelihood of hypotension, defined as a MAP < 65 mmHg.

The model provided a range (lower and upper limits) for each of the 28 values. To evaluate the model using metrics such as which require probabilities rather than 'true' or 'false', we fit a Gaussian curve with the lower and upper limits. This allowed us to calculate probabilities.

For example, in the scenario shown in Fig. 1, the MAP values [68, 58, 57, 59] over four consecutive time points translated into a binary sequence of [false, true, true, true] with a decision threshold of 0.5, meaning that any probability greater than 50% was interpreted as a prediction of hypotension.

We calculated the following metrics for evaluating the binary hypotension predictions: Accuracy quantified the overall correctness of the model across all classes. Sensitivity (true positive rate) and specificity (true negative rate) measured the model's ability to correctly identify positive and negative cases, respectively. The positive predictive value (PPV) and negative predictive value (NPV) reflected the accuracy of positive and negative predictions. The area under the receiver operating characteristic curve (AUROC) assessed the ability of the model to discriminate between classes. Calibration slope, intercept and expected calibration error (ECE) together measured the agreement of the predicted probabilities with the observed outcomes and indicated the probabilistic accuracy of the model. To visualise these metrics, we plotted the receiver operating characteristic (ROC) curve and calibration plot.

Comparison with the XGB model

To establish a benchmark for the TFT model, we also used the extreme gradient boosting (XGB)[22] algorithm on the same training data set used for the TFT model as a way to train several models predicting the binary occurrence of hypotension at one, three, five, and 7 min.

We vectorised time-dependent variables into sequences and transformed them into unit scale. Separate XGB models were trained and optimised to predict occurrences of hypotension at one, three, five, and 7 min into the future.

We assessed the performance of the XGB model using the same metrics as those applied to the TFT model.

Interpretability

The attention mechanism allowed the model to focus on the most relevant aspects of the input data by assigning different levels of attention to different input parameters and acting as a filtering mechanism.[17]

To visualise the model's focus and determine the importance of temporal inputs, we computed the sum of the attention values assigned to all features at each time point. This allowed us to visualise the importance of each time step within the input sequence. In parallel, we assessed the weight of each input parameter across the data set by summing its attention values across all time points, thereby ranking its overall importance to the model's output.

In addition, we conducted experiments to investigate the influence of medication data on the behaviour of the model. After completing the training, we artificially manipulated the input data by omitting medication information and measured the effect of these differences on the predicted MAP over the next 3 min. This approach was only undertaken to provide insight into the extent to which the model was being influenced by medication data.

Statistical analysis

Because of patient privacy concerns and the regulations of the Medical University of Vienna, all data used to train the model are not available for public release in their current format. The external database, which was utilised for validation purposes, is openly available, enabling replication of the validation process.[21] The code for model training and evaluation is available (https://github.com/lorenzkap/MAP_TFT). We performed all calculations with R and Python 3.11.3, TensorFlow 2.12.0, and Scikit-learn 1.2.2.

Role of the funding source

This study was funded by institutional funds of the Medical University of Vienna, Department of Anaesthesia, Intensive Care Medicine and Pain Medicine and the Ludwig Boltzmann Institute Digital Health and Patient Safety.

Results

We screened data from 88,016 anaesthesia cases and included data from 81,122 cases in the final data set. The baseline characteristics of the anaesthesia cases analysed are given in Table 1. The internal data set was split randomly into training (70%), validation (20%), and holdout (10%) test sets, consisting of 56,785, 16,224 and 8113 cases. We tested the final algorithm in an external test set consisting of 5065 cases. Details of the external test set are given in Supplemental Table S3.

Continuous MAP prediction

We trained the TFT model to predict the continuous MAP trajectory for the next 7 min (Fig. 1; Supplemental Fig. S1), here by utilising 52 input features (Supplemental Table S2). In the internal test set, MSE was 0.405 standard deviations and MAE 0.376 standard deviations, corresponding to an average prediction error of 4 mmHg off the actual measurements. In the external test set, the average MSE was 1.165 standard deviations, and the average MAE was 0.622 standard deviations, or 7 mmHg. In both the internal and the external test sets, MAE was reduced when the forecast distance was lower and vice-versa (Fig. 2).

	N = 81,121
Age (years)	52 (34, 70)
Male sex (–)	35,730 (44%)
ASA score	
1	36,272 (27%)
2	36,272 (45%)
3	20,251 (25%)
4	2066 (2.5%)
5	730 (0.9%)
Surgical urgency (–)	
Elective	64,855 (80%)
Emergency	3459 (4.6%)
Urgent	12,808 (16%)
Duration of surgery (min)	132 (6, 296)
Surgical discipline	
General surgery	20,881 (26%)
Orthopaedics/Trauma surgery	16,098 (20%)
Plastic surgery	3153 (3.9%)
ENT	6110 (7.5%)
Maxillofacial surgery	3532 (4.4%)
Neurosurgery	5240 (6.5%)
Gynaecology	8905 (11%)
Obstetrics	5569 (6.9%)
Urology	6849 (8.4%)
Ophthalmology	4123 (5.1%)
Dermatology	656 (0.8%)
Undefined	6 (<0.1%)

1 Median (IQR); n (%)

Table 1: **Patient characteristics: primary data set.**

A key feature of the TFT model was considering past data to predict blood pressure. The TFT model utilised medication data, for example, intravenous anaesthetics or vasopressors, to predict blood pressure. The model reacted to medication and its predictions became better when medication data was present (Fig. 3, Panel a). The model's attention mechanism can filter the data for more relevant time stamps (Fig. 3, Panel c). The top features selected for blood pressure predictions are shown in Fig. 3, Panel b, and the influence of the most common drugs on the prediction of the model in the data set is depicted in Fig. 3, Panel d.

Binary prediction of hypotension

We predicted the likelihood of blood pressure falling below 65 mmHg at one, three, five, and 7 min in the future by using specific quantiles of blood pressure predictions and compared these predictions with those from an XGB model (Fig. 4). In the internal test set, both the TFT and XGB models had area under the receiver operating characteristic curve (AUROC) scores above 0.9 (Table 2; Fig. 4) although the XGB model had slightly superior discrimination compared with the TFT model at the five- and 7-min marks. For both models, discrimination was reduced in the external test set. The TFT model was consistently able to discriminate between timepoints with and without hypotension when the forecast distance was increased from one to 7 min, but discrimination of the XGB model declined with increasing forecast distance, as evidenced by lower AUROC (Table 2; Supplemental Tables S4 and S5).

The calibration plots are shown in Fig. 5. The TFT model demonstrated an ECE ranging from 0.05 to 0.11 in the internal test set and from 0.06 to 0.08 in the external test set (Table 3). In both test sets, the TFT model had a calibration slope of less than one, indicating a tendency to overestimate the likelihood of hypotension (Fig. 5 Panel a, c; Supplemental Table S6). The XGB models showed good calibration in the internal test set (ECE < 0.03). However, the XGB models were poorly calibrated in the external test set (ECE >

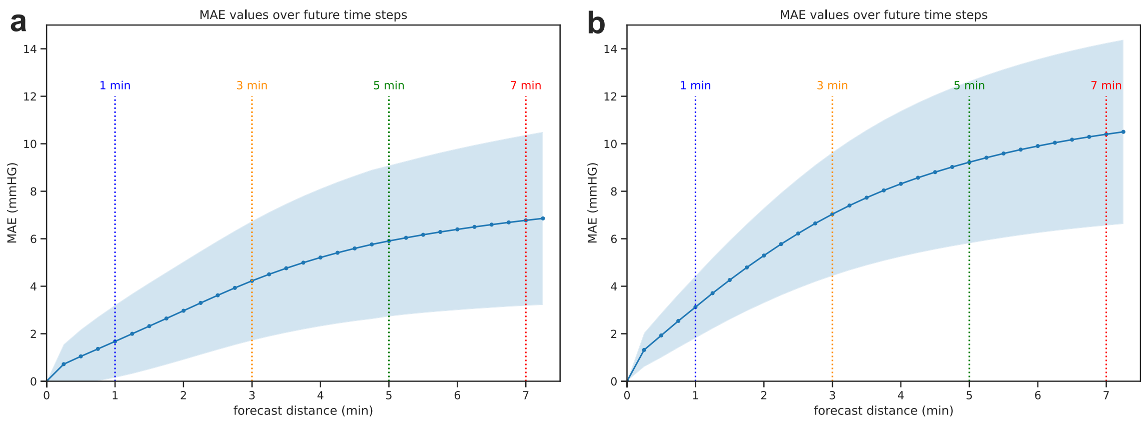

Fig. 2: Performance for continuous blood pressure prediction. Mean absolute error (MAE) of the temporal fusion transformer model for continuous prediction of intraoperative blood pressure in the internal (**a**) and external (**b**) test sets. The standard deviation of all MAEs is indicated by the lighter blue area.

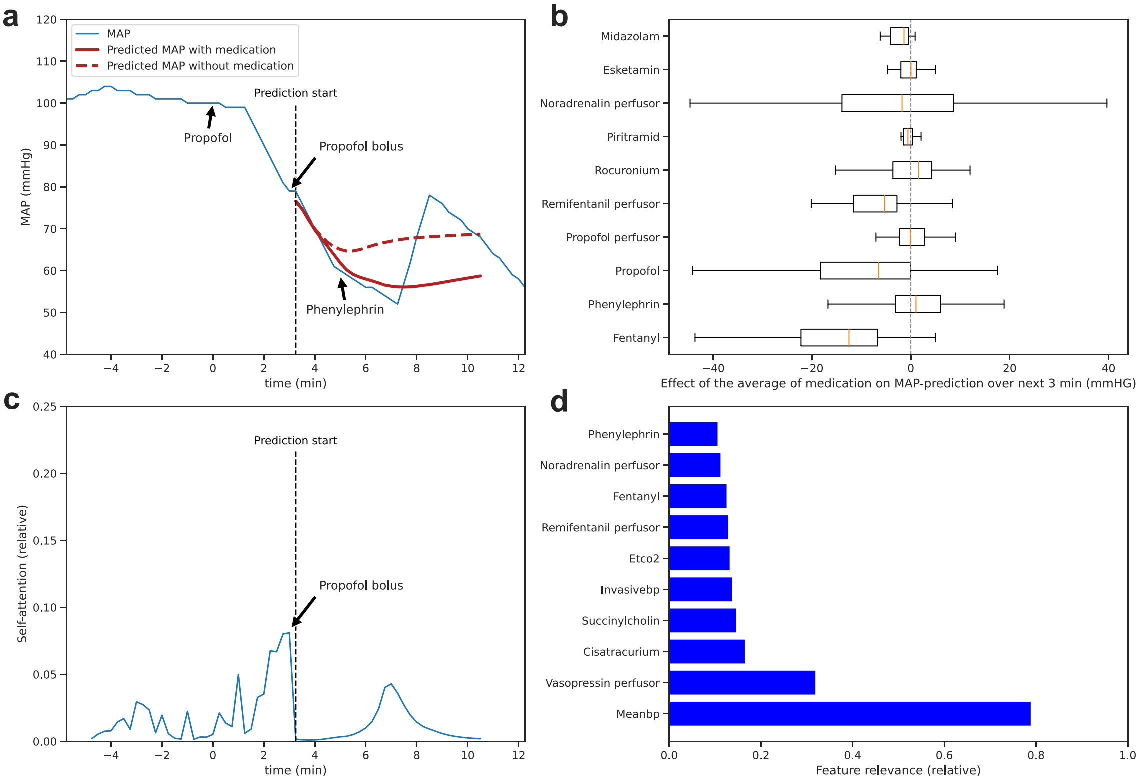

Fig. 3: Importance of medication and attention mechanism. (**a**) is a representative example of continuous mean arterial pressure (MAP) predictions using the temporal fusion transformer (TFT) model and shows that predicted MAP varies significantly when data on the use of propofol is included in the model vs. when these data are omitted. (**b**) shows the impact of the 10 most administered drugs on the predicted MAP over the next 3 min as predicted by the TFT model. The drugs are normalised by their average dosage because of their varying effects per milligram. (**c**) shows the relative importance of each time step in the model's input window. Self-attention in transformer models selectively focuses on the most relevant parts of the input. It highlights a significant increase in the importance of the time steps when propofol is administered, underscoring its influence on the model's output. (**d**) depicts the top 10 features that the model considers as being the most critical to its blood pressure predictions. Among these, historical MAP data stand out as the most influential factor for subsequent MAP predictions.

0.15 for 7 min) and overestimated the occurrence of hypotension (Fig. 5 Panels b, d; Supplemental Table S6).

Discussion

In the present study, we used the TFT algorithm to develop a predictive model 1) for continuously forecasting intraoperative blood pressure trajectories for the next 7 min and 2) for binarily predicting the occurrence of hypotension (defined as MAP below 65 mmHg) within the next one, three, five, and 7 min. We validated our model using internal and external test sets and found that our model predicted MAP with a low predictive error of 4 mmHg, respectively 7 mmHg in the internal and external test sets. Using the dichotomised TFT model, we obtained excellent discrimination and reasonable calibration for binary prediction of the occurrence of hypotension.

Predicting vital sign derangements, such as hypotension, is a well-established problem, and multiple studies from the field of anaesthesia and critical care medicine have used different study designs and computational algorithms to solve it.[8,10,11] For instance, Kendale et al. utilised multiple machine learning techniques to binarily predict the occurrence of hypotension (defined as a single MAP value below 55 mmHg) after the induction of anaesthesia. Jo et al. used deep learning models trained on high-resolution waveform data from VitalDB to predict intraoperative hypotension.[23] Hatib et al. and Davies et al. similarly applied deep learning to binarily predict hypotension (defined by them as MAP below 65 mmHg). Their model, which is commercially available[11,14], provides users with the hypotension prediction index (HPI), a dimensionless number ranging from 0 to 100, which indicates the likelihood of hypotension within the next 15 min. One of the key features distinguishing our TFT model from those works is the fact that our model directly predicts the course of MAP together with an uncertainty interval. In theory, this could be more readily interpretable by clinicians than an

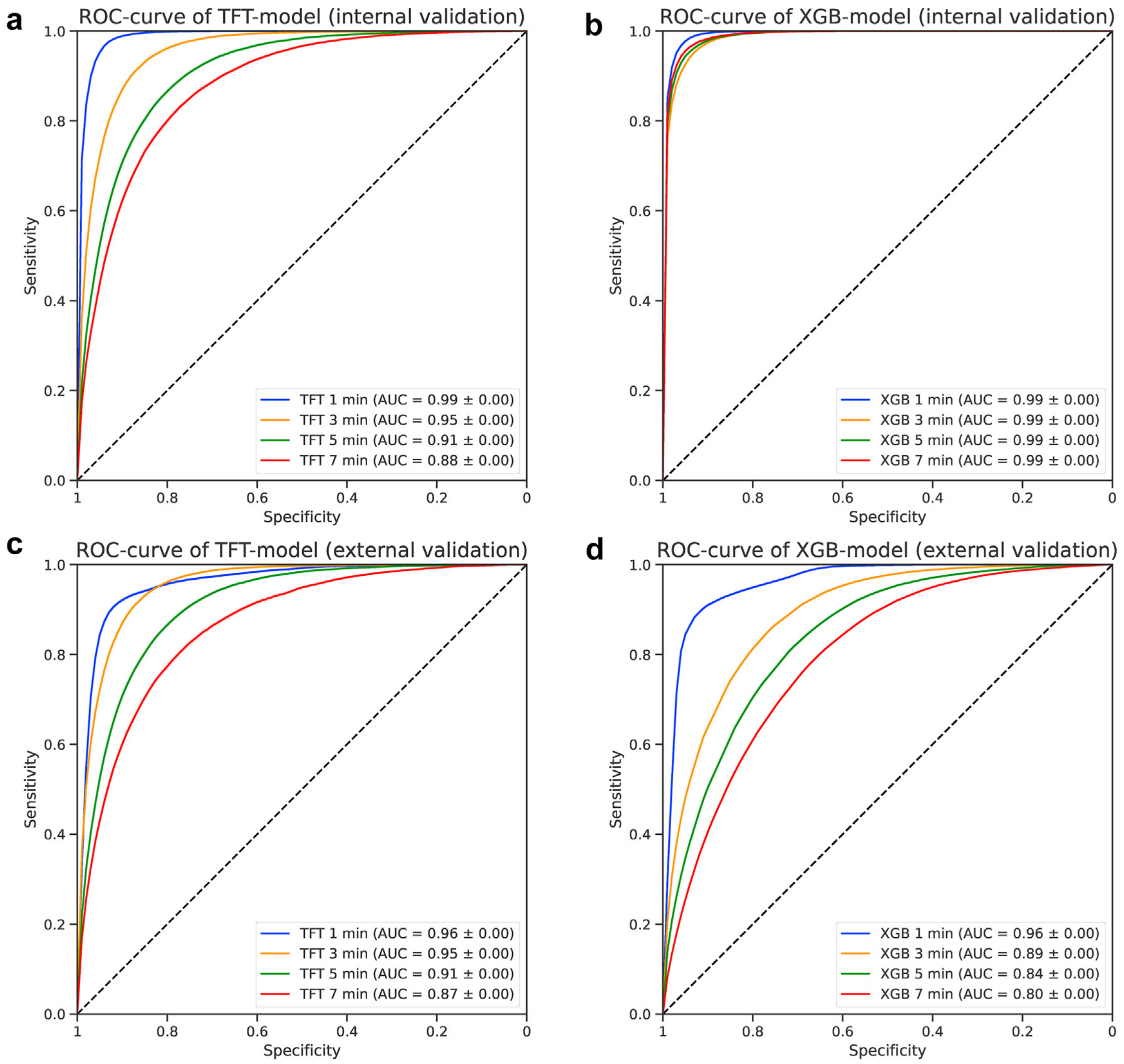

Fig. 4: Performance in binary prediction of hypotension. Receiver operating characteristic (ROC) curves for the temporal fusion transformer (TFT) model (**a, c**) and extreme gradient boosting (XGB) model (**b, d**) across time frames of 1, 3, 5 and 7 min for the prediction of hypotension in the internal and external validation. Area under receiver operating characteristic (AUC) values demonstrate high accuracy for both classifiers internally, with a modest decline externally. The TFT classifier shows a small drop in performance over time, while the XGB-classifier exhibits excellent internal but diminished external performance.

Forecast time	Internal validation		External validation	
	TFT	XGB	TFT	XGB
1 min	0.9883 (0.9880, 0.9886)	0.9941 (0.9939, 0.9943)	0.9598 (0.9590, 0.9607)	0.9607 (0.9602, 0.9612)
3 min	0.9544 (0.9536, 0.9553)	0.9874 (0.9871, 0.9878)	0.9453 (0.9444, 0.9462)	0.8909 (0.8900, 0.8918)
5 min	0.9095 (0.9083, 0.9107)	0.9893 (0.9890, 0.9896)	0.9032 (0.9017, 0.9046)	0.8420 (0.8409, 0.8432)
7 min	0.8800 (0.8785, 0.8816)	0.9908 (0.9905, 0.9910)	0.8667 (0.8648, 0.8686)	0.7981 (0.7968, 0.7994)

Area under the receiver operating characteristic (AUROC) of the temporal fusion transformer (TFT) and the extreme gradient boosting (XGB) model in internal and external validation. The forecast time indicates the time before a hypotensive event. The 95% confidence interval is indicated by the values within the brackets.

Table 2: **AUROC in internal and external test set.**

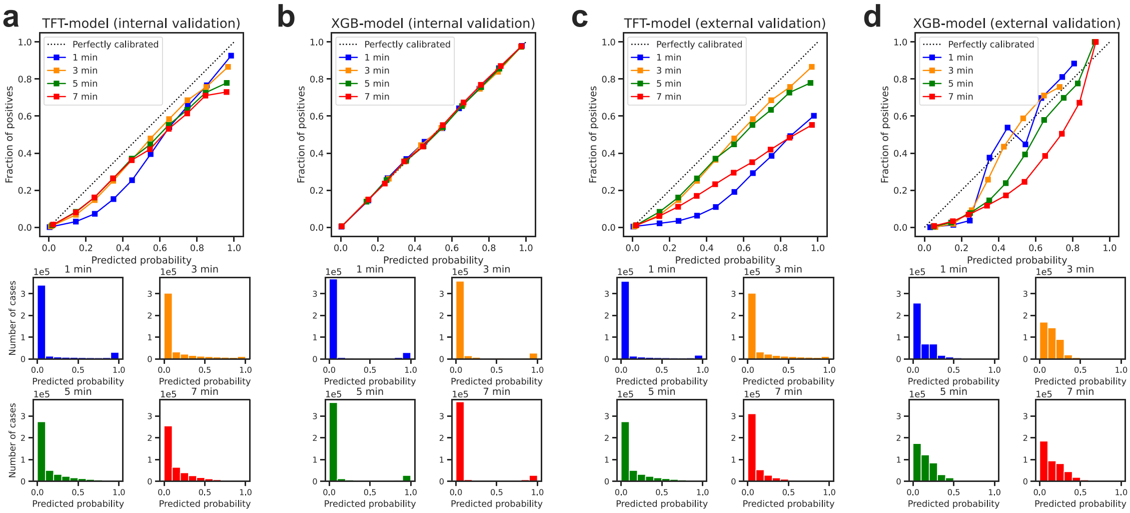

Fig. 5: Calibration curves for binary prediction of hypotension. Calibration curves for the temporal fusion transformer (TFT) model (**a, c**) and extreme gradient boosting (XGB) model (**b, d**) at 1, 3, 5 and 7 min for both internal and external validation for the prediction of hypotension. The graphs compare the predicted probabilities of positives against the actual proportion of positives, with the dotted line representing perfect calibration. The corresponding histograms below the calibration curves show the distribution of predicted probabilities at each time interval. The closer the calibration curve is to the dotted line, the better the calibration of the model. The histograms give an indication of the frequency and confidence of the classifier's predictions.

arbitrary index, and in addition, the length and severity of hypotension is easily visible, which is not the case with the models from Kendale et al. and with the HPI. The second key feature of the TFT model is the use of low-resolution data. In contrast to previous works, which have used waveform data that requires the invasive placement of an arterial line, we utilised vital signs data that is sampled every 15 s. Still, our TFT model showed similar discriminative performance compared with the HPI for predicting hypotension 5 min before it occurred, with an AUROC of 0.909 (TFT) compared with 0.926 (HPI). The higher specificity of the TFT model (0.960 compared with 0.858 for the HPI) could be advantageous because false positive predictions are less likely with our model, potentially reducing alarm fatigue. Notably, the HPI has recently been criticised for selection bias being present during training and validation, leading to data leakage which potentially falsely elevates its performance metrics.[24] As such, it has been suggested that HPI may not be superior to setting the mean blood pressure alarm threshold in the range of 70–75 mmHg.[24] Because our model utilises the TFT algorithm that is specifically designed for the prediction of time series data, we avoided such bias.

We conducted a series of tests on a range of models (LSTM, ARIMA, XGB, transformers) for the continuous MAP forecast. However, the results indicated that these models were not optimal. The TFT model demonstrated superior performance when applied to medical data. Consequently, we concentrated our efforts on the TFT model in our publication.

To the best of our knowledge, only the prediction model from Lee et al. could forecast continuous intraoperative blood pressure values similar to our TFT model; they applied a deep learning technique to predict blood pressure as well as hypotension (i.e., the

Forecast time	Internal validation		External validation	
	TFT	XGB	TFT	XGB
1 min	0.0259 (0.0255, 0.0264)	0.0008 (0.0004, 0.0011)	0.0529 (0.0524, 0.0533)	0.0791 (0.0788, 0.0793)
3 min	0.029 (0.0283, 0.0298)	0.0008 (0.0005, 0.0012)	0.0478 (0.0473, 0.0482)	0.1060 (0.1057, 0.1063)
5 min	0.0346 (0.0337, 0.0353)	0.0007 (0.0004, 0.0010)	0.0465 (0.0459, 0.0469)	0.1076 (0.1073, 0.1079)
7 min	0.0398 (0.0391, 0.0404)	0.0008 (0.0005, 0.0011)	0.0471 (0.0466, 0.0475)	0.1166 (0.1163, 0.1169)

Expected calibration error (ECE) of the temporal fusion transformer (TFT) and the extreme gradient boosting (XGB) model in the internal and external validation. The forecast time indicates the time before a hypotensive event. Low values represent a low error, thus better calibration. The 95% confidence interval is indicated by the values within the brackets.

***Table 3:* Expected calibration error in the internal and external test sets.**

occurrence of blood pressure below 65 mmHg) within the next 5, 10, and 15 min using data from VitalDB, the database we used for external validation.[12] However, in the present study, we obtained lower predictive errors than in their study (MAE 4 mmHg in the internal test set and 7 mmHg in the external test set vs. 7 mmHg in the study from Lee et al.), even though we utilised lower-resolution data (sampled once every 15 s) as opposed to high-quality waveform data. In addition, their model was limited to predicting a single MAP value, whereas our model predicted an entire curve consisting of 28 different values, which can facilitate easier interpretation in the operating room. The TFT model also incorporates data on administered medication, such as hypnotics, analgesics and vasoactive agents, intraoperative ventilation parameters and intraoperative positioning. These features set our model apart from previous studies and are—in our opinion—the most important factor explaining the model's good performance. Our analysis also showed that data on administered medications were critical for the TFT model in predicting the blood pressure trajectory. Data on the use of propofol were especially used to improve MAP predictions, and the predictive error increases, for example, when data on the use of propofol were missing.

Fig. 3, Panel b, illustrates the directional influence of commonly administered drugs on blood pressure. The graph, created by excluding these drugs from the test set and analysing the prediction curves from Fig. 3, Panel a, shows an expected decrease in blood pressure when fentanyl or propofol are administered; however, the effect of noradrenaline varies widely, despite its known pressure-increasing effect. This variability may be attributed to patients entering the dataset with an active noradrenaline perfusor or the fact that the noradrenaline perfusor is often initiated and adjusted early to stabilise blood pressure, then maintained at a consistent level, resulting in minimal fluctuations during surgery. This may mask the actual influence of noradrenaline on blood pressure. Another potential use case of our TFT model could be the calculation of the 'optimal' dose of hypnotics/analgesics during the induction of anaesthesia. Furthermore, the black box problem of machine learning algorithms was alleviated by indicating the probability of the occurrence of hypotension as well as the time-resolved representation (Fig. 3, Panel d) of the essential features for decision-making.[25] For example, the model primarily uses the past MAP-values (Fig. 3, Panel d) for predicting MAP. Furthermore, it can identify significant occurrences such as the administration of propofol (see Fig. 3, Panel c).

To facilitate a comparison with previous studies, we used the results of the TFT model for binary predictions of the occurrence of hypotension. The discrimination of our model was superior to previously published works.[8,11,12] The generalisability of an algorithm was a persistent challenge that complicated the implementation of machine learning algorithms in clinical practice.[26] Our approach to predicting hypotension by directly calculating the MAP curve rather than providing an index offered additional robustness, as confirmed by external validation. Compared with the XGB models, which have previously been shown to have excellent performance in binary classification tasks, such as predicting hypotension[13,27] trained on the same data set, our model demonstrated greater robustness. This was evidenced by its superior performance on the external test set, even though the XGB models performed better on the internal test set and were trained on simpler tasks (hypotension: yes/no).

Although the model was reasonably calibrated in the internal test set, it overestimated the occurrence of hypotension in the external test set (Fig. 5). Miscalibration is a common phenomenon when predictive models are tested in a population that they were not developed in28, highlighting that predictive models should be carefully tested prior to implementation into clinical practise.[29] However, this overestimation is not necessarily an error of the TFT model but is rather a reflection that the model is not anticipating future medical interventions, even though we trained the model on retrospective surgical cases in which clinicians intervened during adverse events. For instance, if the model detected a potential drop in blood pressure, it could predict the onset of hypotension. However, in an actual OR scenario, clinicians often intervene to prevent such events. Therefore, a 'good' model should overestimate the likelihood of hypotension because it does not know these interventions at the time of prediction and therefore cannot and should not take them into account. In addition, these concerns were alleviated by the fact that our model could directly output blood pressure values.

Our study has several strengths and limitations. First, TFT is a state-of-the-art, novel algorithm that can utilise data on administered medication, a factor plausibly related to the occurrence of hypotension. We assembled a large and diverse patient cohort and had surgical cases from many specialties. We adhered to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines[30] for the development and validation of predictive models and performed internal and external validation. However, the present retrospective study used data recorded for routine patient care, which likely introduced errors in our data set. Some data highly relevant for changes in blood pressure were not captured in our data set, such as bleeding, surgical compression of blood vessels or incorrectly documented medication regarding timing of data entry. Similarly, VitalDB lacks information on bolus drugs, which affects the performance of the models in the external validation. A 15-s sampling interval was employed to benefit from higher resolution data and accurately time the effects of medication. However, this may result in inaccuracies due to the mismatch with standard 3-min blood pressure measurements. Although this approach offers increased detail, the use of forward filled values between actual measurements may impact the performance of the model and introduce noise into the learning process. In addition, medical interventions such as administration of vasopressors in response to hypotension were captured in our data set, which may have biased the TFT model towards an expectation of these interventions. Due to the extensive training time requirements, cross-validation was not employed to train the TFT model, which may have an impact on the final results' accuracy. While the TFT model performs well in continuous prediction tasks, the XGB model demonstrated superior results in binary predictions during internal validation, highlighting the importance of selecting the appropriate model for specific needs.

In summary, we applied the novel TFT algorithm to predict intraoperative blood pressure trajectories for the upcoming 7 min. Our model used easily obtainable input data available during routine care—most importantly, data on intraoperatively administered medications—and only required low-resolution data, which can be obtained without the placement of an arterial line. We obtained a low predictive error for continuous blood pressure predictions and—regarding the binary prediction of hypotension—and excellent discrimination with reasonable calibration. Future studies should investigate how our prediction model could be integrated into the anaesthesiologist's workflow and how this would affect patient outcomes.

Contributors

L.K., C.D., N.J., C.H., and O.K. were responsible for the conceptualization of the paper. L.K., C.D., and S.B.1 (Stefan Bartos) accessed and verified the data. The investigation was conducted by L.K., C.D., N.J., S.B.1, S.B.2 (Sybille Behrens), A.B., C.H., and O.K. Methodology was developed by L.K., C.D., N.J., C.H., and O.K. Software was developed by L.K. and N.J. Visualization was done by L.K. The original draft was written by L.K., C.D., N.J., S.B.2, and A.B. Supervision was provided by C.H. and O.K., who also reviewed and edited the paper. All authors agree to be fully accountable for ensuring the integrity and accuracy of the work and have read and approved the final manuscript. The corresponding author had full access to all data in the study and assumed final responsibility for the decision to submit the manuscript for publication.

Data sharing statement

The external database, which was utilised for validation purposes, is openly available, enabling replication of the validation process.[21] The code for model training and evaluation is available (https://github.com/lorenzkap/MAP_TFT).

Declaration of interests

We declare no competing interests.

Acknowledgements

The computational results presented have been achieved using the Vienna Scientific Cluster (VSC).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102797.

References

1 Jor O, Maca J, Koutna J, et al. Hypotension after induction of general anesthesia: occurrence, risk factors, and therapy. A prospective multicentre observational study. *J Anesth*. 2018;32:673–680.

2 Wesselink EM, Kappen TH, Torn HM, Slooter AJC, van Klei WA. Intraoperative hypotension and the risk of postoperative adverse outcomes: a systematic review. *Br J Anaesth*. 2018;121:706–721.

3 Sessler DI, Bloomstone JA, Aronson S, et al. Perioperative Quality Initiative consensus statement on intraoperative blood pressure, risk and outcomes for elective surgery. *Br J Anaesth*. 2019;122:563–574.

4 Nadim MK, Forni LG, Bihorac A, et al. Cardiac and vascular surgery-associated acute kidney injury: the 20th international consensus conference of the ADQI (acute disease quality initiative) group. *J Am Heart Assoc*. 2018;7:e008834.

5 Wachtendorf LJ, Azimaraghi O, Santer P, et al. Association between intraoperative arterial hypotension and postoperative delirium after noncardiac surgery: a retrospective multicenter cohort stud y. *Anesth Analg*. 2022;134:822–833.

6 Maleczek M, Laxar D, Geroldinger A, Kimberger O. Intraoperative hypotension is associated with postoperative nausea and vomiting in the PACU: a retrospective database analysis. *J Clin Med*. 2023;12:2009.

7 Kang AR, Lee J, Jung W, et al. Development of a prediction model for hypotension after induction of anesthesia using machine learning. *PLoS One*. 2020;15:e0231172.

8 Kendale S, Kulkarni P, Rosenberg AD, Wang J. Supervised machine-learning predictive analytics for prediction of post-induction hypotension. *Anesthesiology*. 2018;129:675–688.

9 Solomon SC, Saxena RC, Neradilek MB, et al. Forecasting a crisis: machine-learning models predict occurrence of intraoperative bradycardia associated with hypotension. *Anesth Analg*. 2020;130:1201–1210.

10 Hatib F, Zhongping J, Sai B, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology*. 2018;129:663–674.

11 Davies SJ, Vistisen ST, Jian Z, Hatib F, Scheeren TWL. Ability of an arterial waveform analysis-derived hypotension prediction index to predict future hypotensive events in surgical patients. *Anesth Analg*. 2020;130:352–359.

12 Lee S, Lee H-C, Chu Y, et al. Deep learning models for the prediction of intraoperative hypotension. *Br J Anaesth*. 2021;126:808–817.

13 Kang MW, Kim S, Kim YC, et al. Machine learning model to predict hypotension after starting continuous renal replacement therapy. *Sci Rep*. 2021;11:17169.

14 Maheshwari K, Shimada T, Fang J, et al. Hypotension Prediction Index software for management of hypotension during moderate-to high-risk noncardiac surgery: protocol for a randomized trial. *Trials*. 2019;20:255.

15 Vistisen ST, Enevoldsen J. CON: the hypotension prediction index is not a validated predictor of hypotension. *Eur J Anaesthesiol*. 2024;41:118–121.

16 Lim B, Arık SÖ, Loeff N, Pfister T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int J Forecast*. 2021. https://doi.org/10.1016/j.ijforecast.2021.03.012.

17 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Informat Process Syst*. 2017.

18 Gu A, Gulcehre C, Paine T, Hoffman M, Pascanu R. *Improving the gating mechanism of recurrent neural networks*. 2020.

19 Li J, Yan XS, Chaudhary D, et al. Imputation of missing values for electronic health record laboratory data. *NPJ Digital Med*. 2021;4:147.

20 GitHub - greatwhiz/tft_tf2: temporal fusion transformers for tensorflow 2.x. https://github.com/greatwhiz/tft_tf2.

21 Lee H-C, Park Y, Yoon SB, et al. VitalDB, a high-fidelity multiparameter vital signs database in surgical patients. *Sci Data*. 2022;9:279.

22 Li Z. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. *Comput Environ Urban Syst*. 2022;96:101845.

23 Jo Y-Y, Jang J-W, Kwon J-M, et al. Predicting intraoperative hypotension using deep learning with waveforms of arterial blood pressure, electroencephalogram, and electrocardiogram: retrospective study. *PLoS One*. 2022;17:e0272055.

24 Enevoldsen J, Vistisen ST. Performance of the hypotension prediction index may be overestimated due to selection bias. *Anesthesiology*. 2022;137:283–289.

25 Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Can J Cardiol*. 2022;38:204–213.

26 Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat Biomed Eng*. 2022;6:1330–1345.

27 Fernandes MPB, Armengol de la Hoz M, Rangasamy V, Subramaniam B. Machine learning models with preoperative risk factors and intraoperative hypotension parameters predict mortality after cardiac surgery. *J Cardiothorac Vasc Anesth*. 2021;35:857–865.

28 Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230.

29 Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA*. 2017;318:1377–1384.

30 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.