Clinical and Experimental Emergency Medicine

Clin Exp Emerg Med 2022;9(4):304-313
https://doi.org/10.15441/ceem.22.335

Original Article

eISSN: 2383-4625

Machine learning for the prediction of preclinical airway management in injured patients: a registry-based trial

André Luckscheiter[1], Wolfgang Zink[1], Torsten Lohs[2], Johanna Eisenberger[2], Manfred Thiel[3], Tim Viergutz[4]

[1]Department of Anesthesiology, Intensive Care and Emergency Medicine, Ludwigshafen Municipal Hospital, Ludwigshafen, Germany
[2]Center for Quality Management in Emergency Medical Services Baden-Wuerttemberg (SQR-BW), Stuttgart, Germany
[3]Department of Anesthesiology and Intensive Care Medicine, University Medical Center Mannheim, Mannheim, Germany
[4]Clinic for Anesthesia, Intensive Care and Pain Therapy, BG Trauma Center Tuebingen, Tuebingen, Germany

Objective The aim of this study was to determine the feasibility of using machine learning to establish the need for preclinical airway management for injured patients based on a standardized emergency dataset.

Methods A registry-based, retrospective analysis was conducted of adult trauma patients who were treated by physician-staffed emergency medical services in southwestern Germany between 2018 and 2020. The primary outcome was to assess the feasibility of using the random forest (RF) and Naive Bayes (NB) machine learning algorithms to predict the need for preclinical airway management. The secondary outcome was to use a principal component analysis to determine the attributes that can be used and advanced for future model development.

Results In total, 25,556 adults with multiple injuries were identified, including 1,451 patients (5.7%) who required airway management. Key attributes were auscultation, injury pattern, oxygen therapy, thoracic drainage, noninvasive ventilation, catecholamines, pelvic sling, colloid infusion, initial vital signs, preemergency status, and shock index. The area under the receiver operating characteristics curve was between 0.96 (RF; 95% confidence interval [CI], 0.96–0.97) and 0.93 (NB; 95% CI, 0.92–0.93; $P<0.01$). For the prediction of airway management, RF yielded a higher precision-recall area than NB (0.83 [95% CI, 0.8–0.85] vs. 0.66 [95% CI, 0.61–0.72], respectively; $P<0.01$).

Conclusion To predict the need for preclinical airway management in injured patients, attributes that are commonly recorded in standardized datasets can be used with machine learning. In future models, the RF algorithm could be used because it has robust prediction accuracy.

Keywords Intratracheal intubation; Machine learning; Bayes theorem; Wounds and injuries; Decision trees

Received: 16 June 2022
Revised: 11 September 2022
Accepted: 16 October 2022

Correspondence to: André Luckscheiter
Department of Anesthesiology, Intensive Care and Emergency Medicine, Ludwigshafen Municipal Hospital, Bremserstrasse 79, Ludwigshafen 67063, Germany
E-mail: luckscha@klilu.de

How to cite this article:
Luckscheiter A, Zink W, Lohs T, Eisenberger J, Thiel M, Viergutz T. Machine learning for the prediction of preclinical airway management in injured patients: a registry-based trial. Clin Exp Emerg Med 2022;9(4):304-313. https://doi.org/10.15441/ceem.22.335

Capsule Summary

What is already known

Preclinical airway management is a high risk procedure. Other than a Glascow Coma Scale of less than 9 or acute respiratory insufficiency, there are few methods to predict the need for preclinical airway management.

What is new in the current study

We developed and validated a machine learning model to predict the need for airway management in injured patients.

INTRODUCTION

International guidelines recommend preclinical airway management as a potential life-saving procedure for severely injured patients with traumatic brain injury and a Glasgow Coma Scale (GCS) <9; severe respiratory insufficiency, for example, due to thoracic trauma or airway injuries; or trauma-associated shock.[1-3] However, preclinical airway management is a high-risk procedure due to imminent hypoxia, challenging environmental conditions, and varying clinician experience in managing difficult airway situations.[4,5] Because hemodynamic conditions and the patient's state of awareness can change quickly, preclinical trauma care is a highly dynamic situation. Therefore, an ability to predict or exclude the need for airway management would assist decision-making.

In recent years, several machine learning models that can predict the need for endotracheal intubation in intensive care patients have been published. They are based on electronic medical record systems and common clinical hemodynamic and laboratory parameters.[6-9] In preclinical trauma medicine, no such model exists.

German emergency medical services are divided into paramedic and emergency physician systems (grounded or air), which are alarmed by the rescue coordination center in parallel or sequentially depending on the emergency. Certain medical interventions, such as drug therapy or airway management, are restricted by law to emergency physicians except when needed for resuscitation or when an emergency physician is unavailable. German emergency physicians recruit themselves mainly from fields such as anesthesiology, internal medicine, and surgery. The specialization can be achieved in parallel with main medical specialist training after two years of clinical practice, which must contain at least a 6-month rotation in the accident and emergency department or intensive care unit.[5,10] For quality improvement, the German state of Baden-Wuerttemberg (population, 11.1 million in 2020; area, 35,751 km^2; capital, Stuttgart) created a Center for Quality Management in Emergency Medical Services in 2011. Since then, all paramedics and preclinical emergency physicians have had to provide anonymous, digital documentation to the minimal emergency dataset (MIND).[10,11] The MIND has the advantage of being used throughout Germany, and it also contains international standardized examination findings, diagnoses, and interventions that are used in the German Trauma Registry and the German Resuscitation Registry. Divided into subcategories according to the Advanced Trauma Life Support (ABCDE) algorithm at first contact and hospital admission and supplemented by a free text anamnesis and history (including vital signs diagram) of pharmaceutical therapy and medical interventions, the MIND provides nationwide, standardized, emergency documentation. Although the free text and history sections are not available digitally, the MIND seems suitable for research with machine learning.

Therefore, the aim of this study was to evaluate the feasibility of building machine learning models to predict the need for preclinical airway management in trauma patients. As a first step, attributes of the MIND that define patients who need preclinical airway management were identified. Second, two machine learning algorithms were tested to demonstrate the accuracy of the models.

METHODS

Ethical statements

This study is reported based on the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement.[12] The trial was approved by the Ethics Committee of the State Medical Association of Rhineland-Palatinate (No. 2021-15767-retrospektiv). The study is a retrospective registry analysis with anonymized data. Informed consent was waived due to the retrospective nature of the study.

Design and setting

Adult patients with multiple injuries who were primarily treated by a physician-staffed ground or air ambulance from 2018 to 2020 were selected from the MIND. Dead patients and those requiring resuscitation were excluded. Briefly, the MIND files of the remaining patients were preprocessed for attribute selection using medical causality and a principal component analysis (PCA). With the help of the resulting attributes, Naive Bayes (NB) and random forest (RF) models were trained and tested to find their accuracy in predicting whether those injured patients were given preclinical airway management. Patient selection, dataset creation, and the analyses are illustrated in Fig. 1.

Definition

The MIND does not yet contain anesthesia as an attribute. Therefore, emergency general anesthesia in any injured patient was defined as documentation of invasive airway management, positive end tidal CO_2 without noninvasive ventilation (NIV) at admission, documented invasive ventilation at admission and the use of a muscle relaxant, or any use of a muscle relaxant. The main assumption was the correct indication of preclinical emergency anesthesia.

Attribute selection and data preprocessing

The MIND includes more than 550 anonymized attributes, including specialization of the physician, standardized clinical examination findings, medical diagnoses, injury patterns in relation to particular body parts (classified as none, mild, moderate, severe, or deadly by the attending physicians), blunt or penetrating trauma, and vital signs at first contact and hospital admission, including the GCS, heart rate, systolic blood pressure, respiratory rate, oxygen saturation, end tidal CO_2, temperature, blood glucose level, and pain level. Furthermore, electrocardiogram findings (at first contact and hospital admission), medication (without dosage or timing), treatment (NIV, invasive airway management, thoracic drainage, pelvic sling), infusion therapy (crystalloid/colloid infusion, blood products), age, preemergency status (PES; a preclinically adapted classification of the American Society of Anesthesiologists), time on site, and transport time are recorded in the dataset.[13]

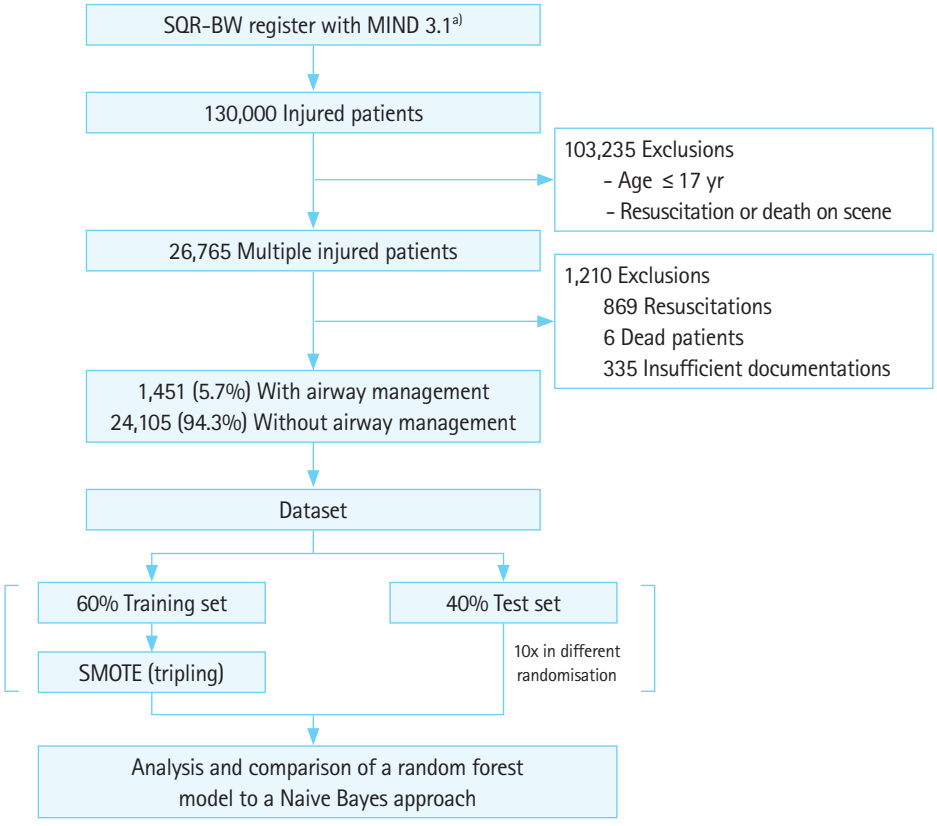


Fig. 1. Flowchart for patient selection, dataset creation, and analysis. SQR-BW, Center for Quality Management in Emergency Medical Services Baden-Wuerttemberg; MIND, minimal emergency dataset; SMOTE, synthetic minority oversampling method. a)A total of 24 attributes included: >550 attributes filtered by causality or potential correlation, then selected by principal component analysis (Wrapper).

The datasets of patients with cardiac arrest were excluded because abstaining from resuscitation could bias the weighting of certain attributes. Only datasets with at least two of the following three attributes, initial GCS, systolic blood pressure, and oxygen saturation, were included because those parameters represent the guidelines' recommendations.[1-3]

In data preprocessing, generally accepted attributes in the training set with potential correlations but no medical causality were excluded from the machine learning analysis (e.g., place of accident), as were causal attributes without any frequent occurrence in one of the two classes. Attributes correlating with indications for airway management were identified using international guidelines about respiratory, neurological, or hemodynamic findings and injury patterns.[2,3,14] However, because critical volume loss and (developing) shock are not directly recorded in the MIND, surrogate parameters such as pelvic sling or tranexamic acid were also included.

The imputation of missing data was not considered due to the nominal character of most attributes. Because the remaining attributes all contributed with different weightings, a PCA was performed on the whole dataset using the wrapper method with a bidirectional search and a C4 decision tree (J48) with tenfold cross-validation (settings in Supplementary Table 2).[15] The Java-based software Weka ver. 3.8.4 (University of Waikato, Hamilton, New Zealand) was used for the PCA and machine learning.[16,17] Statistical comparison of the attributes between the two classes (airway management and no airway management) was performed with chi-square test, U-test, or t-test, as appropriate, in Microsoft Excel (Microsoft Corp., Redmond, WA, USA). A P-value of less than 0.05 was defined as significant. Continuous variables are expressed as means and standard deviations, and categorical variables are expressed as percentages.

Class balancing, training, and testing

The data were split into a 60% training set and 40% test set 10 times with a randomized split procedure to define the performance of the algorithms with different frequencies of invasively ventilated patients. In general, machine learning algorithms tend to learn and predict the majority class, whereas most studies are interested in the minority class. To handle that class imbalance problem for the minority class that received airway management, the synthetic minority oversampling method (SMOTE) algorithm was used to triple the airway management class in the training sets, but not in the test sets. SMOTE synthesis creates one new minority instance out of k = 5 existing minority instances using the k-nearest neighbor approach (Supplementary Table 3).[18] This procedure was chosen because Weka does not offer a cross-validation that uses SMOTE in training but not in testing. Tripling the minority class was an appropriate assessment to improve the predictions and prevent overfitting. For supervised machine learning, the NB and RF methods were chosen (Supplementary Table 4). Both algorithms can handle missing values.

Model performance

All results are presented as means with 95% confidence intervals (CIs). As performance criteria, overall correctness, kappa value, the area under the receiver operator curve (AUC-ROC), sensitivity (need for airway management), specificity (no need for airway management), positive predictive value (PPV) and negative predictive value (NPV), and the precision-recall (PRC) area were chosen.[15] The Matthews correlation coefficient (MCC) was used to measure the quality of the two presented classes of very different sizes (range: −1, total disagreement; 0, random prediction; +1, perfect prediction).[19] The cost-benefit calculation for the RF algorithm was performed automatically for the lowest overall error rate. The performance across all 10 test sets was averaged and compared with a t-test ($P < 0.05$ as significant, calculated in Microsoft Excel).

RESULTS

Out of more than 130,000 injured patients, 26,765 patients with multiple injuries were selected. Of the selections, 869 resuscitations, 6 fatal cases, and 335 insufficiently documented datasets were then excluded, leaving 25,556 datasets with 1,451 cases (5.67%) of airway management.

Data preprocessing identified 31 attributes with potential correlation or medical causality. In the PCA, 24 attributes were selected, among them auscultation, injury pattern without the upper limbs or soft parts, oxygen therapy, NIV, tranexamic acid and catecholamines, pelvic sling, vital signs, PES, and shock index. With the exception of initial systolic blood pressure and respiratory rate ($P > 0.05$), the groups with and without airway management differed significantly (Table 1). For further information about nonselected attributes see Supplementary Table 1.

In overall correctness, the RF outperformed the NB (97.8 [95% CI, 97.57–98.03] vs. 93.55 [95% CI, 93.11–93.99], respectively; $P < 0.01$). The RF reached a significantly higher kappa value (0.78 [95% CI, 0.75–0.8]) than the NB (0.54 [95% CI, 0.52–0.56]; $P < 0.01$). In the AUC-ROC analysis, the RF reached 0.96 (95% CI, 0.96–0.97), and the NB reached 0.93 (95% CI, 0.92–0.93; $P < 0.01$) (Fig. 2A). Furthermore, the RF model had a significantly higher MCC than the NB approach (0.78 [95% CI, 0.76–0.8] vs. 0.56 [95% CI, 0.54–0.57], respectively; $P < 0.01$).

Table 1. Clinical findings and medical treatments for both classes with the attributes selected through the principal component analysis

Attribute	Airway management		P-value
	Yes (n = 1,451)	No (n = 24,105)	
Auscultation 1			< 0.01[a]
Obstruction/gasping/apnea	15.0	0.3	
Bronchial spasm	18.0	0.3	
Rhonchi	2.0	0.2	
Other	31.0	13.0	
Auscultation 2			
Dyspnea ± cyanosis	37.0	3.0	< 0.01[a]
Head injury			
None	36.0	65.0	< 0.01[a]
Mild	5.0	20.0	< 0.01[a]
Moderate	15.0	13.0	0.05[a]
Severe	44.0	2.0	< 0.01[a]
Face injury			< 0.01[a]
None	77.0	83.0	
Mild	5.0	10.0	
Moderate	11.0	7.0	
Severe	7.0	0.6	
Cervical spine injury			
None	90.0	88.0	0.07
Mild	2.0	6.0	< 0.01[a]
Moderate	4.0	5.0	0.40
Severe	4.0	0.7	< 0.01
Thoracic/lumbar spine injury			< 0.01[a]
None	90.0	85.0	
Mild	1.0	5.5	
Moderate	4.0	8.0	
Severe	4.0	1.0	
Thoracic injury			
None	68.0	77.0	< 0.01[a]
Mild	3.0	9.0	< 0.01[a]
Moderate	10.0	12.0	0.10
Severe	19.0	2.0	< 0.01[a]
Abdominal injury			
None	85.0	92.0	< 0.01[a]
Mild	1.0	2.0	< 0.01[a]
Moderate	4.0	4.0	0.50
Severe	10.0	1.0	< 0.01[a]
Pelvic injury			
None	83.0	87.0	< 0.01[a]
Mild	2.0	5.0	< 0.01[a]
Moderate	4.0	6.0	0.02[a]
Severe	1.0	2.0	< 0.01[a]
Lower limb injury			< 0.01
None	76.0	72.0	
Mild	4.0	12.0	
Moderate	7.0	12.0	
Severe	13.0	3.0	
Oxygen therapy	57.0	35.0	< 0.01[a]
Noninvasive ventilation	32.0	0.2	< 0.01[a]
Thoracic drainage	14.0	0.2	< 0.01[a]
Colloid infusion	7.0	0.2	< 0.01[a]
Tranexamic acid	40.0	4.0	< 0.01[a]
Pelvic sling	27.0	4.0	< 0.01[a]
Catecholamine	43.0	1.0	< 0.01[a]
Systolic blood pressure (mmHg)	137 ± 29	138 ± 28	0.29
Oxygen saturation (%)	94 ± 7	95 ± 6	< 0.01[a]
Heart rate (beats/min)	90 ± 20	89 ± 19	< 0.01[a]
Respiratory rate (breaths/min)	16 ± 5	16 ± 5	0.24
Pain level (0–10)[b]	5 (0–10)	5 (0–10)	< 0.01[a]
Shock index	0.7 ± 0.3	0.6 ± 0.6	0.03[a]
Preemergency status (1–4)[c]	2 (1–3)	2 (1–3)	< 0.01[a]
Glasgow Coma Scale (3–15)	15 (14–15)	15 (15–15)	< 0.01[a]
Age (yr)[d]	54.88 ± 21.44	55.80 ± 22.28	0.13
Male sex[d]	72.0	60.0	< 0.01[a]

Values are presented as percentage, mean ± standard deviation, or median (interquartile range).
[a]Statistically significant value (P < 0.05). [b]No pain, 0. [c]Healthy, 1; moribund, 4. [d]Baseline characteristics not used in the algorithm.

In predicting the use of airway management, the difference between the NB and RF results was not statistically significant (0.75 [95% CI, 0.73–0.76] vs. 0.73 [95% CI, 0.71–0.76], respectively; P = 0.38). The best PPV was gained with the RF (0.85 [95% CI, 0.84–0.87]; NB, 0.46 [95% CI, 0.44–0.49]; P < 0.01). This also resulted in a larger PRC area for the RF (0.83 [95% CI, 0.80–0.85]; NB, 0.66 [95% CI, 0.61–0.72]; P < 0.01) (Fig. 2B).

Both algorithms yielded a very high specificity (RF, 0.993 [95% CI, 0.992–0.994] vs. NB, 0.947 [95% CI, 0.942–0.952]; P < 0.01), a high NPV (RF, 0.984 [95% CI, 0.980–0.987] vs. NB, 0.984 [95% CI, 0.983–0.985]; P = 0.85), and a high PRC area (RF, 0.996 [95% CI, 0.996–0.997] vs. NB, 0.992 [95% CI, 0.992–0.993]; P < 0.01) (Table 2).

The average threshold of the RF model was 0.51 (95% CI, 0.49–0.53). Due to the decision process used by the NB, no average threshold can be given for it. The three most important attributes in the RF were systolic blood pressure (0.306 ± 0.019), head injury (0.305 ± 0.013), and initial heart rate (0.294 ± 0.018) (Fig. 3).

DISCUSSION

This study set out to develop a decision model for determining the necessity of preclinical airway management in adult trauma patients. Commonly recorded preclinical attributes such as injury pattern, certain examination findings, vital signs, and emergency medical interventions were found to be most influential in forecasting the need for preclinical airway management. Both models

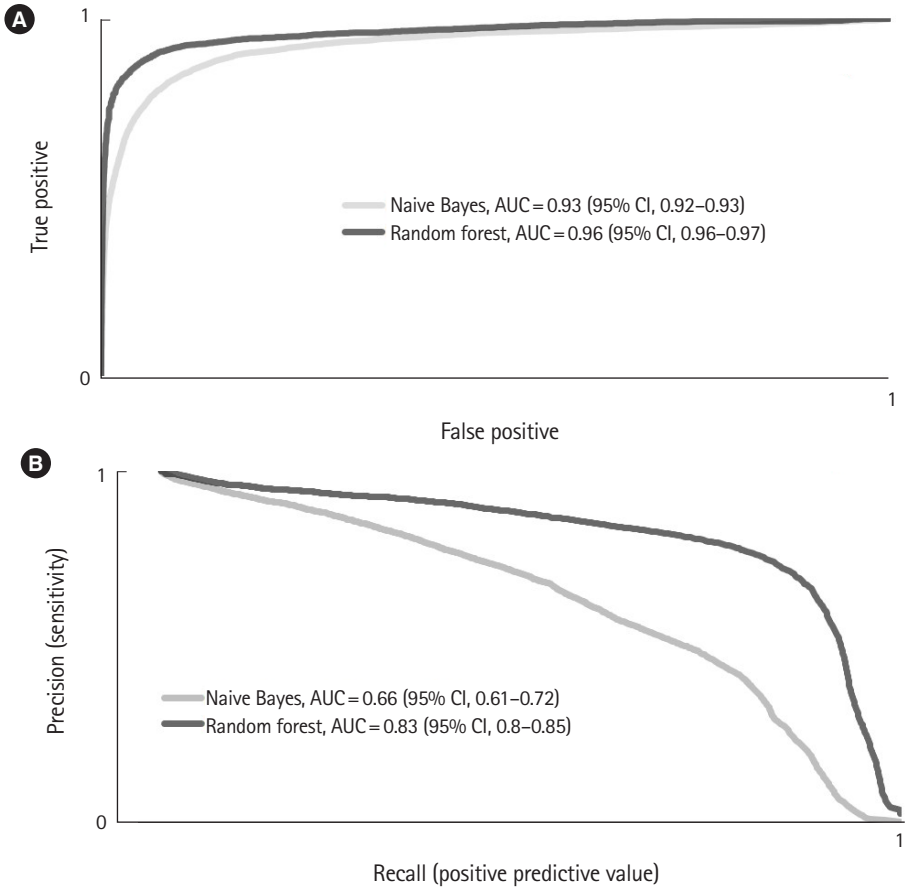


Fig. 2. Averaged (A) receiver operator curves for the overall performance and (B) precision-recall curves for the prediction of airway management by the Naive Bayes and random forest algorithms. AUC, area under the curve; CI, confidence interval.

Table 2. Model performance and evaluation of random forest versus Naive Bayes

Variable	Random forest	Naive Bayes	P-value
Overall correctness (%)	97.80 ± 0.37 (97.57–98.03)	93.55 ± 0.71 (93.11–93.99)	< 0.01[a]
Kappa	0.78 ± 0.04 (0.75–0.80)	0.54 ± 0.03 (0.52–0.56)	< 0.01[a]
AUC-ROC	0.96 ± 0.01 (0.96–0.97)	0.93 ± 0 (0.92–0.93)	< 0.01[a]
MCC	0.78 ± 0.04 (0.76–0.80)	0.56 ± 0.02 (0.54–0.57)	< 0.01[a]
Sensitivity	0.73 ± 0.05 (0.71–0.76)	0.75 ± 0.02 (0.73–0.76)	0.38
Positive predictive value	0.85 ± 0.03 (0.84–0.87)	0.46 ± 0.03 (0.44–0.49)	< 0.01[a]
PRC area[b]	0.83 ± 0.04 (0.80–0.85)	0.66 ± 0.09 (0.61–0.72)	< 0.01[a]
Specificity	0.993 ± 0.002 (0.992–0.994)	0.947 ± 0.008 (0.942–0.952)	< 0.01[a]
Negative predictive value	0.984 ± 0.006 (0.980–0.987)	0.984 ± 0.001 (0.983–0.985)	0.85
PRC area[b]	0.996 ± 0.001 (0.996–0.997)	0.992 ± 0.001 (0.992–0.993)	< 0.01[a]

Values are presented as standard deviation (95% confidence interval).
AUC-ROC, area under the receiver operator curve; MCC, Matthews correlation coefficient; PRC, precision-recall.
[a]Statistically significant value ($P < 0.05$). [b]Given for the prediction and exclusion of airway management.

developed here showed excellent results in excluding the need for airway management, but only the RF model had satisfactory accuracy in predicting it. Therefore, the feasibility of using machine learning to predict the need for airway management in preclinical trauma patients has been confirmed, but the models need to be advanced. Nonetheless, even before a final model can be implemented in the electronic medical records, the attributes determined here can already be used clinically to alert emergency physicians about trauma patients at increased risk of requiring airway management. For example, the absence of severe head or

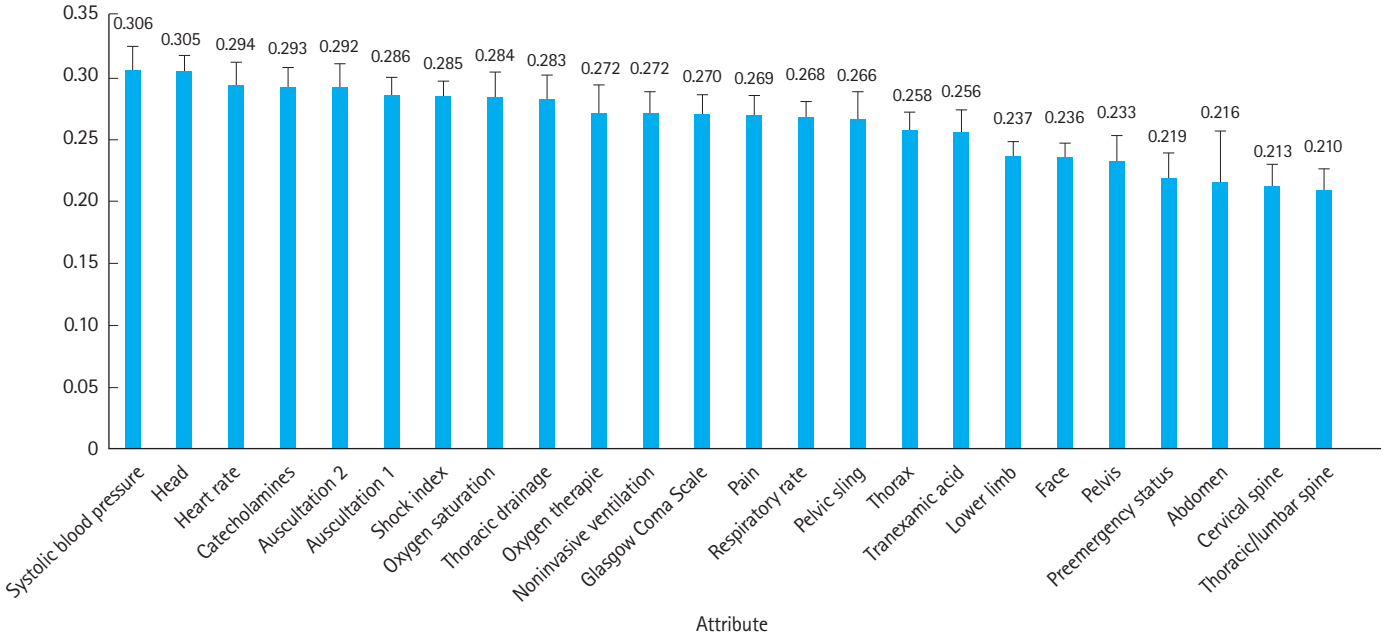


Fig. 3. Attribute weighting in the random forest model, given as means with standard deviation error bars.

thoracic injury, catecholamine therapy, thoracic drainage, or NIV could justify a later evaluation of airway protection. To the best of our knowledge, this analysis is the first to use machine learning to forecast airway management in a preclinical environment. However, several factors need to be considered to interpret and advance the results.

Database, attribute selection, and model comparison

The more distinct the pathological findings in the initial parameters, the better the classification by the algorithms could be. However, differences in attributes such as GCS or oxygen saturation were marginal, and their averages were physiological, which was partly reported in other clinical modeling studies.[8,20,21] This could be explained by belated documentation of paramedically stabilized vital signs.

Attribute choice is always a compromise between overgeneralization (selecting only attributes with strong correlation or causality) and overfitting (selecting many attributes, even those with weak correlation). The PCA in this study filtered in attributes with strong indirect correlations with airway management. For example, the use of catecholamines can be interpreted as a surrogate for hemodynamic instability before or after airway management in emergency anesthesia. Other surrogates were colloid infusion, pelvic sling, and tranexamic acid for potential blood loss (attribute tourniquet not included in MIND). NIV can be discussed as a surrogate for respiratory failure or a method of preoxygenation. Although the shock index is only to some extent reliable for the diagnosis of shock, it had weight in combination with other attributes.[22,23] Because preclinical emergency physicians in Germany usually lack point-of-care and radiographic findings, they have to use a less-reliable clinical examination with baseline vital signs for their time-critical decision-making. The surrogate parameters used in this study can therefore be seen as a replacement for real-time vital signs. They also reflect to some extent the recommendations for airway management in patients with traumatic respiratory disorder, brain injury, and shock.[1-3] Future prediction models in preclinical airway management should combine attributes emphasized in the guidelines with selected surrogates that reflect the dynamics of preclinical emergency medicine to compensate for any lack of real-time parameters.

Compared with other studies, a main distinction of this study is the restriction to initial vital signs and adaptation to preclinical conditions.[2,3] Siu et al.[20] used an additional blood gas analysis with sequential organ failure assessments at multiple time points for their RF model to predict the need for intubation in the first 24 hours after a critical care admission (sensitivity, 0.88; specificity, 0.66; AUC-ROC, 0.86; PPV, 0.73; NPV, 0.85). Arvind et al.[6] indicated a AUC-ROC of 0.84 and PRC area of 0.3 for their RF model for predicting mechanical ventilation in COVID-19 patients based on vital signs and a blood gas analysis. In neonatal intensive care, Clark et al.[8] demonstrated a boosted logistic regression model with an AUC-ROC of 0.84. Politano et al.[21] could predict urgent intubation in a trauma intensive care unit with an AUC-ROC of 0.770 to 0.865 with the help of a boosted logistic regression us-

ing multiple sampling windows for vital signs along with age, oxygen partial pressure, and days since extubation.

Model performance

With regard to the performance of both algorithms, several factors about their basic method of calculation and the prevalence of airway management must be considered. In this study, the ROC curve alone overestimates the model performance because of the class imbalance problem (94% without emergency anesthesia) and the very high specificities and negative predictive values. Therefore, the goodness of class prediction can best be evaluated by the PRC area, which showed that the RF had a robust predication accuracy.[24]

The basic assumption of the NB is the independence of all attributes without any correlation. Such a level of independence is almost never found in real-world data. In this study, the auscultation findings, respiratory rate, and oxygen saturation all influence one another, as do the GCS score and face and/or head injury. The decision process in favor of or against a class is performed by comparing the summed probability of the test case to the summed probability of the class, which leads to the shown bad calibration. The advantage of an NB approach is its fast calculation and simple implementation. Also, the arithmetic means and variance are parameterized independently of all other variables.[15]

Unlike in the NB, independence is not a basic assumption of an RF. Decision trees have the advantage of using the same attributes on different levels in different dependencies. In contrast to a single decision tree model, an RF uses the bagging procedure, by which multiple random trees each calculate a prediction. Those are then averaged to reach a final decision. This explains not only why RF got better outcomes than NB but also the weights of certain attributes whose differences were marginal. Those same effects also appear in the PCA, because it also uses a decision tree model. Therefore, the RF is robust to outliers, works well with nonlinear data, and has a lower risk of overfitting than single decision trees. As a result, the RF could handle even the relatively small prevalence of airway management cases in the test sets, achieved a good PRC area, and had a robust performance.[15,25] Given the prevalence between the different test sets, the RFs differ, and a final model cannot be given.

Further limitations

Due to the former and following limitations, this study represents only a first attempt to build a sustainable, general model for predicting preclinical airway management. Overreliance on machine learning in high-risk situations can result in potential patient hazards. Future models are also needed for internal and neurological patients. These results were developed in a physician-staffed emergency medical system and therefore cannot be simply transferred to paramedic systems.[26] The weighting of certain attributes could be changed by alterations in clinical practice. The timing of interventions is missing from MIND, which limits the applicability of the models presented here. Unlike previous prediction models for resuscitation, attributes such as trauma site were not included in the data used here. Whereas in resuscitation, the site of cardiac arrest is directly linked to bystander cardio-pulmonary resuscitation, there is no such correlation for trauma site or mechanism and airway management, only for trauma severity.[3,27] Unfortunately, that severity can only be assessed by the primary physical exam and not by later radiographic findings and hospital data. Although this study used data from a statewide emergency medical service, no independent external test set from another German region was used here. Therefore, predications of stability with regard to noise and overfitting must be restrained. Unlike in other studies, the imputation of missing values in this study was not reasonable, mainly due to static nominal, binary, or ordinal attributes.[6,20] Whether emergency physicians postponed endotracheal intubation because of a potentially difficult airway or a lack of experience cannot be stated because no further clinical records were available.[5] Also, the correct indication for airway management and primary assessment according to the ABCDE algorithm could not be checked in every single case due to the retrospective design and dataset structure. In machine learning, unsupervised deep learning neural networks have recently outperformed supervised approaches such as the RF. However, those deep learning models require a large amount of data and computing power. Network creation is complex, unstandardized, and time-consuming. Because this study focused on a simple binary problem, and the data structure was inconsistent, RF and NB were chosen. The supplementary data contain a first approach to a deep learning neural network, but it performed worse than the RF in predicting the need for airway management (Supplementary Table 5 and Supplementary Fig. 1). Nonetheless, a deep learning application might be suitable for future models, especially with real-time attributes.[15]

CONCLUSION

In conclusion, this study has shown the feasibility of using a machine learning model to predict the need for airway management in injured patients. The RF model combined a satisfactory prediction performance with an excellent ability to exclude the need for airway management in trauma patients. Because the many attributes available can be a hindrance in quickly assessing trauma patients, models such as those presented here could already

be used as surveillance tools in the background or to send the intubation probability to the hospital, where additional resources could be activated. Embedded in a continuous electronic medical record and expanded by data about internal patients, real-time parameters and point-of-care tests, an RF-based prediction model could be made more reliable and support preclinical decision-making or quality management. In the future, patients at risk could be identified at an early time with the help of such a machine learning model.

SUPPLEMENTARY MATERIAL

Supplementary Table 1. All recorded attributes and their values together with the class comparison and reason for exclusion
Supplementary Table 2. Settings of the principal component analysis in Weka
Supplementary Table 3. Settings of the SMOTE algorithm in Weka
Supplementary Table 4. Settings of the random forest and Naive Bayes model in Weka
Supplementary Table 5. Performance of two deep learning networks before and after attribute selection
Supplementary Fig. 1. Averaged receiver operator curves (ROC) for (A) the overall performance and (B) the averaged precision-recall (PRC) curves for the prediction of airway management of the Naive Bayes, the random forest algorithm, and the deep learning neural network (one dense layer with six neurons) after attribute selection.
Supplementary materials are available at https://doi.org/10.15441/ceem.22.335. Further supplementary data, including single random forest models, are available upon reasonable request via e-mail. Due to data protection, the datasets cannot be published, but research with the database is possible upon request to the Center for Quality Management in Emergency Medical Services Baden-Wuerttemberg (SQR-BW).

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

FUNDING

This work was supported by the Department of Anesthesiology, Operative Intensive Care Medicine and Emergency Medicine, Ludwigshafen Municipal Hospital.

AUTHOR CONTRIBUTIONS

Conceptualization: AL; Data curation: AL, TL, JE; Formal analysis: AL; Funding acquisition: WZ; Investigation: AL; Methodology: AL; Project administration: TV; Resources: TL, JE; Software: AL; Supervision: WZ, MT, TV; Validation: AL; Visualization: AL; Writing–original draft: AL; Writing–review & editing: WZ, TL, JE, MT, TV.
All authors read and approved the final manuscript.

ORCID

André Luckscheiter https://orcid.org/0000-0002-5724-7130
Wolfgang Zink https://orcid.org/0000-0002-4224-8384
Torsten Lohs https://orcid.org/0000-0003-3476-0430
Johanna Eisenberger Not available
Manfred Thiel https://orcid.org/0000-0002-6267-4380
Tim Viergutz https://orcid.org/0000-0001-8104-1047

REFERENCES

1. Crewdson K, Rehn M, Lockey D. Airway management in prehospital critical care: a review of the evidence for a 'top five' research priority. Scand J Trauma Resusc Emerg Med 2018;26:89.
2. Rehn M, Hyldmo PK, Magnusson V, et al. Scandinavian SSAI clinical practice guideline on pre-hospital airway management. Acta Anaesthesiol Scand 2016;60:852-64.
3. Polytrauma Guideline Update Group. Level 3 guideline on the treatment of patients with severe/multiple injuries: AWMF Register-Nr. 012/019. Eur J Trauma Emerg Surg 2018;44(Suppl 1):3-271.
4. Shavit I, Levit B, Basat NB, Lait D, Somri M, Gaitini L. Establishing a definitive airway in the trauma patient by novice intubators: a randomised crossover simulation study. Injury 2015;46:2108-12.
5. Luckscheiter A, Lohs T, Fischer M, Zink W. Airway management in preclinical emergency anesthesia with respect to specialty and education. Anaesthesist 2020;69:170-82.
6. Arvind V, Kim JS, Cho BH, Geng E, Cho SK. Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. J Crit Care 2021;62:25-30.
7. Bolourani S, Brenner M, Wang P, et al. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. J Med Internet Res 2021;23:e24246.
8. Clark MT, Vergales BD, Paget-Brown AO, et al. Predictive monitoring for respiratory decompensation leading to urgent un-

planned intubation in the neonatal intensive care unit. Pediatr Res 2013;73:104-10.
9. Rahimian F, Salimi-Khorshidi G, Payberah AH, et al. Predicting the risk of emergency admission with machine learning: development and validation using linked electronic health records. PLoS Med 2018;15:e1002695.
10. Luckscheiter A, Lohs T, Fischer M, Zink W. Preclinical emergency anesthesia : a current state analysis from 2015-2017. Anaesthesist 2019;68:270-81.
11. Messelken M, Schlechtriemen T, Arntz HR, et al. Minimal data set in German Emergency Medicine MIND3. Notf Rett Med 2011;14:647-54.
12. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594.
13. Abouleish AE, Leib ML, Cohen NH. ASA provides examples to each ASA physical status class. ASA Newsl 2015;79:38-49.
14. Timmermann A, Bottiger BW, Byhahn C, et al. German guideline for prehospital airway management (short version). Anasthesiol Intensivmed 2019;6:316-36.
15. Witten IH, Frank E, Hall MA, Pal CJ. Data mining: practical machine learning tools and techniques. 4th ed. Cambridge, MA: Morgan Kaufmann; 2017.
16. Langer T, Favarato M, Giudici R, et al. Development of machine learning models to predict RT-PCR results for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in patients with influenza-like symptoms using only basic clinical data. Scand J Trauma Resusc Emerg Med 2020;28:113.
17. Yadav K, Sarioglu E, Choi HA, Cartwright WB 4th, Hinds PS, Chamberlain JM. Automated outcome classification of computed tomography imaging reports for pediatric traumatic brain injury. Acad Emerg Med 2016;23:171-8.
18. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-57.
19. Chicco D, Totsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min 2021;14:13.
20. Siu BM, Kwak GH, Ling L, Hui P. Predicting the need for intubation in the first 24 h after critical care admission using machine learning approaches. Sci Rep 2020;10:20931.
21. Politano AD, Riccio LM, Lake DE, et al. Predicting the need for urgent intubation in a surgical/trauma intensive care unit. Surgery 2013;154:1110-6.
22. Olaussen A, Blackburn T, Mitra B, Fitzgerald M. Review article: shock index for prediction of critical bleeding post-trauma: a systematic review. Emerg Med Australas 2014;26:223-8.
23. Tran A, Yates J, Lau A, Lampron J, Matar M. Permissive hypotension versus conventional resuscitation strategies in adult trauma patients with hemorrhagic shock: a systematic review and meta-analysis of randomized controlled trials. J Trauma Acute Care Surg 2018;84:802-8.
24. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432.
25. Basu S, Faghmous JH, Doupe P. Machine learning methods for precision medicine research designed to reduce health disparities: a structured tutorial. Ethn Dis 2020;30(Suppl 1):217-28.
26. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2:e489-92.
27. Grasner JT, Meybohm P, Lefering R, et al. ROSC after cardiac arrest: the RACA score to predict outcome after out-of-hospital cardiac arrest. Eur Heart J 2011;32:1649-56.

Supplementary Table 1. All recorded attributes and their values together with the class comparison and reason of exclusion

Attribute	Airway management		P-value	Reason of exclusion
	Yes (n = 1,451)	No (n = 24,105)		
Trauma site				No causality
Unknown	1.31	0.71	0.01[a]	
Home	24.60	28.43	< 0.01[a]	
Retirement home	2.00	3.07	0.02[a]	
Workplace	7.79	6.84	0.16	
Medical practice	0.21	0.54	0.09	
Street	44.59	44.31	0.80	
Public space	10.61	8.97	0.03[a]	
Hospital	1.45	1.24	0.48	
Mass event	0.14	0.19	0.67	
Educational institution	2.83	0.87	< 0.01[a]	
Sport facility	1.03	1.47	0.18	
Birth center	0.55	0.39	0.33	
Other	2.89	2.98	0.80	
Transportation				No causality
Ground	59.06	86.34	< 0.01[a]	
Air	1.59	1.82	0.50	
Ambulant	39.35	11.84	< 0.01[a]	
Emergency vehicle			< 0.01[a]	No causality
Ground ambulance	55.75	89.36		
Air ambulance	44.25	10.64		
Specialist and qualification				Preprocessing, left out for generalization
Anesthetist	65.61	53.58	< 0.01[a]	
Anesthetist with qualification in intensive care medicine	12.68	10.14	< 0.01[a]	
Other	19.92	33.34	< 0.01[a]	
Other with qualification in intensive care medicine	1.79	2.94	0.01[a]	
State of awareness				Preprocessing
Other	0.34	0.40	0.75	
Awake	69.06	84.22	< 0.01[a]	
Unconscious	5.93	1.50	< 0.01[a]	
Reacts to speech	7.51	6.16	0.04[a]	
Reacts to pain	3.86	1.90	< 0.01[a]	
Sedated	7.31	1.56	< 0.01[a]	
Unknown	6.00	4.26	< 0.01[a]	
Dementia				No causality
Yes	0.76	1.39	< 0.01[a]	
No	99.24	98.61		
Pathologic neurologic examination			< 0.01[a]	Preprocessing
Yes	9.92	6.47		
No	90.08	93.53		
Skin				Preprocessing
Other	11.99	8.60	< 0.01[a]	
Normal	29.57	66.85	< 0.01[a]	
Exsiccosis	1.31	3.56	< 0.01[a]	
Oedema	0.41	0.61	0.35	
Cold sweat	9.51	4.11	< 0.01[a]	
Missing value	47.21	16.28	< 0.01[a]	
Aggressive			0.30	Preprocessing
Yes	8.34	7.57		
No	91.66	92.43		

(*Continued on the next page*)

Supplementary Table 1. (Continued)

Attribute	Airway management		P-value	Reason of exclusion
	Yes (n = 1,451)	No (n = 24,105)		
Confusion			< 0.01[a]	Preprocessing
Yes	9.10	12.75		
No	90.90	87.25		
Acute CNS deficiency				Preprocessing
None	96.76	97.64	0.03[a]	
Seizure	1.03	1.03	0.99	
Stroke/bleeding	2.14	0.93	< 0.01[a]	
Other	0.07	0.39	0.049[a]	
Pulmonary embolism			0.50	Too rare
Yes	0	0.03		
No	100	99.97		
Acute cardiac disease			< 0.01[a]	Too rare
Yes	1.38	2.61		
No	98.62	97.39		
Bronchial spasm			0.10	Too rare
Yes	0	0.16		
No	100	99.84		
Aspiration/hemoptysis				
Yes	0.76	0.02	< 0.01[a]	Too rare
No	99.24	99.98		
Pneumothorax			0.04[a]	Too rare
Yes	0.07	0.01		
No	99.93	99.99		
Intoxication			< 0.01[a]	Too rare
Yes	0.90	2.15		
No	99.10	97.85		
Anaphylaxis			0.34	Too rare
Yes	0	0.06		
No	100	99.94		
Sepsis			0.48	Too rare
Yes	0	0.03		
No	100	99.97		
Hypothermia/hyperthermia			0.39	Too rare
Yes	0.69	0.52		
No	99.31	99.48		
Palliative condition			0.50	Too rare
Yes	0	0.02		
No	100	99.98		
Acute infection			0.40	Too rare
Yes	0.14	0.23		
No	99.86	99.77		
Acute abdomen			0.20	Too rare
Yes	0.41	0.26		
No	99.59	99.74		
Acute nontraumatic disease (summarized)			0.02[a]	Too rare
Yes	9.17	11.06		
No	90.83	88.94		
Upper limb			< 0.01[a]	Wrapper
None	80.36	64.33		
Mild	4.96	16.64		
Moderate	8.68	16.66		
Severe	6.00	2.37		

(*Continued on the next page*)

Supplementary Table 1. (Continued)

Attribute	Airway management		P-value	Reason of exclusion
	Yes (n = 1,451)	No (n = 24,105)		
Severe limb trauma (summarized)			< 0.01[a]	Wrapper
Yes	15.58	3.99		
No	84.42	96.01		
Soft parts				Wrapper
None	89.32	93.03	< 0.01[a]	
Mild	3.58	4.52	0.10[a]	
Moderate	3.79	1.98	< 0.01[a]	
Severe	3.31	0.47	< 0.01[a]	
Burn			< 0.01	Too rare
Yes	1.17	0.48		
No	98.83	99.52		
Acute inhalation injury			< 0.01	Too rare
Yes	0.76	0.11		
No	99.24	99.89		
Electrical accident			0.80	Too rare
Yes	0.14	0.12		
No	99.86	99.88		
Chemical burn			0.60	Too rare
Yes	0	0.02		
No	100	99.98		
Diving accident			0.35	Too rare
Yes	0	0.01		
No	100	99.99		
Other trauma			0.35	Too rare
Yes	1.10	1.40		
No	98.90	98.60		
Neck collar			< 0.01[a]	Wrapper
Yes	63.89	33.36		
No	36.11	66.64		
Mechanism of trauma				No causality
Blunt	57.89	78.45	< 0.01[a]	
Penetrating	4.14	5.58	0.02[a]	
Unknown	37.97	15.96	< 0.01[a]	
Circumstances of trauma				No causality, further details needed
Biker	11.65	9.94	0.03[a]	
Violent felony	4.48	1.68	< 0.01[a]	
Vehicle occupant	18.47	17.47	0.30	
Pedestrian	3.79	4.28	0.37	
Fall				
< 3 m	24.47	30.98	< 0.01[a]	
> 3 m	10.54	7.60	< 0.01[a]	
Bicyclist	12.06	11.82	0.78	
Assault	1.72	1.40	0.31	
Stabbed	0.34	0.64	0.16	
Shot	0	0.06	0.34	
Machine accident	0.55	0.71	0.47	
Burying	0.34	0.18	0.15	
Other modes of transport	0.83	0.92	0.71	
Other	4.89	6.37	0.02[a]	
Unknown	5.86	5.93	0.43	

(*Continued on the next page*)

Supplementary Table 1. (Continued)

Attribute	Airway management		P-value	Reason of exclusion
	Yes (n = 1,451)	No (n = 24,105)		
Crystalloid infusion			< 0.01[a]	Too general
Yes	93.87	86.02		
No	6.13	13.98		
Blood products			< 0.01[a]	Too rare
Yes	0.55	0.01		
No	99.45	99.99		
Blood glucose level (mg/dL)	159.77 ± 101.29	139.51 ± 81.73	< 0.01[a]	No causality
Age (yr)	54.88 ± 21.44	55.8 ± 22.28	0.13	Wrapper
Temperature (°C)	37.59 ± 9.28	38.38 ± 36.5	0.14	No causality
Systolic blood pressure < 90 mmHg			< 0.01[a]	Preprocessing
No	93.59	96.22		
Yes	6.41	3.78		
Glasgow Coma Scale < 9			< 0.01[a]	Preprocessing
No	85.87	96.67		
Yes	14.13	3.33		
Transport time (min)	10.58 ± 10.13	11.28 ± 10.53	0.01	No causality
Total	35.15	26.61	< 0.01[a]	Preprocessing
< 10	32.67	33.60		
11–20	7.31	12.45		
21–30	3.65	3.37		
31–45	1.03	0.85		
> 45	20.19	23.11		
Time on side (min)	24.32 ± 19.75	20.35 ± 14.97	< 0.01[a]	No causality
Total	6.00	12.56	< 0.01[a]	Preprocessing
< 15	41.42	45.48		
15–30	22.81	17.23		
31–45	8.34	3.51		
46–60	2.69	1.07		
> 60	18.75	20.15		
Prehospital time (min)	44.81 ± 38.25	41.40 ± 36.75	< 0.01[a]	No causality
Total	1.03	2.24	< 0.01[a]	Preprocessing
< 30	9.65	13.76		
30–45	23.64	24.07		
46–60	29.22	24.31		
61–90	6.27	4.73		
> 90	30.19	30.89		
Male sex	72.00	60.00	< 0.01[a]	No causality

Values are presented in percentage or mean ± standard deviation.
Some attributes were combined because they appeared too rarely. For example, the diagnosis of hypertensive emergency, pulmonary oedema, myocardial infarction, and other acute cardiac diseases were combined to "acute cardiac disease."
CNS, central nervous system.
[a]Statistically significant value (P < 0.05).

Supplementary Table 2. Settings of the principal component analysis in Weka

Setting	Principal component analysis	J48
Classifier	J48	-
Do not check capabilities	False	False
Evaluation measure	Accuracy	-
Calc out of bag	False	-
No. of folds	5	3
Threshold	0.01	-
Search method	Best first	-
Best first direction	Bidirectional	-
Lookup cache size	1	-
Search termination	5	-
Batch size	-	100
Binary splits	-	False
Collapse tree	-	True
Confidence factor	-	0.25
Debug	-	False
No. of decimal places	-	2
Do not make split point actual value	-	False
Reduced error pruning	-	False
Save instance data	-	False
Subtree raising	-	True
Unpruned	-	False
Laplace smoothing	-	False
MDL correction	-	True

MDL, minimum description length.

Supplementary Table 3. Settings of the SMOTE algorithm in Weka

Setting	SMOTE
Class value	0 (Minority class)
Debug	False
Do not check capabilities	False
Nearest neighbors	5
Percentage	200

SMOTE, synthetic minority oversampling method.

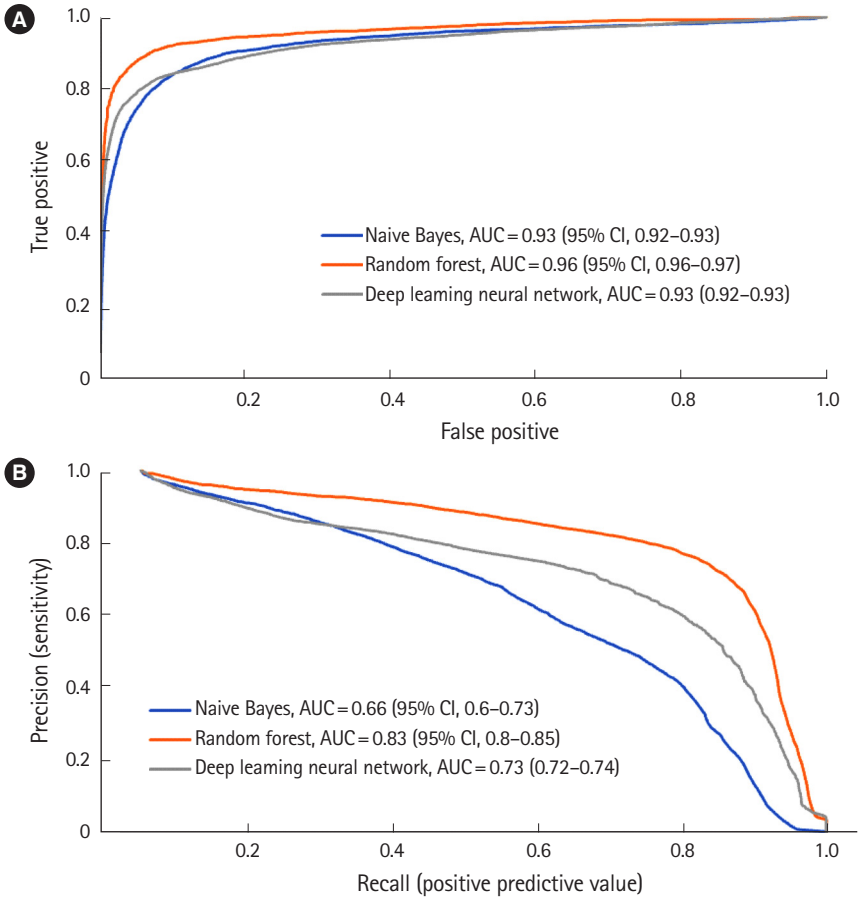
Supplementary Table 4. Settings of the random forest and Naive Bayes model in Weka

Setting	Random forest	Naive Bayes
Batch size	100	100
Break ties randomly	False	-
Calc out of bag	False	-
Debug	False	False
Do not check capabilities	False	False
No. of decimal places	2	2
Max depth	Unlimited	-
No. of execution slots	1	-
No. of randomly chosen attributes	0 (=log_2(#predictors)+1)	-
No. of iterations	100	-
Bag size percent	100	-
Supervised discretization	-	False
Kernel estimator	-	False

Supplementary Table 5. Performance of two deep learning networks before and after attribute selection

Variable	Deep learning neural network	
	Before attribute selection[a]	After attribute selection[b]
Overall correctness	96.51 ± 0.31 (96.29–96.73)	96.68 ± 0.17 (96.56–96.8)
Kappa	0.66 ± 0.03 (0.64–0.68)	0.68 ± 0.01 (0.67–0.69)
AUC-ROC	0.90 ± 0.01 (0.89–0.91)	0.93 ± 0.01 (0.92–0.93)
MCC	0.66 ± 0.03 (0.64–0.68)	0.68 ± 0.01 (0.67–0.69)
Sensitivity	0.73 ± 0.04 (0.70–0.75)	0.74 ± 0.04 (0.71–0.76)
Positive predictive value	0.64 ± 0.03 (0.61–0.66)	0.69 ± 0.09 (0.62–0.76)
PRC area	0.67 ± 0.03 (0.65–0.69)	0.73 ± 0.01 (0.72–0.74)
Specificity	0.95 ± 0.08 (0.89–1.01)	0.98 ± 0 (0.98–0.98)
Negative predictive value	0.99 ± 0 (0.98–0.99)	0.99 ± 0 (0.98–0.99)
PRC area	0.99 ± 0 (0.99–0.99)	0.98 ± 0.02 (0.97–1.00)

AUC-ROC, area under the receiver operator curve; MCC, Matthews correlation coefficient; PRC, precision-recall.
[a]Two dense layers with 30 neurons each. [b]One dense layer with six neurons.



Supplementary Fig. 1. Averaged receiver operator curves (ROC) for (A) the overall performance and (B) the averaged precision-recall (PRC) curves for the prediction of airway management of the Naive Bayes, the random forest algorithm, and the deep learning neural network (one dense layer with six neurons) after attribute selection. AUC, area under the curve; CI, confidence interval.