BMJ Health & Care Informatics

Artificial intelligence-based prediction of transfusion in the intensive care unit in patients with gastrointestinal bleeding

Riccardo Levi ,[1] Francesco Carli,[2] Aldo Robles Arévalo,[3] Yuksel Altinel,[4] Daniel J Stein,[5] Matteo Maria Naldini,[6] Federica Grassi,[7] Andrea Zanoni,[8] Stan Finkelstein,[9,10] Susana M Vieira,[3] João Sousa,[3] Riccardo Barbieri,[1] Leo Anthony Celi [11,12]

To cite: Levi R, Carli F, Arévalo AR, *et al*. Artificial intelligence-based prediction of transfusion in the intensive care unit in patients with gastrointestinal bleeding. *BMJ Health Care Inform* 2021;**28**:e100245. doi:10.1136/bmjhci-2020-100245

Received 24 September 2020
Revised 26 October 2020
Accepted 27 November 2020

For numbered affiliations see end of article.

Correspondence to
Mr Riccardo Levi;
riccardo.levi@mail.polimi.it

ABSTRACT

Objective Gastrointestinal (GI) bleeding commonly requires intensive care unit (ICU) in cases of potentialhaemodynamiccompromise or likely urgent intervention. However, manypatientsadmitted to the ICU stop bleeding and do not require further intervention, including blood transfusion. The present work proposes an artificial intelligence (AI) solution for the prediction of rebleeding in patients with GI bleeding admitted to ICU.

Methods A machine learning algorithm was trained and tested using two publicly available ICU databases, the Medical Information Mart for Intensive Care V.1.4 database and eICU Collaborative Research Database using freedom from transfusion as a proxy for patients who potentially did not require ICU-level care. Multiple initial observation time frames were explored using readily available data including labs, demographics and clinical parameters for a total of 20 covariates.

Results The optimal model used a 5-hour observation period to achieve an area under the curve of the receiving operating curve (ROC-AUC) of greater than 0.80. The model was robust when tested against both ICU databases with a similar ROC-AUC for all.

Conclusions The potential disruptive impact of AI in healthcare innovation is acknowledge, but awareness of AI-related risk on healthcare applications and current limitations should be considered before implementation and deployment. The proposed algorithm is not meant to replace but to inform clinical decision making. Prospective clinical trial validation as a triage tool is warranted.

Summary

What is already known?

- ► Gastrointestinal bleeding is a severe event that requires admission to the ICU.
- ► Many patients in the ICU for gastrointestinal bleeding undergo only increased monitoring without intervention.
- ► ICU stay is associated with increased cost and morbidity.

What does this paper add?

- ► An algorithmic approach using artificial intelligence on readily available electronic data can accurately predict ICU transfusion need.
- ► Using this approach to identify patients at low risk for ongoing bleeding and transfusion could be validated prospectively to identify patients who may not require ICU-level care.

INTRODUCTION

Gastrointestinal (GI) haemorrhage is a common condition that frequently requires hospitalisation, often in the intensive care unit (ICU)[1] with considerable associated morbidity. In particular, ICU admission is associated with increased costs and a greater rate of complications and poor outcomes compared with ward admission.[2–4] Some patients are initially admitted to the ICU for haemodynamic instability but stabilise without further intervention and are discharged to the ward the following day.

Previous instruments, such as the Rockall or the Blatchford score[5] have been applied to triage patients based on the likelihood of mortality, recurrent/ongoing bleeding, need for hospitalisation and requirement for endoscopic intervention. However, these models are validated only for upper GI bleeding with a focus on endoscopic intervention and mortality and do not assist in informing level of monitoring for hospitalised patients. Currently, there is no model to assist in triaging patients with GI bleeding including those with an undifferentiated source to an appropriate acuity of care.

We identified the need for blood transfusion as a surrogate for persistent bleeding. Previous prospective studies have shown that up to half of patients with GI bleeding may not require transfusion.[6] We used an ICU database to train a prediction model but focused on the first few hours on arrival as a proxy of the patient's state in the emergency department.

The use of artificial intelligence (AI) represents an opportunity for more effective and efficient care delivery by predicting disease trajectory and complications.[7–12] Previous work in GI bleeding has used methods such as artificial neural networks,[13 14] support vector machines[13] to predict the need for intervention; and fuzzy models[15] to identify which lab test is likely to contribute information gain and influence clinical management of patients with GI bleeding in the ICU. This study focused on using machine learning to predict transfusion to better identify those patients who continue to bleed.

METHODS

This study is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology statement.[16]

Database description

Data were collected from the Medical Information Mart for Intensive Care-III (MIMIC-III) V.1.4[17] and in the eICU Collaborative Research Database V.2.0 (eICU-CRD).[18] Both databases contain information from patients admitted to the ICU. The MIMIC-III database collects detailed haemodynamic and clinical parameters from all ICU patients admitted to a single major academic medical centre between 2008 and 2014, whereas the eICU-CRD is a multicentre database with high granularity data for over 200 000 admissions to ICUs monitored by an eICU[19] across the USA.

Ethical approval

Both databases are previously de-identified and have been reviewed by the institutional review boards (IRB) of their hosting organisations and determined to be exempt from subsequent IRB.

Definition of outcome

The outcome of this study is ongoing GI bleeding after admission to the ICU. Since this outcome variable is not encoded, blood transfusions were used as surrogate marker.

Software

Models were developed in Python V.3.7 using data science packages including pandas V.0.25.3 (data wrangling),[20] NumPy V.1.17.5 (computations),[21] SciPy V.1.4.1 (hypothesis testing),[22] Scikit-learn V.0.22.1 (modelling)[23] and Hyperopt V.0.2.3 (hyperparameter optimisation).[24]

Data preparation

We included non-pregnant adult patients (≥18 years old) admitted to the ICU and diagnosed with GI bleeding based on the International Classification of Diseases (ICD-9) codes (see table A1, online supplemental digital content 1,). For patients with multiple ICU admissions within a single hospitalisation event, only the first ICU stay was considered. The inclusion criteria for each database are further detailed in figure 1.

Missing records were imputed with the last observation available carried forward. Patients missing their first value were imputed with the intra-subject median. In order to take into account the dynamics of the observed features within the training window (eg, increasing, decreasing trends), we adopted a feature engineering approach (see text, online supplemental digital content 2). Also, non-normally distributed features (skewness >3) were log-transformed[25] in order to obtain a normal distribution for improved model performance.

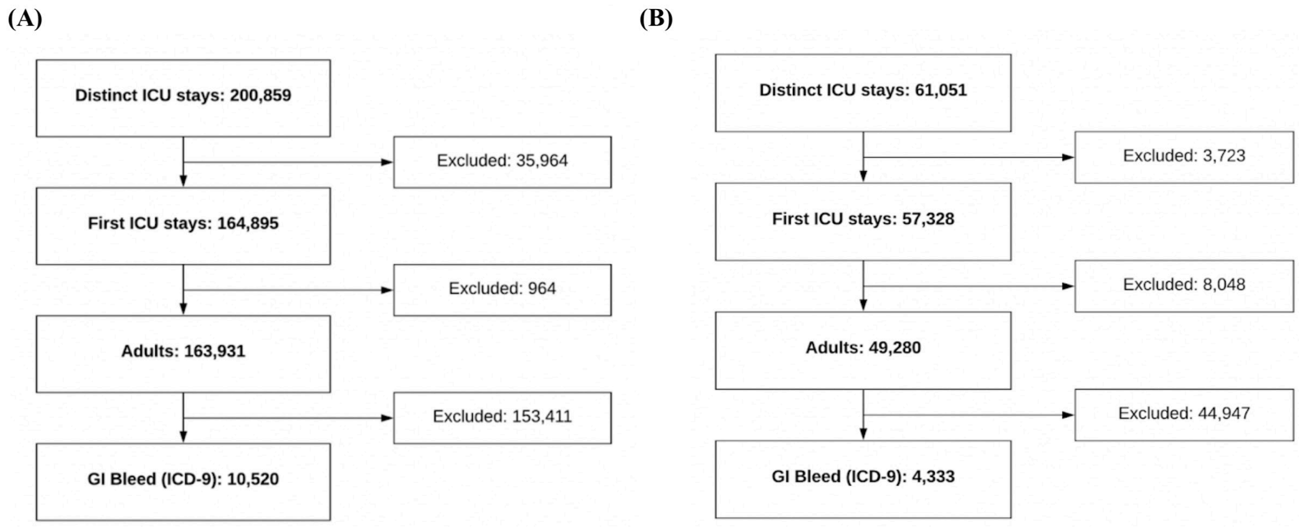

Figure 1 Inclusion criteria for the cohort extracted from the (A) eICU-CRD and (B) MIMIC-III. eICU-CRD, eICU Collaborative Research Database; ICD-9, International Classification of Diseases-9; ICU, intensive care unit; GI, gastrointestinal; MIMIC-III, Medical Information Mart for Intensive Care-III.

Table 1 List of covariates, the output variable and demographic information for each cohort. Continuous variables are stated as mean (IQR), otherwise are the number of occurrences. only a subset of these variables (selected by recursive feature elimination procedure) enters in the final models.

	MIMIC-III (n=4314)	eICU-CRD (n=10 306)
Demographics		
Age at admission (years)	83.5 (56–81)	76.7 (56–79)
Gender (n)		
Male	2491	5927
Female	1823	4379
Output variable (transfusion)		
Transfused patients (n, % wrt total number of patients)	2077 (48.15%)	2712 (26.31%)
Covariates		
Heart rate (bpm)	92.9 (79.0–105.7)	94.0 (79.9–106.5)
Mean blood pressure (mm Hg)	78.9 (68.5–87.8)	78.4 (67.6–87.5)
Systolic blood pressure (mm Hg)	114.5 (99.0–129.0)	108.1 (93–121)
Diastolic blood pressure (mm Hg)	60.3 (54.7–65.2)	62.6 (56.0–68.2)
Respiratory rate (breaths/min)	21.2 (18.0–24.0)	21.9 (17.8–24.4)
Haematocrit (%)	28.4 (23.8–32.6)	26.5 (20.7–31.6)
Haemoglobin (g/L)	97 (80–112)	87 (67–104)
White blood cell (×10^9/L)	11.8 (7.2–14.1)	11.7 (7.4–14.4)
Platelet (×10^9/L)	227.5 (137.0–286.0)	207 (129.0–263.0)
Creatinine (mg/dL)	1.79 (0.85–1.88)	1.73 (0.80–1.90)
Blood urea nitrogen (mg/dL)	39.5 (19.0–51.0)	39.2 (19.0–51.0)
Potassium (mEq/L)	4.34 (3.80–4.70)	4.38 (3.80–4.80)
Bicarbonate (mEq/L)	22.6 (20.0–26.0)	22.7 (20.0–26.0)
Amount blood transfused (mL)	601.0 (375.0–750.0)	571.9 (324.0–700.0)
Glucose (mg/dL)	160.2 (106.0–174.0)	153.2 (105.0–176.0)
Albumin (g/dL)	3.17 (3.2–3.2)	2.96 (2.8–3.1)
Temperature (°C)	36.3 (36.0–36.7)	36.4 (36.4–36.5)
Partial thromboplastin time (s)	37.3 (26.1–37.9)	35.3 (26.0–37.0)

eICU-CRD, eICU Collaborative Research Database; ICU, intensive care unit; MIMIC-III, Medical Information Mart for Intensive Care-III.

Feature selection has been performed by recursively discarding features that do not reduce accuracy performance when eliminated. This procedure is called recursive feature elimination (RFE), a method used to remove non-predictive covariates with a greedy approach[26] (see text, online supplemental digital content 3). Final input datasets gather 4333 first ICU admissions from the MIMIC-III database and 10520 first ICU admissions from the eICU-CRD along with 20 covariates. Input variables include several laboratory analyses and demographic information that are available in each database. Detailed information of these features is described in table 1.

Prediction time windows

Several time windows were assessed for data extraction of the training/testing data and the data for the output variable (blood transfusion) that was predicted. Four different time windows starting from ICU admission (hour 0) were evaluated: training time from 0 to 3 hours/prediction time 4–24 hours, training time 0–4 hours/prediction time 5–24 hours, training time 0–5 hours/prediction time 6–24 hours, training time 0–6 hours/prediction time 7–24 hours. The training timeframe contains the covariates recorded during that time frame for each ICU stay. All training time windows include information recorded prior to the ICU admission (up to −1 hour). The prediction time window is when the surrogate variable (blood transfusion) was recorded (see figure 2).

This analysis helped us to find the optimal training/prediction time windows. The selected time windows were those that achieved the best predictive performance. In addition to that, the best training time window is the one that gathered the highest amount of data before a blood transfusion. Except from that, there is no other contextual detail that was considered during this analysis.

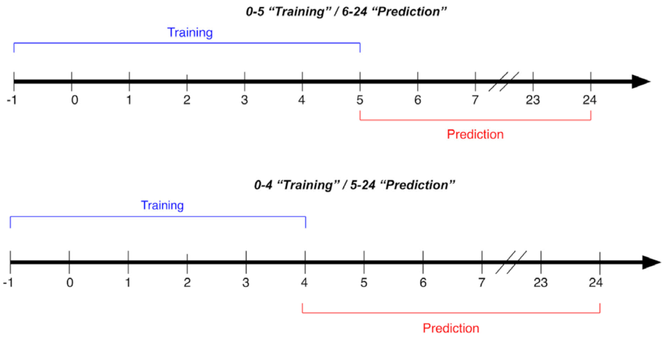
0-5 “Training” / 6-24 “Prediction”
Training
-1
0
1
2
3
4
5
6
7
23
24
Prediction
0-4 “Training” / 5-24 “Prediction”
Training
-1
0
1
2
3
4
5
6
7
23
24
Prediction

Figure 2 Graphical schema of the time windows.

Training and testing partitions

Several training/testing partitions and strategies were designed in order to fully exploit the information contained in both datasets. Specifically, both datasets are randomly divided into a test (25% of records) and training set (75% of records). A model is fitted on each of the training sets and on a combination of the two. All training subsets were split to perform 10-fold cross validation and to optimise model's hyperparameters. The testing subsets had data that were not used for training/validation.

Three different training sets were considered: (1) including MIMIC-III data only (n=3235); (2) including eICU-CRD data only (n=7729) and (3) a training set composed by 29.17% of MIMIC-III and 70.83% of eICU-CRD (n=10 964). The performance of the models is then gauged on both the test sets, allowing for an external validation of the classifiers for a total of three models per each considered time window:

1. Train on MIMIC-III, internal validation on MIMIC, external validation on eICU-CRD.
2. Train on eICU-CRD, internal validation on eICU-CRD, external validation on MIMIC-III.
3. Train on MIMIC-III and eICU-CRD, internal validation on MIMIC-III and eICU-CRD.

Predictive models

In order to improve the performance of individual machine learning models, the final classifier is determined as an ensemble of machine learning models combined together. To select the models for this ensemble, we assessed several classifiers. Hyperparameter tuning was performed through Bayesian optimisation[27] with a stratified 10-fold cross validation, where class imbalance is taken into account in the parameters of the models. This tuning is carried out with a customised loss function that takes into account accuracy and F1 score (see text, online supplemental digital content 4). This delivers a model based (and hence non arbitrary) procedure to find cut-off-thresholds that optimise jointly the accuracy, specificity and sensitivity of the model. By specifying the weights of F1 score and accuracy inside the custom loss function the model could be oriented to avoid false negative predictions (higher F1 score and recall) with a high accuracy. However, since the model also provides the probability that a patient will bleed the physician could in principle perform standard sensitivity–specificity trade-off decisions.

Given that eICU-CRD exhibits target imbalance (26% transfused patients against 74% non-transfused patients) classifiers trained on this dataset are imbalance-aware in order not to skew predictions towards the majority class (ie, predicting all patients as low risk patients, which is not desirable).

Permutation feature importance[28] of the five most important covariates is estimated for each model. Moreover, the partial dependence function[29] function of the outcome with respect to the most important variable is estimated (see text, online supplemental digital content 5).

In order to assess the goodness of the classifier during testing, we estimated the model's accuracy, sensitivity (recall or true classification positive rate), specificity (true negative classification rate) and area under the curve of the receiving operating curve (ROC-AUC).

To conclude, models are calibrated through Platt's scaling[30 31] to obtain reliable probability estimates. The effects of the calibration can be diagnosed visually with the calibration curves (see text, online supplemental digital content 6).

RESULTS

The best results are achieved when the models are trained on the MIMIC-III dataset (see table A2, online supplemental digital content 7), and the lowest values are observed in the models trained on the eICU-CRD data (see table A3, online supplemental digital content 8). When both datasets are merged (see table A4, online supplemental digital content 9), the performance does not improve considerably, but we can observe a significant improvement in terms of sensitivity. Of note, the sensitivity obtained in the models trained with MIMIC-III is the highest among all other models; which indicates that it is better to detect true positive cases or patients that would require transfusion.

It is also interesting to highlight that the models trained on MIMIC-III (see table A2, online supplemental digital content 7) have a greater discriminative power on the eICU-CRD testing set than the models trained only on the eICU-CRD data (see table A3, online supplemental digital content 8) and even if these are tested on the same database. Thus, a model trained on MIMIC-III is capable of generalising better to patients that the model does not train on.

These observations could be explained by the fact the MIMIC-III input dataset is not skewed (48.14% of the entries required transfusion) as the input dataset from the eICU-CRD (26.31% of the entries required transfusion). This imbalance could skew the model predictions towards the majority class (the most frequent label in the population) that are the patients that did not bleed (not required transfusion).

Table 2 Results for the time window composed by the pair training time of 0–5 hours/prediction time 6–24 hours

	Testing sets							
	ROC-AUC		Accuracy		Specificity		Sensitivity	
Training sets	MIMIC-III	eICU-CRD	MIMIC-III	eICU-CRD	MIMIC-III	eICU-CRD	MIMIC-III	eICU-CRD
MIMIC-III	0.8141	0.7634	0.7470	0.5021	0.6482	0.3502	0.8536	0.9277
eICU-CRD	0.8017	0.7858	0.7470	0.7060	0.7982	0.6872	0.6917	0.7581
MIMIC-III+eICU-CRD	0.8035	0.7908	0.7488	0.6884	0.7143	0.6535	0.7861	0.7861

eICU-CRD, eICU Collaborative Research Database; MIMIC-III, Medical Information Mart for Intensive Care-III; ROC-AUC, area under the curve of the receiving operating curve.

To avoid these misclassifications, the decision threshold was tuned during the optimisation procedure. In case the models were optimised only in terms of accuracy, it could have pushed the model to predict the majority class (non-transfused). By using the customised loss function, it was forced to jointly maximise precision and the recall of the final model notwithstanding the accuracy.

Looking at the results reported in table 2, online supplemental tables A5–A7 (see tables A, online supplemental digital content 10–12) we notice that the performances of all the time windows are satisfying and the overall best ones are obtained when the training phase is performed with data collected in the time window 0–5 hours and the prediction time window is from 6 to 24. Hence, in the following, we will mainly focus on this subdivision.

The models achieve greater ROC-AUC values when they are tested on the MIMIC-III dataset (>0.80) compared with the models tested on the eICU-CRD (0.76–0.79) as shown in table 2. Only accuracy and specificity improve when the models are trained in the eICU-CRD, but no improvement is detected in terms of sensitivity. The highest true classification positive rate is achieved in the models trained on the MIMIC-III, a critical metric being that it indicates how good are the models to predict the need of transfusions (true positives). We remark that this behaviour was expected since the eICU-CRD dataset has a larger variety of patients and hospitals than on the MIMIC-III. Therefore, adding more training data with different characteristics is beneficial for the former but not for the latter.

The highest value of ROC-AUC is achieved when the model is both trained and tested in the MIMIC-III (0.81) as verified in figure 3 as well. When the same model is tested in the eICU-CRD dataset, we observed lower ROC-AUC values. This metric is improved (0.79) when the model is trained with both datasets, but tested in the same dataset. In terms of the ability to predict transfusion, the model trained in MIMIC-III and tested on the eICU-CRD dataset achieves the best sensitivity (0.93).

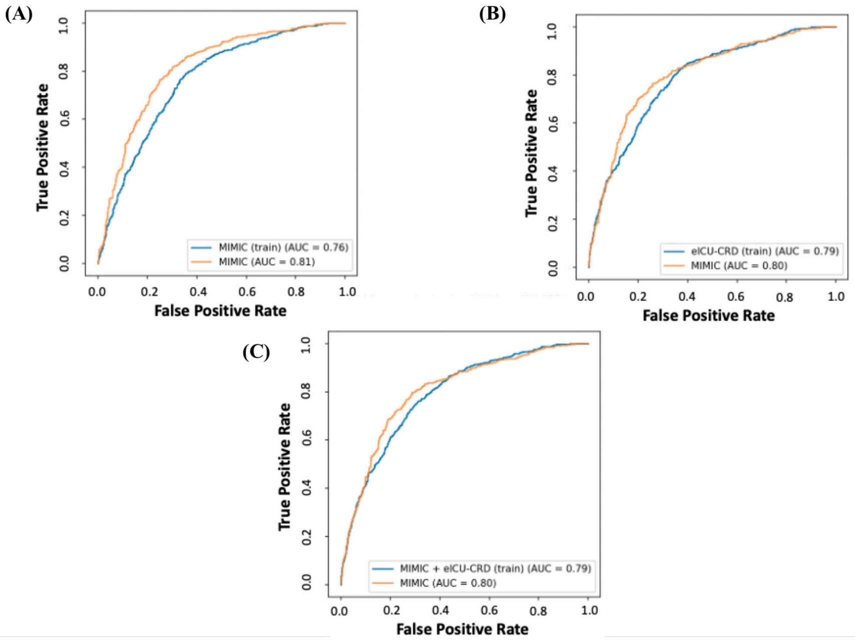

Figure 3 ROC plot for all the test sets. Model is trained on (A) the MIMIC-III training set, (B) the eICU-CRD and (C) on the training set that contains both the MIMIC-III and the eICU-CRD. AUC, area under the curve; eICU-CRD, eICU Collaborative Research Database; ICU, intensive care unit; MIMIC-III, Medical Information Mart for Intensive Care-III; ROC, receiving operating curve.

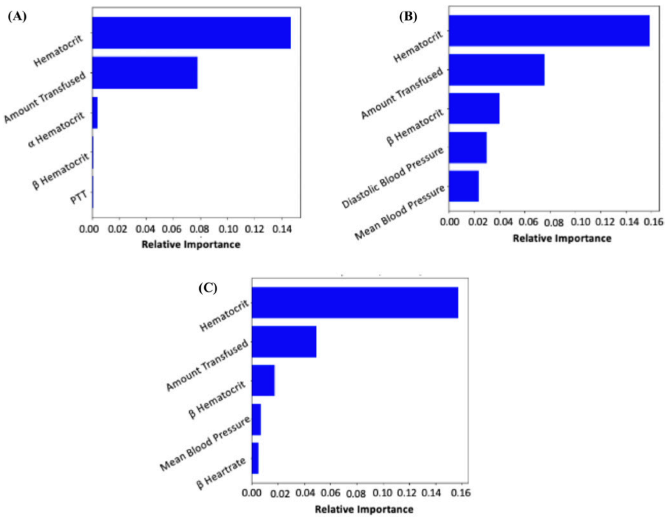

Figure 4 Feature importance plots for all the training sets. Model is trained on (A) the MIMIC-III training set, (B) the eICU-CRD and (C) on the training set that contains both the MIMIC-III and the eICU-CRD. eICU-CRD, eICU Collaborative Research Database; ICU, intensive care unit; MIMIC-III, Medical Information Mart for Intensive Care-III.

The most important features (see figure 4) to predict the need of transfusion are the haematocrit and the amount of blood already transfused during the training time window (0–5 hours) with the corresponding time pattern features (slope and intercept of haematocrit). Because of the importance of haematocrit, the interaction between this feature and the output variable was assessed visually in the partial dependence plots shown in figure 5.

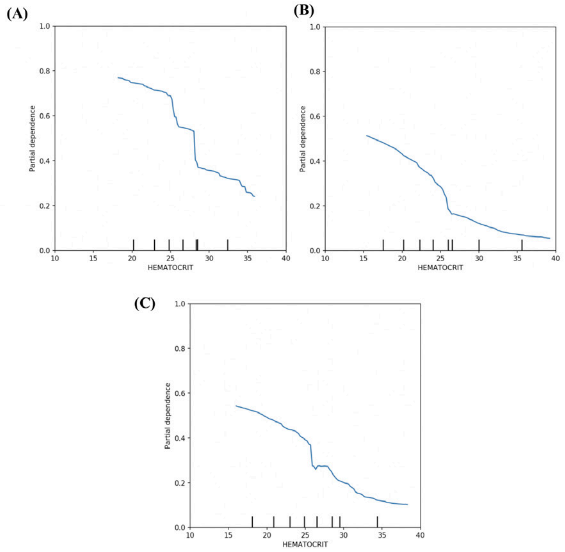

Figure 5 Partial dependence plot of the need of transfusion on haematocrit for all the training sets. Model is trained on (A) the MIMIC-III training set, (B) the eICU-CRD and (C) on the training set that contains both the MIMIC-III and the eICU-CRD. eICU-CRD, eICU-CRD, eICU Collaborative Research Database; ICU, intensive care unit; MIMIC-III, Medical Information Mart for Intensive Care.

Despite the three plots do not have identical shapes, the same trend is verified in the three plots: haematocrit is inversely proportional to the output variable. That implies that if values of haematocrit decreases, the probability of needing blood transfusion increases. Moreover, the partial dependence function shown in figure 5 highlights the presence of a discriminative threshold in the model with respect to haematocrit. It indicates that if the value of haematocrit is greater than this threshold, the probability of bleeding increases substantially. We remark that the value for this threshold seems to be dependent on the dataset that is used for training, where this shift is more noticeable (figure 5A).

DISCUSSION

GI bleeding remains a common reason for ICU admission. In a dataset consisting of over 10 000 patients admitted to the ICU with GI haemorrhage (both upper and lower), under half require transfusion during their ICU admission.[32] We present a model based on observations from the first 5 hours of ICU admission to predict the need for transfusion in the next 24 hours of admission with a high level of accuracy (overall AUC of 0.80). The patient's vital signs and laboratory test findings during the first few hours in the ICU are a good proxy of the measurements in the emergency department.

In the clinical setting, the need for transfusion has been an outcome of interest for GI haemorrhage. Prior work from Villanueva *et al*[6] found that even in active upper GI bleeding, up to half of patients do not require transfusion. Furthermore, it has been established that while the minority of patients with upper GI bleeding require hospitalisation, this can be a significant driver of costs. By identifying patients who will no longer require transfusion, it is possible to safely triage these patients to a regular ward, or even discharged to home if ambulatory monitoring can be provided.

Previous work in this area has focused either on upper or lower GI bleeding separately. In a 2016 analysis by Robertson *et al*,[32] the Rockall, AIMS65 and Glasgow-Blatchford Score (GBS) were all used to predict outcomes for upper GI bleeding. In their population, a total of 62% of the patients required a blood transfusion. They found the GBS to be the best predictor with an ROC-AUC of 0.90. Both the AIMS65 (ROC-AUC 0.72) and full (ROC-AUC 0.68)/pre-endoscopy (ROC-AUC 0.66) Rockall scores were considerably less accurate. However, the use of these scores to predict the need for transfusion has limitations. First, the only score with an ROC-AUC over 0.8, the GBS was validated only on upper GI bleeding (primarily ulcer-related in the initial validation). Furthermore, relying on clinical data input from the healthcare providers, for example, presence of melena, presentation with syncope, presence of heart failure, introduces opportunities for error and bias. Attempts to generalise the use of GBS to lower GI bleeding have found some success but focuses primarily on the prediction of mortality and need for an intervention instead of transfusion, and with suboptimal accuracy.

The sensitivity, or recall, of the models trained on MIMIC-III is the highest among all other models. A high recall means the algorithm identifies the majority of patients who require transfusion. For the use case presented, sensitivity is more important than precision, or the true positivity rate. When several models have similar ROC-AUC, sensitivity should be prioritised over precision. The consequence of missing patients who eventually bleed and sending them to the regular floor or even discharging them home is worse than over-calling potential persistent bleeders and getting them admitted to the ICU. The context in which the algorithm will be used and for what purpose are crucial to the model building.

Even when models are externally validated in another dataset, there is no guarantee that it will perform well in another patient population. External validation does not circumvent the need to evaluate algorithms trained elsewhere using local data prior to deployment. The performance of any predictive model is dependent on the database used to train the algorithm, and thus, the features available as candidate variables. The relationship between the features and the output of an algorithm is influenced by local practice patterns. In addition, model performance should be continuously monitored after deployment as accuracy almost always wanes over time, requiring model re-calibration.[33]

We submit the potential disruptive impact of AI-based technologies in precision medicine and in clinical decision-support systems. Nonetheless, we are aware of AI-related risks on healthcare applications and the pitfalls that have occurred in the past.[34] Although we reduced the risk of misclassification in the design of our models, we propose a human in the loop system for decision support. A final decision still rests on the healthcare provider after a careful clinical assessment which now includes input from the algorithm. Moreover, before implementation to a real clinical setting, the algorithm requires regulatory approval, human factors engineering to incorporate it into the workflow and prospective evaluation of its impact on hard clinical endpoints including patient harm from false negative predictions.

There are key strengths to the model we presented. First, the calculation can be completely automated without clinician input of symptoms and past medical history. Furthermore, it does not require identification of the source of bleeding–upper versus lower. The model performed well on held out test sets from two different databases, one of them collected from more than 200 hospitals across the USA.

Despite model validation on two databases, the algorithm is not guaranteed to perform accurately in a different institution. We present a reproducible methodology that other hospitals can employ to develop their own algorithm, as different patient demographics and practice patterns would undoubtedly modify the relationship of the features with the outcome being predicted, that is, the need for blood transfusion. At the very least, medical AI algorithms require evaluation on data from the local population prior to prospective evaluation using hard clinical endpoints.

Going forward, this work presents a methodology to build a clinical AI-based model that potentially can be implemented for prediction of the need for transfusion. The algorithm is not meant to replace but to inform decision making, specifically around identification of patients who may not benefit from an ICU-level of care. A prospective trial is warranted to assess the utility of this model in clinical usage.

Author affiliations
[1]Department of Electronic, Information and Bioengineering, Politecnico di Milano, Milano, Italy
[2]Department of Informatics, Università degli Studi di Torino, Torino, Piemonte, Italy
[3]IDMEC, Instituto Superior Tecnico, Universidade de Lisboa, Lisbon, Portugal
[4]General Surgery Department, Istanbul Bagcilar Training and Research Hospital, Istanbul, Turkey
[5]Department of Internal Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[6]San Raffaele Telethon Institute for Gene Therapy, Milano, Lombardia, Italy
[7]School of Medicine and Surgery, Università degli Studi di Milano-Bicocca, Milano, Lombardia, Italy
[8]Institute of Mathematics, Ecole Polytechnique Federale de Lausanne, Lausanne, VD, Switzerland
[9]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[10]Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Boston, MA, USA
[11]Laboratory for Computational Physiology, Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA
[12]Division of Pulmonary Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

Twitter Leo Anthony Celi @MITCriticalData

Acknowledgements Part of the analysis was performed during the fall course HST.953, Collaborative Data Science in Medicine, at the Massachusetts Institute of Technology and at the first European Society of Intensive Care Medicine datathon in Milan, Italy.

Contributors All authors contributed to writing the manuscript. FC, RL, ARA collaborated on the data extraction, modeling, visualisation and analysis. YA, AZ, MMN, FG guided data extraction. DJS and LAC interpreted, validated results, design of the work and supervised data extraction. SMV, JS, RB and SF reviewed the paper and supervised the work.

Funding ARA is supported by Fundação da Ciência e da Tecnologia (FCT-Portuguese National Science Foundation) through the PhD fellowship PD/BD/114107/2015. ARA, SMV and JS are supported by FCT, through IDMEC, under LAETA, project UIDB/50022/2020; and by FEDER under Programa Operacional Regional de Lisboa, project LISBOA-01-0145-FEDER-031474. LAC is funded by the National Institute of Health through NIBIB R01 EB017205. The project commenced during the HST.953 course (Collaborative Data Science) in Medicine) at the Massachusetts Institute of Technology.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The Massachusetts Institute of Technology (No. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) both approved the use of the MIMIC-III for research. The eICU-CRD was certified as meeting Safe Harbor standards by an independent privacy expert (Privacert, Cambridge, Massachusetts, USA) (Health Insurance Portability and Accountability Act Certification No. 1031219-2).

Data availability statement The data that support the findings of this paper employs data from MIMIC-III and eICU-CRD databases. The access to these datasets is controlled and researchers should request access on the PhysioNet website (https://physionet.org/about/database/). The code is available upon reasonable request contacting R. Levi at riccardo.levi@mail.polimi.it.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Riccardo Levi http://orcid.org/0000-0001-9030-0071
Leo Anthony Celi http://orcid.org/0000-0001-6712-6626

REFERENCES

1 Rahman SI-U, Saeian K. Nonvariceal upper gastrointestinal bleeding. *Crit Care Clin* 2016;32:223–39.
2 Esrailian E, Gralnek IM. Nonvariceal upper gastrointestinal bleeding: epidemiology and diagnosis. *Gastroenterol Clin North Am* 2005;34:589–605.
3 Minne L, Abu-Hanna A, de Jonge E. Evaluation of SOFA-based models for predicting mortality in the ICU: a systematic review. *Crit Care* 2008;12:R161.
4 Akaraborworn O, Chaiwat O, Chatmongkolchart S, *et al*. Prediction of massive transfusion in trauma patients in the surgical intensive care units (THAI-SICU study). *Chin J Traumatol* 2019;22:219–22.
5 Blatchford O, Murray WR, Blatchford M. A risk score to predict need for treatment for upper-gastrointestinal haemorrhage. *Lancet* 2000;356:1318–21.
6 Villanueva C, Colomo A, Bosch A, *et al*. Transfusion strategies for acute upper gastrointestinal bleeding. *N Engl J Med* 2013;368:11–21.
7 The Lancet Digital Health. Walking the tightrope of artificial intelligence guidelines in clinical practice. *Lancet Digit Health* 2019;1:e100.
8 Cosgriff CV, Celi LA, Stone DJ. Critical care, critical data. *Biomed Eng Comput Biol* 2019;10:117959721985656–7.
9 Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319:1317.
10 Kindle RD, Badawi O, Celi LA, *et al*. Intensive care unit telemedicine in the era of big data, artificial intelligence, and computer clinical decision support systems. *Crit Care Clin* 2019;35:483–95.
11 Gholami B, Haddad WM, Bailey JM. Ai in the ICU: in the intensive care unit, artificial intelligence can keep Watch. *IEEE Spectr* 2018;55:31–5.
12 Ruffle JK, Farmer AD, Aziz Q. Artificial Intelligence-Assisted Gastroenterology— promises and pitfalls. *Off J Am Coll Gastroenterol* 2019;114.
13 Chu A, Ahn H, Halwan B, *et al*. A decision support system to facilitate management of patients with acute gastrointestinal bleeding. *Artif Intell Med* 2008;42:247–59.
14 Das A, Ben-Menachem T, Farooq FT, *et al*. Artificial neural network as a predictive instrument in patients with acute nonvariceal upper gastrointestinal hemorrhage. *Gastroenterology* 2008;134:65–74.
15 Cismondi F, Celi LA, Fialho AS, *et al*. Reducing unnecessary lab testing in the ICU with artificial intelligence. *Int J Med Inform* 2013;82:345–58.
16 Vandenbroucke JP, von Elm E, Altman DG, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLoS Med* 2007;4:e297.
17 Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.
18 Pollard TJ, Johnson AEW, Raffa JD, *et al*. The eICU Collaborative research database, a freely available multi-center database for critical care research. *Sci Data* 2018;5:180178.
19 Celi LA, Hassan E, Marquardt C, *et al*. The eICU: It's not just telemedicine. *Crit Care Med* 2001;29:N183–9 http://journals.lww.com/00003246-200108001-00007
20 McKinney W. Data structures for statistical computing in python. proC. 9th python SCI. *Conf* 2010.
21 der WSvan, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng* 2011;13:22–30.
22 Virtanen P, Gommers R, Oliphant TE, *et al*. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods* 2020;17:261–72.
23 Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011.
24 Bergstra J. Distributed Asynchronous Hyperparameter Optimization [Internet]. Available: https://pypi.org/project/hyperopt/ [Accessed cited 2020 May 30].
25 Yeo I-K, Johnson RA. A new family of power transformations to improve normality or symmetry. *Biometrika* 2000;87:954–9.
26 Guyon I, Weston J, Barnhill S, *et al*. Gene selection for cancer classification using support vector machines. *Mach Learn* 2002;46:389–422.
27 Bergstra J, Yamins D, Cox DD. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. JMLR.org*, 2013: p. I–115–I–123.
28 Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
29 Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. New York, NY: Springer New York, 2009.
30 Niculescu-Mizil A, Caruana R. *Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning - ICML '05*. New York, USA: ACM Press, 2005: 625–32.
31 Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif* 2000;10.
32 Robertson M, Majumdar A, Boyapati R, *et al*. Risk stratification in acute upper Gi bleeding: comparison of the AIMS65 score with the Glasgow-Blatchford and Rockall scoring systems. *Gastrointest Endosc* 2016;83:1151–60.
33 Futoma J, Simons M, Panch T. The myth of generalizability in clinical research and machine learning in healthcare. *Lancet Digit Heal*. In Press 2020.
34 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.