BMJ Open

# Training machine learning models to predict 30-day mortality in patients discharged from the emergency department: a retrospective, population-based registry study

Mathias Carl Blom,[1] Awais Ashfaq,[2,3] Anita Sant'Anna,[2] Philip D Anderson,[4,5] Markus Lingman[3,6]

**To cite:** Blom MC, Ashfaq A, Sant'Anna A, *et al*. Training machine learning models to predict 30-day mortality in patients discharged from the emergency department: a retrospective, population-based registry study. *BMJ Open* 2019;**9**:e028015. doi:10.1136/bmjopen-2018-028015

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-028015).

Received 18 November 2018
Revised 13 April 2019
Accepted 05 July 2019



For numbered affiliations see end of article.

**Correspondence to**
Markus Lingman;
Markus.Lingman@regionhalland.se

## ABSTRACT

**Objectives** The aim of this work was to train machine learning models to identify patients at end of life with clinically meaningful diagnostic accuracy, using 30-day mortality in patients discharged from the emergency department (ED) as a proxy.

**Design** Retrospective, population-based registry study.

**Setting** Swedish health services.

**Primary and secondary outcome measures** All cause 30-day mortality.

**Methods** Electronic health records (EHRs) and administrative data were used to train six supervised machine learning models to predict all-cause mortality within 30 days in patients discharged from EDs in southern Sweden, Europe.

**Participants** The models were trained using 65 776 ED visits and validated on 55 164 visits from a separate ED to which the models were not exposed during training.

**Results** The outcome occurred in 136 visits (0.21%) in the development set and in 83 visits (0.15%) in the validation set. The model with highest discrimination attained ROC–AUC 0.95 (95% CI 0.93 to 0.96), with sensitivity 0.87 (95% CI 0.80 to 0.93) and specificity 0.86 (0.86 to 0.86) on the validation set.

**Conclusions** Multiple models displayed excellent discrimination on the validation set and outperformed available indexes for short-term mortality prediction interms of ROC–AUC (by indirect comparison). The practical utility of the models increases as the data they were trained on did not require costly de novo collection but were real-world data generated as a by-product of routine care delivery.

## Strengths and limitations of this study

- In this study, we report the performance of supervised machine learning models that were trained on a population-based retrospective real-world material of high completeness with minimal loss to follow-up.
- The models make use of standard data elements readily capturable in many electronic health record systems for training, which we believe facilitates their implementation across systems and reduces susceptibility to institution-specific biases.
- The models were tuned using cross-validation and thereafter validated on an external sample from a site to which they were previously unexposed, improving external validity.
- Prospective validation is needed to fully assess model impact in clinical practice.
- Given the flexibility of machine learning models and the resulting risk of overfitting, models should be retrained if implemented at a new site and periodically when used in clinical practice.

## BACKGROUND

As healthcare costs increase in the USA and across the globe,[1–3] evidence suggests that advances in healthcare technologies and increased utilisation of these technologies are important drivers.[3] While technological advancements may result in improved diagnostics and treatments, the return on investment of healthcare spending in terms of life expectancy has decreased over time.[4] In turn, this questions whether new medical technologies are always used wisely.

The definition of value in healthcare suggests that value is eroded when patients with low probability of benefit are overtreated with risky or costly procedures,[5] potentially causing net harm. The fee-for-service model has been implicated in promoting such value erosion by incentivizing volume and price of care irrespective of its quality.[6] Although randomised trials on the topic are lacking, observational studies of variation in US healthcare spending have failed to show an association between higher spending and better quality of care.[7 8] Rather, higher spending has been associated with poorer care experiences.[9 10] Associations between

more aggressive treatment near end of life (EOL) and poorer quality of life in cancer patients,[11 12] as well as indications that aggressive treatment may not always be in line with patient preferences[13–16] even suggest that patient autonomy may be jeopardised at EOL. We are not aware of firm evidence linking overtreatment to the recently observed fall in US life expectancy.[17]

We argue that the first step in improving EOL care and reducing overtreatment at EOL is to identify terminally ill patients who could benefit from proactive discussions about their preferences in order to reduce the risk of overtreatment. While surrogate decision-making such as advance directives and do not resuscitate orders are already part of clinical practice, previous work indicates that they are used too infrequently and sometimes fail to take patients' preferences into account.[14 18] Buying into the hypothesis that patients who are given an opportunity to communicate their EOL preferences are more likely to receive EOL care that are in line with their preferences,[14 19] we aimed to train supervised machine learning models to identify patients at EOL. Our ambition is that the final models can subsequently be used to systematically identify patients who may benefit from a discussion about EOL care without significantly adding to the workload of healthcare practitioners. We set out to study patients discharged from the emergency department (ED) as this population is both accessible for screening and contain terminally ill patients without clear advance directives, whose conditions deteriorate.

## METHODS

### Study design

The study was conducted as a retrospective, population-based registry study utilising data from a comprehensive healthcare analysis platform in Region Halland, southern Sweden. A consecutive sample of ED visits in the region from 1 January 2015 to 31 December 2016 were included. Data were collected using an analysis platform that connects various sources, including medical (electronic health records, EHR) and administrative data from healthcare providers in the region. Data were linked to the Swedish population register to assess the outcome. All-cause 30-day mortality in patients discharged from the ED was used for the primary outcome as we believe it serves as a reasonable proxy for patients at EOL. Discharged patients were deliberately selected as they largely reflect situations where the attending physician judges that acute inpatient admission is of limited benefit. Visits resulting in admission to inpatient departments or referral to other hospitals on ED discharge were excluded, as well as visits where the patient died in the ED, and visits to the psychiatric ED. No interventions or treatments were administered. The study was approved by The Regional Ethical Review Board in Lund, Dnr 2016/517. Individual informed consent was not requested, but patients were given an opportunity to opt out from participation (12 patients exercised this option). The population of the studied region is 320 000 but expands during summer due to tourism. The Region hosts two separate EDs that are open 24/7.

### Independent variables

The selection of independent variables was conducted a priori and was based on published literature and directed acyclic graphs as agreed on by a committee of physicians, researchers and informaticians. Descriptive statistics for the independent variables are shown in table 1 and variable definitions are available in the online supplementary appendix. The unit of analysis is one ED visit. Complete-case analysis was deployed as the proportion missing values was low.

### Statistical analysis

Six different algorithms were selected for model training, based on their principally different approaches to prediction. These were L2 regularised logistic regression (LR),[20] support vector machine (SVM),[21] K-nearest neighbours(KNN) classifier,[22] boosted gradient trees (AB),[23] random forests (RF)[24] and neural network (MLP).[25] All selected predictors were fed into each of the models. As prediction algorithms assume that training sets have reasonably evenly distributed classes of the outcome, skewed data sets pose risks of biasing the algorithm towards the majority class. To mitigate this, we oversampled the minority class in the development set[26] for KNN to equal proportions. For the other algorithms, we used an embedded cost matrix in the model function that penalised misclassified samples from the minority more than from the majority[27] (proportional to the inverse probability of belonging to the minority class). Despite acknowledging the ongoing debate on reporting standards for rare event classifiers, we chose to optimise models for area under the ROC (ROC–AUC) as it makes for a straightforward comparison to models published by others and is recommended by the authorities for evaluating diagnostic tests.[28] Once the optimal set of hyperparameters was identified through systematic grid-search (using fivefold cross-validation to reduce variance), the performance of each model was evaluated on the validation set. Performance on the development and validation set was compared to assess whether models were overfit or underfit. The development set consisted of visits to one ED in the region and the validation set consisted of visits to another. 95% CIs were obtained by identifying the fifth and 95th percentiles of a probability distribution of each relevant measure, obtained by refitting the final models on bootstrapped samples of the validation set (drawn with replacement over 1000 iterations).[29] For face-validity, the relative importance of each predictor was assessed using the internal estimates of variable importance inherent to the RF algorithm.[24] Continuous variables were normalised before being fed into the models. Observations were designated predicted positive if the predicted probability of

**Table 1** Descriptive statistics

| | Complete data set* n=123 975 | Validation set n=55 164 | Development set n=65 776 | | | |
|---|---|---|---|---|---|---|
| **Variable** | **N missing (%)** | **% exposed†** | **% exposed** | **% experiencing outcome in exposed** | **% experiencing outcome in unexposed** | **P value‡** |
| Female | 0 (0.0) | 49.5 | 49.0 | 0.19 | 0.22 | 0.48 |
| Arrived by ambulance | 0 (0.0)§ | 13.6 | 11.1 | 0.87 | 0.12 | <0.001 |
| Referred by physician | 0 (0.0) | 14.0 | 10.1 | 0.36 | 0.19 | 0.006 |
| Triage priority 1 | 0 (0.0) | 0.8 | 0.9 | 1.48 | 0.19 | <0.001 |
| Triage priority 2 | 0 (0.0) | 13.1 | 14.8 | 0.41 | 0.17 | <0.001 |
| Radiology order in ED | 0 (0.0)¶ | 18.1 | 12.8 | 0.27 | 0.20 | 0.19 |
| Left against medical advice | 0 (0.0) | 5.0 | 5.1 | 0.09 | 0.21 | 0.18 |
| Discharged night-time | 0 (0.0) | 30.4 | 33.5 | 0.18 | 0.22 | 0.36 |
| Discharged weekend | 0 (0.0) | 31.0 | 33.0 | 0.17 | 0.23 | 0.12 |
| Discharged summer | 0 (0.0) | 15.2 | 14.7 | 0.11 | 0.22 | 0.04 |
| Discharged winter | 0 (0.0) | 23.3 | 23.4 | 0.22 | 0.20 | 0.73 |
| Male provider | 3385 (2.73) | 44.2 | 43.9 | 0.24 | 0.18 | 0.09 |
| Junior physician | 3385 (2.73) | 22.5 | 25.2 | 0.25 | 0.19 | 0.22 |
| Non-physician provider | 3385 (2.73) | 7.1 | 14.3 | 0.11 | 0.22 | 0.03 |
| Mortality | 0 (0.0) | 0.15 | 0.21 | N/A | N/A | N/A |
| | | Median (IQR) | Median (IQR) | Median (IQR) in subjects experiencing outcome | Median (IQR) in subjects not experiencing outcome | P** |
| Age (years) | 0 (0.0) | 42.0 (20.0, 66.0) | 31.0 (12.0, 58.0) | 81.0 (71.8, 89.0) | 31.0 (12.0, 58.0) | <0.001 |
| Comorbidity score | 3035 (2.45) | 0.0 (0.0, 0.0) | 0.0 (0.0, 0.0) | 2.0 (1.0, 6.0) | 0.0 (0.0, 0.0) | <0.001 |
| ED census (N) | 0 (0.0) | 29.0 (20.0, 36.0) | 30.0 (22.0, 37.0) | 33.0 (25.0, 39.0) | 30.0 (22.0, 37.0) | 0.02 |
| Hospital bed occupancy (%) | 0 (0.0) | 92.0 (87.8, 96.6) | 89.1 (84.1, 93.5) | 90.1 (83.9, 93.8) | 89.1 (84.1, 93.5) | 0.87 |

*N before excluding missing values.
†Proportion of subjects sharing characteristic indicated in 'variable' column.
‡P-value for difference in outcome, exposed vs unexposed, non-adjusted, development set. Arrived by ambulance, referred by physician, triage priority 1 and 2, discharged summer, non-physician provider with p<0.05.
§Database-linkage between source table and ambulance dispatches for 14 918 (12.0%) subjects.
¶Database-linkage between source table and radiology orders for 18 435 (14.9%) subjects.
**P-value for difference in predictor distribution, subjects experiencing outcome vs subjects not experiencing outcome, non-adjusted, development set. Age, comorbidity score and ED census with p<0.05.
ED, emergency department.

the outcome was ≥50%. Performance was reported as sensitivity and specificity in accordance with STARD[30] and benchmarked across models by comparing 95% CIs. Univariate comparisons were conducted using the Wilcoxon rank sum test for continuous variables and the $\chi^2$ test for indicator variables. Multicollinearity was addressed using Spearman's r. Statistical analyses were undertaken in Python 3.6, scikit-learn 20.0[31] and Keras.[32] Data analysis was conducted by one author (AA) with supervision from MCB and ASA. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis reporting guidelines were used.[33]

## RESULTS

### Descriptive statistics

The development set included 65 776 observations and the validation set 55 164 observations, after excluding 3035 observations with missing information for comorbidity score. Of note, 3385 observations lacked information on provider experience, but as these variables were constructed as indicators, missing values for the source variable were not excluded. See table 2 for a detailed description of the construction of the study cohort. Patients in the validation set were older than patients in the development set and more of them were referred to the ED and subject to radiology orders,

Table 2 Exclusion analysis

| | Change (N) | Cohort size (N) |
|---|---|---|
| All ED visits 2015–2016 in database | N/A | 177 833 |
| Including all ED visits with discharge destination 'home' | +109 745 | 109 745 |
| Including all ED visits with discharge destination 'referred' | +8070 | 117 815 |
| Including all ED visits with discharge destination 'LAMA' | +6644 | 124 459 |
| Excluding ED visits with discharge destination 'admitted to hospital' | −112 | 124 347 |
| Excluding visits to odontology | −339 | 124 008 |
| Excluding ED visits with where patient has unknown gender | −7 | 124 001 |
| Excluding ED visits where patient age is not >0.00 years | −26 | 123 975 |
| Excluding missing values | −3035 | 120 940 |
| Final sample | N/A | 120 940 |

ED, emergency department; LAMA, leave against medical advice; N/A, not applicable.

while fewer of them were cared for by a junior provider (see table 1).

ED census and night-time discharge, along with hospital bed occupancy and weekend discharge, displayed moderate correlations (coefficients −0.46 and −0.52) (see online supplementary figure S1). All models converged and did not indicate multicollinearity.

## Model performance

All models performed excellently on the development set, ranging from ROC–AUC 0.92 (95% CI 0.91 to 0.94) for KNN to 1.00 (1.00 to 1.00) for AB. The substantial decrease in performance of MLP and AB on the validation set indicated overfitting to the development set. The decrease in performance of these two models was driven by sensitivity, that is, an inability to correctly identify cases, which is in line with expectations for imbalanced tasks (ie, the low prevalence of cases incited the models to predict both cases and non-cases as negative). However, ROC–AUC was excellent for the remaining models on the validation set (LR, SVM, RF and KNN), suggesting little or no overfitting to the development set (see table 3 and figure 1). Detailed information about algorithm training is provided in the online supplementary appendix. Final models, source code and instructions are made available on request.

Patient age and comorbidity score displayed the highest relative importance among the independent variables, followed by arriving in the ED by ambulance (see figure 2). These findings are aligned with an expectation that older and comorbid patients are at increased risk of death as well as that arriving by ambulance may indicate a more serious condition. A posthoc sensitivity analysis that was undertaken on the final RF algorithm by retraining it on the top five features only (age, comorbidity score, arrival by ambulance, ED census and hospital bed occupancy, selected based on the mean decrease in Gini impurity) suggested only a small reduction in performance from limiting the number of features (ROC–AUC 0.937, 95% CI 0.922 to 0.949).

Table 3 Algorithm performance (development and validation set)

| | Development set | | | Validation set | | |
|---|---|---|---|---|---|---|
| | ROC–AUC (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) | ROC–AUC (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) |
| KNN | 0.923 (0.907 to 0.937) | 0.856 (0.792 to 0.910) | 0.850 (0.827 to 0.871) | 0.925 (0.904 to 0.941) | 0.891 (0.815 to 0.952) | 0.844 (0.818 to 0.865) |
| SVM | 0.944 (0.931 to 0.956) | 0.921 (0.881 to 0.956) | 0.854 (0.851 to 0.856) | 0.945 (0.933 to 0.956) | 0.869 (0.802 to 0.931) | 0.858 (0.855 to 0.860) |
| MLP | 0.975 (0.967 to 0.979) | 1.00 (0.963 to 1.000) | 0.922 (0.896 to 0.934) | 0.867 (0.828 to 0.905) | 0.500 (0.366 to 0.655) | 0.925 (0.899 to 0.937) |
| RF | 0.962 (0.953 to 0.970) | 0.750 (0.684 to 0.815) | 0.954 (0.950 to 0.958) | 0.934 (0.920 to 0.946) | 0.737 (0.647 to 0.824) | 0.907 (0.902 to 0.912) |
| AB | 1.000 (1.000 to 1.000) | 1.000 (1.000 to 1.000) | 1.000 (1.000 to 1.000) | 0.499 (0.499 to 0.513) | 0.000 (0.000 to 0.027) | 0.999 (0.998 to 0.999) |
| LR | 0.940 (0.926 to 0.953) | 0.714 (0.650 to 0.774) | 0.944 (0.943 to 0.946) | 0.942 (0.928 to 0.954) | 0.890 (0.835 to 0.944) | 0.861 (0.859 to 0.863) |

AB, boosted gradient trees; KNN, K-nearest neighbours; LR, logistic regression; MLP, neural network; RF, random forests; SVM, support vector machine.

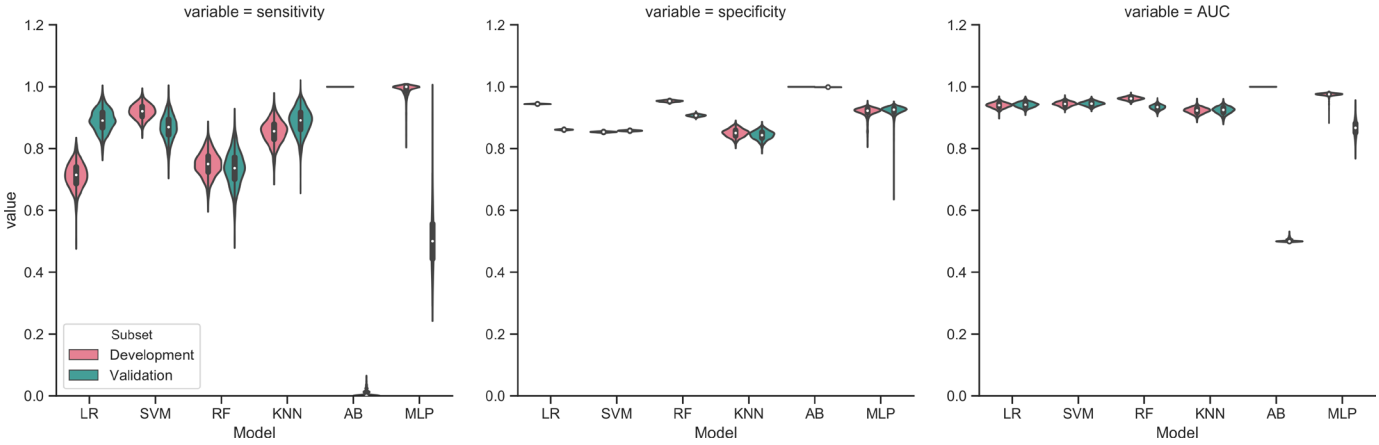


**Figure 1** Algorithm performance (development and validation set). AB, boosted gradient trees; KNN, K-nearest neighbours; LR, logistic regression; MLP, neural network; RF, random forests; SVM, support vector machine.

## DISCUSSION

Four of the machine learning models predicted all-cause 30-day mortality with excellent discrimination on the validation set (ROC–AUC >0.900). This exceeds several previously reported models (by indirect comparison, as clinical data sets are not available), such as ROC–AUC 0.860 of a frequently cited algorithm for short-term mortality prediction proposed by Gagne *et al*[34] as well as ROC–AUC 0.930 of models aimed at identifying patients who may benefit from palliative care proposed by Avati *et al*[35] and an array of models trained on less heterogeneous patient subgroups that exhibit lower class imbalance (ie, higher baseline risk). A non-exhaustive sample of such models include the contributions made by *Miro* (ROC–AUC 0.836),[36] Makar *et al* (ROC–AUC 0.828)[37] and Elfiky *et al* (ROC–AUC 0.940).[38] Additionally, as the models proposed here are trained on data produced as a by-product of routine care delivery, we argue that our contributions are less resource intensive to implement in clinical practice than many traditional risk scores that require costly de novo data collection. Moreover, our models are distinguished by maintaining performance when validated on a distribution that they were unexposed to during training, which contrasts the common approach of validating on a random heterogeneous sample from the training distribution.[35–39]

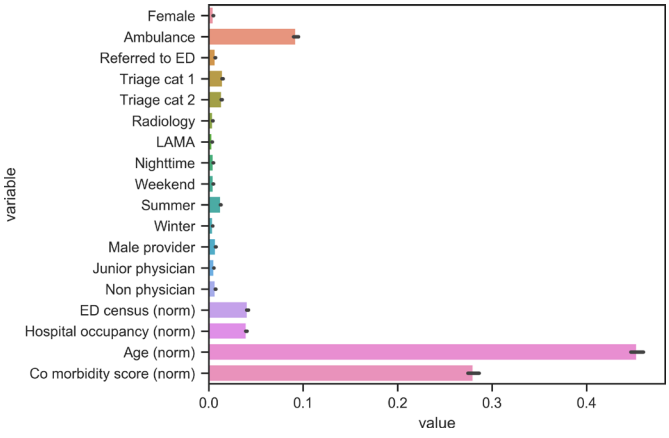


**Figure 2** Variable importance using the RF algorithm. ED, emergency department; LAMA, leave against medical advice; RF, random forests.

Many clinicians recognise the challenges in hosting timely discussions about patients' EOL preferences, which is reflected in findings suggesting that advance care planning often occurs too late or not at all. In turn, we believe this contributes to overtreatment and care that is not in line with patient preferences.[2 40 41] We hope that our models can aid physicians who face such challenges to systematically identify patients at EOL to schedule for more timely planning, without significantly adding to their workload.

While screening healthy populations traditionally demands tests with high specificity, the desired level depends on the scheduled intervention. If the intervention scheduled for patients deemed high-risk by our models is a non-invasive follow-up visit to primary care, we argue that high sensitivity is more relevant than high specificity, as the direct physical risks to the patient are minimal. Depending on the cost of delivering the intervention, individual healthcare systems may want to fine-tune the prediction threshold to achieve a lower false-positive rate (and lower costs of the intervention) at the expense of sensitivity. At the discretion of the primary care physician, a follow-up visit could focus on advance care planning or on an overall evaluation, which likely adds value to the elderly patients with multiple comorbidities that constitute most of the high-risk patients. An evaluation in primary care could also benefit patients who are of high risk of death due to an acute condition that was not correctly identified in the ED. While the latter patient group is not the main focus of this work, the models can be retrained on a refined population to learn identify such erroneous discharges. Using follow-up in primary care as the intervention would also address the suggested benefits of involving primary care in advance care planning.[41] It is already not uncommon to arrange follow-up in primary care after an ED visit, which makes us believe that scheduling patients with high predicted risk of death

for such follow-up after ED discharge fits well within the general process of care. Moreover, an overall risk-assessment is already part of the emergency physician's duties at discharge, which makes automated screening using our models fit well within the ED clinical workflow. While classic risk stratification tools developed in the past have been making use of linear equations that lend themselves well to translation into risk scores that can be retrieved from memory, the flexibility of machine learning models makes such use less straightforward. However, current methods for deploying predictive models in hospital information systems would allow models like these to be accessed through an application interface in healthcare workers' clinical workflow, much like is the case with decision support systems or clinical systems used for placing for example, radiology orders.

While a case has been made in the past for targeting EOL care as a means of reducing overall healthcare spending, recent work has challenged the overall impact of such a strategy[2 39] and we do not expect that implementing our models in clinical practice will prevent accelerating costs of care. Rather, we hope that the models can promote value in healthcare by bringing patients, physicians and families closer to meaningful EOL discussions. Additionally, the scarcity of evidence supporting EOL interventions[42] poses a need for prospective trials, and the models may prove useful as a computable phenotype to identify study subjects for future research.

### Strengths and limitations

One effect of the flexibility allowed by machine learning models is that they may overfit to the characteristics of the development set and therefore not perform similarly across sites.[43] To mitigate this situation, we implemented cross-validation and validated model performance out of sample on data from a separate hospital, that the models were previously unexposed to. Also, the use of standard data-elements routinely captured in most EHR systems makes our models less susceptible to being overfit to the practices of a specific institution, as compared with models that make predictions from a wider array of data elements that tend to be more institution specific (eg, text in EHR notes that may reflect individual physicians' documentation style or biases). As variations in local processes or populations are expected to occur over time, our models should be continuously monitored and periodically retrained to maintain performance when implemented in clinical practice. The inverse-probability weighting scheme maintained in this exercise makes it unlikely that algorithm performance is significantly impacted by retraining on data sets displaying different levels of class-imbalance.

Before deployment, we also suggest that the models are subject to prospective validation across several sites, and to a formal cost–benefit analysis in order to identify associated interventions that are safe, effective and add value. Further customisation of the models is achievable by optimising the decision threshold to produce the most favourable trade-off between false positives and false negatives in any given population, taking into account the characteristics of the intervention scheduled to follow algorithm predictions. Additionally, combining several models into an ensemble predictor for increased flexibility may improve performance further still.

## CONCLUSIONS

In this paper, we report performance of supervised machine learning models that predict 30-day mortality in patients discharged from the ED with excellent discrimination. The models outperform other indexes previously developed for short-term mortality prediction in terms of ROC–AUC (by indirect comparison) without being dependent on costly de novo data collection, which makes them readily implementable in clinical practice.

**Author affiliations**
[1]Department of Clinical Sciences Lund, Medicine, Lund University, Medical Faculty, Lund, Sweden
[2]Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, Sweden
[3]Halland Hospital, Region Halland, Halmstad, Sweden
[4]Department of Emergency Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA
[5]Harvard Medical School, Boston, Massachusetts, USA
[6]Department of Molecular and Clinical Medicine/Cardiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

**Acknowledgements** We wish to acknowledge contributions made to this study by Thomas Wallenfeldt (CGI group Inc) and Ziad Obermeyer, MD. (Brigham and Women's Hospital, Harvard Medical School).

**Contributors** MCB and ML came up with the study idea and drafted the first version of the study protocol. ASA, AA, ML and MCB developed the analysis plan. AA conducted all analyses for the paper with supervision from MCB and ASA. MCB, AA, ASA, PDA and ML provided critical input on the study protocol. MCB, AA, ASA, PDA and ML took part in interpreting preliminary results and drafting the manuscript.

**Funding** This work was partly funded by Region Halland, Sweden. The authors also wish to recognise the Health Technology Center (HCH) and Center for Applied Intelligent Systems Research (CAISR) at Halmstad University for support from the project HiCube - behovsmotiverad halsoinnovation.The initial stage of MCBs involvement in the work was funded by a grant for post-doctoral research from the Tegger Foundation. The funders/sponsors had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation review, or approval of the manuscript and decision to submit the manuscript for publication.

**Competing interests** None declared.

**Patient consent for publication** This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Technical appendix, statistical code and final models available upon request. Individual-level patient data may not and therefore will not be shared.

## REFERENCES


1. Moses H, Matheson DHM, Dorsey ER, *et al*. The anatomy of health care in the United States. *JAMA* 2013;310:1947–63.
2. Aldridge MD, Kelley AS. Epidemiology of serious illness and high utilization of health care. in: Institute of medicine of the National academies. *Dying in America: Improving quality and honoring individual preferences near the end of life. Washington, DC. The National Academies Press* 2015:487–531.
3. Bodenheimer T. High and rising health care costs. Part 2: technologic innovation. *Ann Intern Med* 2005;142:932–7.
4. Cutler DM, Rosen AB, Vijan S. The value of medical spending in the United States, 1960-2000. *N Engl J Med* 2006;355:920–7.
5. Porter ME. What is value in health care? *N Engl J Med* 2010;363:2477–81.
6. Schroeder SA, Frist W. National Commission on physician payment reform. phasing out fee-for-service payment. *N Engl J Med* 2013;368:2029–32.
7. Fisher ES, Wennberg DE, Stukel TA, *et al*. The implications of regional variations in Medicare spending. Part 2: health outcomes and satisfaction with care. *Ann Intern Med* 2003;138:288–99.
8. Yasaitis L, Fisher ES, Skinner JS, *et al*. Hospital quality and intensity of spending: is there an association? *Health Aff* 2009;28(Supplement 1):w566–72.
9. Mittler JN, Landon BE, Fisher ES, *et al*. Market variations in intensity of Medicare service use and beneficiary experiences with care. *Health Serv Res* 2010;45:647–69.
10. Wennberg JE, Bronner K, Skinner JS, *et al*. Inpatient care intensity and patients' ratings of their hospital Experiences: What could explain the fact that Americans with chronic illnesses who receive less hospital care report better hospital experiences? *Health Aff* 2009;28:103–12.
11. Wright AA, Zhang B, Ray A, *et al*. Associations between end-of-life discussions, patient mental health, medical care near death, and caregiver bereavement adjustment. *JAMA* 2008;300:1665–73.
12. Zhang B, Wright AA, Huskamp HA, *et al*. Health care costs in the last week of life: associations with end-of-life conversations. *Arch Intern Med* 2009;169:480–8.
13. Groff AC, Colla CH, Lee TH. Days spent at home — a patient-centered goal and outcome. *N Engl J Med* 2016;375:1610–2.
14. Silveira MJ, Kim SYH, Langa KM. Advance directives and outcomes of surrogate decision making before death. *N Engl J Med* 2010;362:1211–8.
15. Teno JM, Fisher ES, Hamel MB, *et al*. Medical care inconsistent with patients' treatment goals: association with 1-year Medicare resource use and survival. *J Am Geriatr Soc* 2002;50:496–500.
16. Pritchard RS, Fisher ES, Teno JM, *et al*. Influence of patient preferences and local health system characteristics on the place of death. *J Am Geriatr Soc* 1998;46:1242–50.
17. – Murphy SL, Xu J, Kochanek KD, *et al*. Mortality in the United States, 2017. *U.S. Department of Health and Human Services, National Center for Health Statistics* 2018;328.
18. Yuen JK, Reid MC, Fetters MD. Hospital do-not-resuscitate orders: why they have failed and how to fix them. *J Gen Intern Med* 2011;26:791–7.
19. Mack JW, Weeks JC, Wright AA, *et al*. End-Of-Life discussions, goal attainment, and distress at the end of life: predictors and outcomes of receipt of care consistent with preferences. *JCO* 2010;28:1203–8.
20. James G, Witten D, *et al*. Linear Model Selection and Regularization. In: *An introduction to statistical learning with applications in R*. 1th edn. New York: Springer, 2013: 203–64.
21. Hastie T, Tibshirani R, Friedman J. Support Vector Machines and Flexible Discriminants. In: *The elements of statistical learning*. 2th edn. New York: Springer, 2009: 417–58.
22. Hastie T, Tibshirani R, Friedman J. Prototype Methods and Nearest-Neighbors. In: *The elements of statistical learning*. 2th edn. New York: Springer, 2009: 459–84.
23. Hastie T, Tibshirani R, Friedman J, –. Boosting and Additive Trees. In: *The elements of statistical learning*. 2th edn. New York: Springer, 2009: 337–89.
24. Breiman L, Forests R. *Mach Learn* 2001;45:5–32.
25. Networks – Neural. In: Hastie T, Tibshirani R, Friedman J. *Editors. The elements of statistical learning 2ed. New York. NY: Springer* 2009:389–416.
26. Leevy JL, Khoshgoftaar TM, Bauder RA, *et al*. A survey on addressing high-class imbalance in big data. *Journal of Big Data* 2018;5.
27. Ling CX, Sheng VS. Class Imbalance Problem. In: Sammut C, Webb GI, eds. Boston, MA: Encyclopedia of Machine Learning Springer, 2011.
28. Docket No 2003D-0044. *Statistical guidance on reporting results from studies evaluating diagnostic tests*. Rockville, MD: Food and Drug Administration, Center for Devices and Radiological Health, 2007.
29. Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc* 1987;82:171–85.
30. Bossuyt PM, Reitsma JB, Bruns DE, *et al*. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527.
31. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *Journal of machine learning research* 2011;12:2825–30.
32. Chollet F. Keras: deep learning for humans. GitHub, 2015. Available: https://github.com/fchollet/keras
33. Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Circulation* 2015;131:211–9.
34. Gagne JJ, Glynn RJ, Avorn J, *et al*. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J Clin Epidemiol* 2011;64:749–59.
35. Avati A, Jung K, Harman S, *et al*. Improving palliative care with deep learning. *International Conference on Bioinformatics and Biomedicine* 2017:311–6.
36. Miró Òscar, Rossello X, Gil V, *et al*. Predicting 30-day mortality for patients with acute heart failure in the emergency department: a cohort study. *Ann Intern Med* 2017;167:698–705.
37. Makar M, Ghassemi M, Cutler DM, *et al*. Short-Term mortality prediction for elderly patients using Medicare claims data. *Int J Mach Learn Comput* 2015;5:192–7.
38. Elfiky AA, Pany MJ, Parikh RB, *et al*. Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. *JAMA Network Open* 2018;1:e180926.
39. Einav L, Finkelstein A, Mullainathan S, *et al*. Predictive modeling of U.S. health care spending in late life. *Science* 2018;360:1462–5.
40. Connors AF, Dawson NV, Desbiens NA, *et al*. For the support principal Investigators. A controlled trial to improve care for seriously ill hospitalized patients: the study to understand prognoses and preferences for outcomes and risks of treatments (support). *JAMA* 1995;274:1591–8.
41. Lynn J, DeVries KO, Arkes HR, *et al*. Ineffectiveness of the support intervention: review of explanations. *J Am Geriatr Soc* 2000;48:S206–S213.
42. Halpern SD. Toward evidence-based end-of-life care. *N Engl J Med* 2015;373:2001–3.
43. Obermeyer Z, Lee TH. Lost in thought — the limits of the human mind and the future of medicine. *N Engl J Med* 2017;377:1209–11.