BMJ Open

Developing learning algorithms to predict 30-day mortality in patients discharged from the Emergency Department

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-028015
Article Type:	Research
Date Submitted by the Author:	18-Nov-2018
Complete List of Authors:	Blom, Mathias; Lund University Medical Faculty, Department of Clinical Sciences Lund, Medicine Ashfaq, Awais; Halmstad University, Center for Applied Intelligent Systems Research (CAISR) Sant'Anna, Anita; Halmstad University, Center for Applied Intelligent Systems Research (CAISR) Anderson, Philip; Brigham & Women's Hospital, Department of Emergency Medicine; Harvard Medical School Lingman, Markus; Sahlgrenska Academy, University of Gothenburg, Department of molecular and clinical Medicine/Cardiology; Halland Hospital, Region Halland
Keywords:	ACCIDENT & EMERGENCY MEDICINE, BIOTECHNOLOGY & BIOINFORMATICS, EPIDEMIOLOGY, PALLIATIVE CARE, PUBLIC HEALTH

Developing learning algorithms to predict 30-day mortality in patients discharged from the Emergency Department

Mathias C. Blom M.D. Ph.D.[1], Awais Ashfaq M.Sc.[2], Anita Sant'Anna Ph.D.[2], Philip D. Anderson M.D.[3], Markus Lingman M.D. Ph.D[4].

Corresponding author:

Lingman, Markus, M.D. Ph.D.

Markus.Lingman@regionhalland.se

[1] – Department of Clinical Sciences Lund, Medicine, Lund University, Lund, Sweden.

[2] – Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, Sweden.

[3] – Department of Emergency Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, US.

[4] – Institute of Medicine, Dept of Molecular and Clinical Medicine/Cardiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. Halland Hospital, Region Halland, Sweden

Keywords: Emergency Medicine, Mortality, Machine Learning, Advance Care Planning,

Word count: 2,283 (excluding abstract, figures, tables and legends)

Abstract

Objectives: The aim of this work was to develop predictive algorithms for identifying patients at end of life (EOL) 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 different supervised learning algorithms to predict all-cause mortality within 30 days in patients discharged from EDs in southern Sweden, Europe.

Participants: Algorithms were developed using 65,776 visits and validated on 55,164 visits from a separate ED to which the algorithms were not exposed during development.

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

Conclusions: Multiple algorithms displayed excellent discrimination on the validation set and outperformed available indexes for short-term mortality prediction. The practical utility of the algorithms increases as the required data were captured electronically and did not require *de novo* data collection.

Article summary

Strengths and limitations of this study

- In this study, we report the performance of supervised learning algorithms that were developed on a population-based retrospective material of high completeness with minimal loss to follow-up.
- The algorithms developed make use of standard data elements, which we believe facilitates their implementation across systems and reduces susceptibility to institution-specific biases.
- The algorithms were developed using cross-validation and thereafter validated on an external sample from a site to which the algorithms were unexposed during development, improving external validity.
- Prospective validation is needed to fully assess algorithm performance in clinical practice.
- Given the flexibility of machine learning algorithms and the resulting risk of overfitting, algorithms should be periodically retrained if implemented in clinical practice.

Background

Research suggests increasing healthcare costs in the U.S. and across the globe [1-3]. Implicated drivers include the elderly, patients with complex co-morbidities and functional limitations [1-2], as well technological advancements that increase the ambitions of care [3]. Although technological breakthroughs may result in improved diagnostics and treatments, trends indicate that the marginal benefit of healthcare spending may decrease over time [4], which questions whether interventions are always used wisely. Given that value in healthcare is defined in terms of both quality and costs, value may be eroded when patients with low probability of benefit are subjected to risky or costly procedures [5].

The fee-for-service model has been implicated in promoting such erosion by incentivizing volume and price irrespective of quality [6] and although randomized trials are lacking, observational studies of variation in U.S. 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 EOL and poorer quality of life in cancer patients [11-12], as well as indications that aggressive treatment may not be in line with patient preferences [13-16] suggest that patient autonomy is sometimes jeopardized at EOL and that an unmet medical need for advance care planning exists.

We argue that the first step in improving EOL care is to identify relevant patients, and therefore aimed to develop diagnostic supervised learning algorithms to identify patients at EOL in a source population that is readily accessible for screening. Given that the Emergency Department (ED) interfaces with multiple functions in a healthcare system and is strategically located in the

process of care, we propose that ED patients constitute a relevant source population. A second objective is to validate the algorithms out of sample on retrospectively collected data.

Methods

Study Design

The study was conducted as a retrospective, population-based registry study utilizing data from a comprehensive healthcare analysis platform in Region Halland, southern Sweden. A consecutive sample of ED visits in the region from Jan 01 2015 to Dec 31 2016 were included. All-cause 30-day mortality in patients discharged from the ED was used as a proxy for EOL (primary outcome). Discharged patients were deliberately selected as they largely reflect situations where acute inpatient admission is of limited benefit. Visits resulting in admission to inpatient departments or referral to other hospitals upon 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). Data analysis was conducted by one author (A.A.). 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. 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.

Independent variables

The selection of independent variables was conducted *a priori* and was based on published literature and directed acyclic graphs as agreed upon by a committee of physicians, researchers and informaticians. Descriptive statistics for the independent variables are shown in Table 2 and variable definitions are available in the supplementary appendix. The unit of analysis is one ED visit. Complete-case analysis was deployed as the proportion missing values was very low.

Statistical analysis

Six different algorithms were selected for training, based on their principally different approaches to prediction. These were L2 regularized logistic regression (LR) [17], support vector machine (SVM) [18], K-nearest neighbors classifier (KNN) [19], boosted gradient trees (AB) [20], Random Forest™ (RF) [21] and Neural Network (MLP) [22]. Although commonly used in the machine learning literature and potentially boosting performance, our desire to facilitate interpretability made us refrain from combining the palette of models into one ensemble. All selected predictors were fed into each of the algorithms. As prediction algorithms assume that training sets have evenly distributed classes of the outcome, skewed datasets pose risks of biasing the algorithm towards the majority class. To mitigate this, we over-sampled the minority class in the development set [23] for KNN to equal proportions. For the other algorithms, we used an embedded cost matrix in the model function that penalized misclassified samples from the minority more than from the majority [24] (proportional to the inverse probability of belonging to the minority). Algorithms were optimized for area under the ROC-curve (AUC-ROC) as it reflects the trade-off between sensitivity and specificity and is recommended for evaluating diagnostic tests [25]. Once the optimal set of hyper-parameters was identified through systematic search (using 5-fold cross validation to reduce variance), the performance of each

algorithm was evaluated on the validation set. Performance on development and validation set was used to assess whether models were over- or underfit. The development set consisted of visits to one ED in the region and the validation set consisted of visits to another. 95% CI:s were obtained using the bootstrap [26]. For face-validity, the relative importance of each predictor was assessed using the RF algorithm [21]. Continuous variables were normalized before being fed into the algorithms. If the predicted probability of the outcome was ≥50%, the observation was designated as predicted positive. Performance was reported as sensitivity and specificity in accordance with STARD [27] and benchmarked across algorithms by comparing 95% CI:s. Univariate comparisons were conducted using the Wilcoxon rank sum test for continuous variables and the chi2 test for indicator variables. Multicollinearity was addressed using Spearman's rho. Statistical analyses were undertaken in Python™ 3.6, scikit-learn 20.0 [28] and Keras [29].

Results

Descriptive statistics

The development set included 65,776 observations and the validation set 55,164 observations, after excluding 3,035 observations with missing information for co-morbidity score. 3,385 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 1 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, while fewer of them were cared for by a junior provider (see Table 2).

ED census and nighttime discharge, along with hospital bed occupancy and weekend discharge, displayed moderate correlations (coefficients -0.46 and -0.52) (see Figure S1). All algorithms converged and did not indicate multicollinearity.

Model performance

All algorithms performed excellent on the development set, ranging from ROC-AUC 0.92 (95% CI 0.91, 0.94) for KNN to 1.00 (1.00, 1.00) for AB. The substantial decrease in performance of MLP and AB on the validation set indicated that they were overfit to the development set. The decrease in performance of these two algorithms was driven by sensitivity, i.e. an inability to correctly identify cases, which is in line with expectations for imbalanced tasks. However, ROC-AUC was excellent for the remaining algorithms on the validation set (LR, SVM, RF, 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 supplementary appendix. Final models, source code and instructions are made available upon request.

Patient age and co-morbidity 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 co-morbid patients are at increased risk of death as well as that arriving by ambulance may indicate a more serious condition.

Discussion

Four of the learning algorithms predicted all-cause 30-day mortality in patients discharged to home from the ED, with excellent discrimination on the validation set. They outperform popular

algorithms for short-term mortality prediction [30] as well as specific algorithms used in the clinic, which require costly *de novo* data collection [31] and other machine learning algorithms aimed at identifying patients who may benefit from palliative care [32]. Our algorithms also outperform algorithms developed in select patient subgroups that exhibit higher baseline risk [33-35]. Apart from achieving excellent performance in a broad study population, our algorithms stand out by reaching this performance when validated on a distribution that the algorithms were unexposed to during development, in contrast to validation on a random subsample from the distribution used for development.

Most clinicians recognize the challenges in making accurate bedside predictions about the timing of death, 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 care that is not in line with patient preferences [2,36-37]. With sensitivity close to 90%, our KNN, SVM and LR algorithms can help physicians systematically identify a clinically meaningful proportion of patients at EOL with negligible direct risks to the patient. In contrast, RF displayed higher specificity at the expense of somewhat lower sensitivity, thereby limiting the false positive rate (FPR).

While screening of healthy populations traditionally demands tests with high specificity, its absolute level depends on the scheduled intervention. If the intervention scheduled for patients deemed high-risk by our algorithms is a 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 FPR (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 co-morbidities that constitute most of the high-risk patients. An evaluation in primary care could also benefit false positives that result from patients at high risk of death due to an acute condition, that have been discharged from the ED erroneously. Using follow-up in primary care as the intervention would also address the importance in involving primary care in advance care planning [37]. It is already not uncommon to arrange follow-up in primary care after an ED visit, which makes us believe that scheduling predicted positives for such follow-up after discharge from the ED 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 algorithms fit well with the ED workflow as well.

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 [2,38] and we do not expect that implementing our algorithms in clinical practice will prevent accelerating costs of care. Rather, we hope that the algorithms can promote value in healthcare by helping patients and families make more informed decisions about interventions that come with significant side-effects. In addition, as the scarcity of evidence supporting EOL interventions [39] poses a need for prospective trials, the algorithms 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 algorithms is that they may overfit to the characteristics of the development set and therefore not perform similarly across sites [40]. To mitigate this situation, we implemented cross-validation and assessed algorithm performance on data from a separate hospital, that the algorithms were previously unexposed to. Also, the use of standard data-elements makes our algorithms less susceptible to being overfit to the practices of a specific institution, as compared to algorithms that make predictions from a wider array of data elements that tend to be more institution specific (e.g. text in EHR notes etc.). As variations in local processes or populations are expected to occur over time, our algorithms should be continuously monitored and periodically retrained to maintain performance if implemented in clinical practice. We also suggest that our algorithms be 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.

Conclusions

In this paper we report performance of supervised learning algorithms, that predict 30-day mortality in patients discharged from the Emergency Department with excellent discrimination. The algorithms outperform other indexes previously developed for short-term mortality prediction without being dependent on costly *de novo* data collection, which makes them readily implementable in clinical practice. Moreover, this is accomplished with data from an external validation site to which the algorithms were previously not exposed. Although a multitude of uses are possible, we propose that the main utility of the algorithms is to identify patients in scope of advance care planning, to ensure that end of life care is in line with patient preferences.

References

1 – Moses H 3rd, Matheson DHM, Dorsey ER, George BP, Sadoff D, Yoshimura S. The anatomy of health care in the United States. JAMA 2013;310(18):1947-1963.

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;142:932-937.

4 – Cutler DM, Rosen AB, Vijan S. The Value of Medical Spending in the United States, 1960-2000. N Engl J Med 2006;355(9):920-927

5 – Porter ME. What Is Value in Health Care? N Engl J Med 2010;363(26):2477-2481.

6 – Schroeder SA, Frist W, National Commission on Physician Payment Reform. Phasing Out Fee-for-Service Payment. N Engl J Med 2013;368(21):2029-2032.

7 – Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder ÉL. The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care. Ann Intern Med, 2003;138(4):288-299.

8 – Yasaitis L, Fisher ES, Skinner JS, Chandra A. Hospital quality and intensity of spending: is there an association?. Health Aff (Millwood) 2009;28(4):w566-w572.

9 – Mittler JN, Landon BE, Fisher ES, Cleary PD., Zaslavsky AM. Market variations in intensity of Medicare service use and beneficiary experiences with care. Health services research 2010;45(3): 647-669.

10 – Wennberg JE, Bronner K, Skinner JS, Fisher ES, Goodman DC. 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 (Millwood) 2009;28(1):103-112.

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(14):1665-1673.

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(5):480-488.

13 – Groff AC, Colla CH, Lee TH. Days spent at home—a patient-centered goal and outcome. N Engl J Med 2016;375(17):1610-1612.

14 – Silveira MJ, Kim SY, Langa KM. Advance directives and outcomes of surrogate decision making before death. N Engl J Med 2010;362(13):1211-1218.

15 – Teno JM, Fisher ES, Hamel MB, Coppola K, Dawson NV. Medical care inconsistent with patients' treatment goals: Association with 1-year Medicare resource use and survival. JAGS 2002;50(3):496-500.

16 – Pritchard RS, Fisher ES, Teno JM, et al, for the SUPPORT Investigators. Influence of patient preferences and local health system characteristics on the place of death. JAGS 1998;46(10):1242-1250.

17 – Linear Model Selection and Regularization. In: James G, Witten D, Hastie T, Tibshirani R. Editors. An introduction to Statistical Learning with applications in R. 1ed. New York. NY: Springer. 2013. 203-264.

18 – Support Vector Machines and Flexible Discriminants. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 417-458.

19 – Prototype Methods and Nearest-Neighbors. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 459-484.

20 – Boosting and Additive Trees. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 337-389.

21 – Breiman L. Random Forests. Machine learning 2001;45(1):5-32.

22 – Neural Networks. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 389-416.

23 – Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 2002;16:321-357.

24 – Wang H, Cui Z, Chen Y, Avidan M, Abdallah AB, Kronzer A. Predicting Hospital Readmission via Cost-sensitive Deep Learning. IEEE/ACM Transactions on Computational Biology and Bioinformatics. April 2018.

25 – Statistical guidance on reporting results from studies evaluating diagnostic tests. Rockville, MD: Food and Drug Administration, Center for Devices and Radiological Health. 2007. (Docket No. 2003D-0044)

26 – Efron B. Better bootstrap confidence intervals. Journal of the American statistical Association 1987;82(397):171-185.

27 – Bossuyt PM, Reitsma JB, Bruns DE, et al, for the STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.

28 – Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research 2011;12: 2825-2830.

29 – Chollet F. Keras: Deep Learning for humans. GitHub, 2015. (https://github.com/fchollet/keras)

30 – Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss SA. combined comorbidity score predicted mortality in elderly patients better than existing scores. J Clin Epidemiol 2011;64(7), 749-759.

31 – Dunn W, Jamil LH, Brown LS, et al. MELD accurately predicts mortality in patients with alcoholic hepatitis. Hepatology 2005;41(2):353-358.

32 – Avati A, Jung K, Harman S, Downing L, Ng A, Shah N. Improving Palliative Care with Deep Learning. International Conference on Bioinformatics and Biomedicine (BIBM), 2017. pp. 311-316.

33 – Miró Ò, 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(10):698-705.

34 – Makar M, Ghassemi M, Cutler DM, Obermeyer Z. Short-term mortality prediction for elderly patients using Medicare claims data. Int J Mach Learn Comput 2015;*5*(3):192-197.

35 – Elfiky A, Pany M, Parikh R, Obermeyer Z. 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(3):e180926.

36 – 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(20):1591-1598.

37 – Lynn J, DeVries KO, Arkes HR, et al. Ineffectiveness of the SUPPORT Intervention: Review of Explanations. JAGS 2000;48(5):S206-S213.

38 – Einav L, Finkelstein A, Mullainathan S, Obermeyer Z. Predictive modeling of US health care spending in late life. Science 2018;360:1462-1465.

39 – Halpern SD. Toward evidence-based end-of-life care. N Engl J Med 2015;373(21):2001-2003.

40 – Obermeyer Z, Lee TH. Lost in thought—the limits of the human mind and the future of medicine. N Engl J Med 2017;377(13):1209-1211.

Acknowledgements

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

Funding

This work was partly funded by Region Halland and Halmstad University, Sweden. 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.

Conflicts of interests

We have read and understood BMJ policy on declaration of interests and declare that we have no competing interests.

Author contributions

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

Patient involvement

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.

Data statement

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

Tables

Table 1: 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”	+8,070	117,815
Including all ED visits with discharge destination “LAMA”	+6,644	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	-3,035	120,940
Final study cohort	N/A	120,940

Table 2: Descriptive statistics

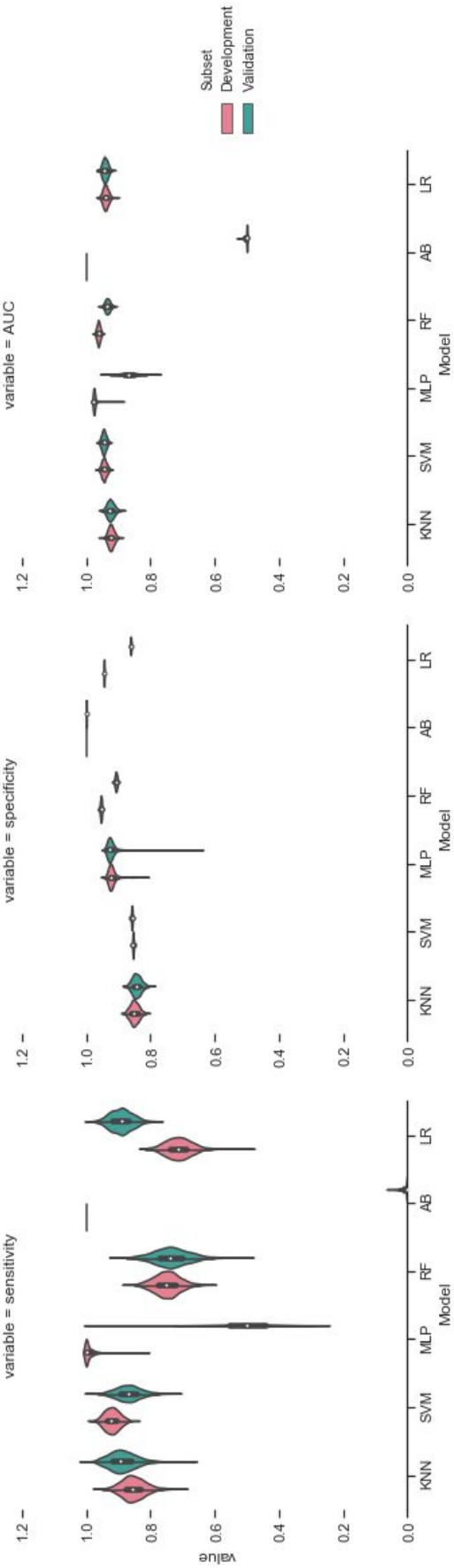
	Complete dataset[1] n=123,975	**Validation set** n=55,164	**Development set** n=65,776			
Variable	N missing (%)	% exposed	% exposed	% experiencing outcome in exposed	% experiencing outcome in unexposed	P[4]
Female	0 (0.0)	49.5	49.0	0.19	0.22	0.48
Arrived by ambulance	0 (0.0)[2]	13.6	11.1	0.87	0.12	<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)[3]	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 nighttime	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	3,385 (2.73)	44.2	43.9	0.24	0.18	0.09
Junior physician	3,385 (2.73)	22.5	25.2	0.25	0.19	0.22
Non-physician provider	3,385	7.1	14.3	0.11	0.22	0.03

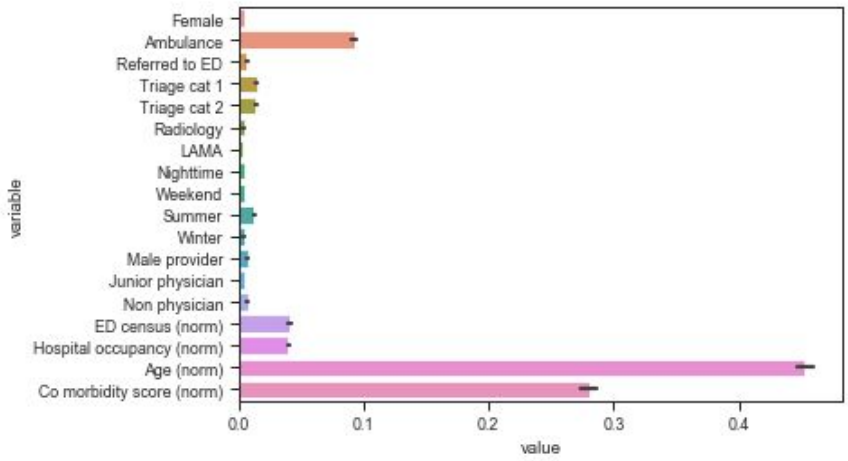
	(2.73)					
Mortality	0 (0.0)	0.15	0.21	N/A	N/A	N/A
		Median (IQR)	Median (IQR)	Median (IQR) in subjects experiencing outcome	Median (IQR) in subjects not experiencing outcome	P[5]
Age [years]	0 (0.0)	42.0 (20.0, 66.0)	31.0 (12.0, 58.0)	81.0 (71.8, 89.0)	31.0 (12.0, 58.0)	<0.001
Co-morbidity score	3,035 (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

[1] N before excluding missing values
[2] Database-linkage between source table and ambulance dispatches for 14,918 (12.0%) subjects
[3] Database-linkage between source table and radiology orders for 18,435 (14.9%) subjects
[4] P-value for difference in outcome, exposed vs unexposed, non-adjusted, development set. Arrived by ambulance, referred by physician, triage priority 1 & 2, discharged summer, non-physician provider with P<0.05.
[5] P-value for difference in predictor distribution, subjects experiencing outcome vs subjects not experiencing outcome, non-adjusted, development set. Age, Co-morbidity score and ED census with P<0.05.

Table 3: Algorithm performance (development and test set)

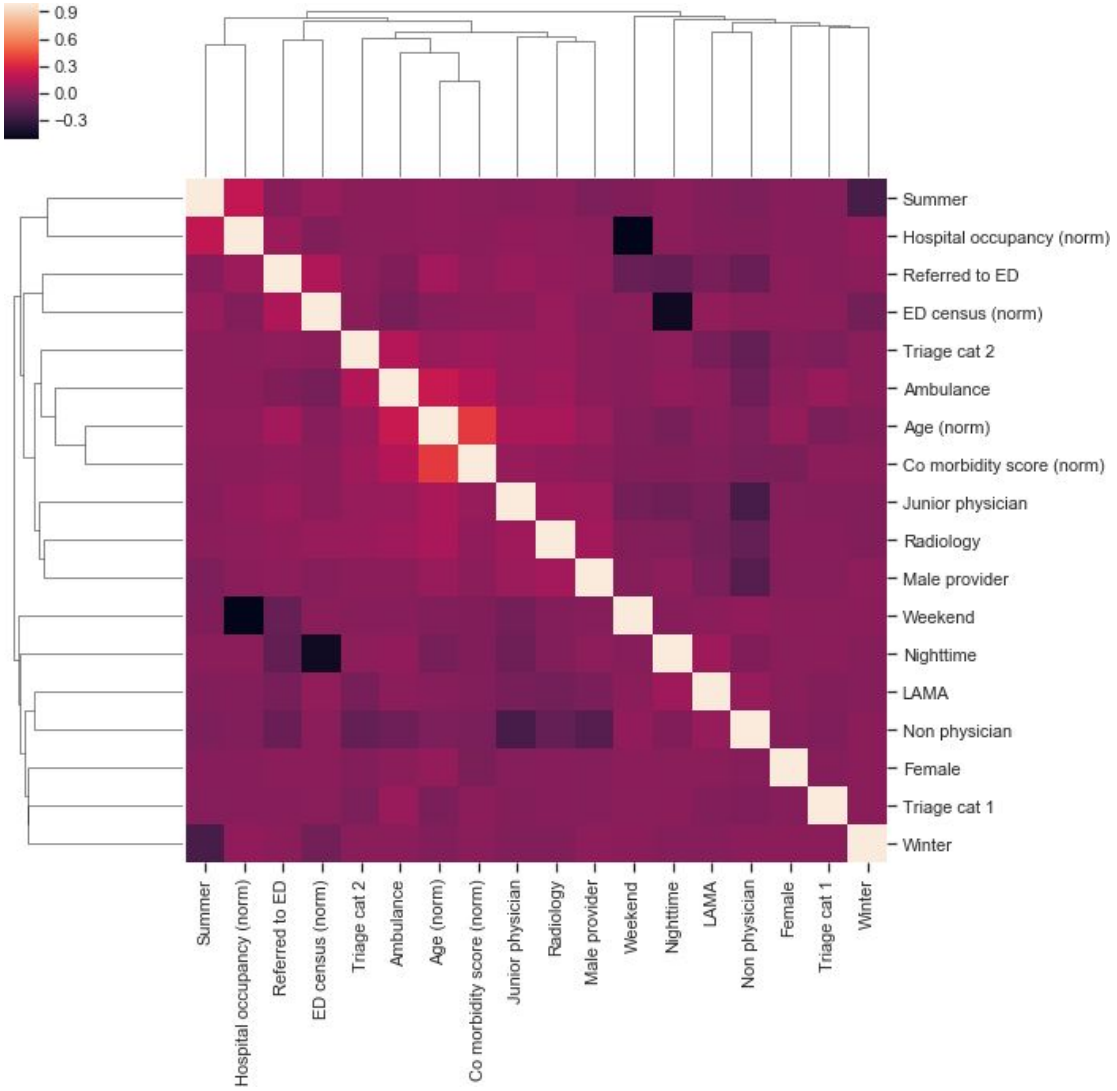
	Development set			Validation set		
	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
KNN	0.923 (0.907, 0.937)	0.856 (0.792, 0.910)	0.850 (0.827, 0.871)	0.925 (0.904, 0.941)	0.891 (0.815, 0.952)	0.844 (0.818, 0.865)
SVM	0.944 (0.931, 0.956)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





Supplementary figures

Figure S1: Correlation matrix of predictors



Correlation coefficients (range -1, 1) for independent variables.

Supplementary Appendix

Construction of independent variables

Individual level Electronic Health Record (EHR) data from all ED visits in Region Halland during the period Jan 01 2015 to Dec 31 2016 were linked to records on inpatient visits, ambulance referrals and radiology orders. All tables were accessed through a recently constructed healthcare analytics platform, in Microsoft SQL Server 2014. Inpatient visits were linked to ED visits by unique personal identifiers derived from a subject's national Personal Identification Number (PIN) and a time criterion (inpatient registration +-3h of ED discharge), as were ambulance referrals (ambulance arrival +- 15min of ED arrival). Hospital bed occupancy was linked by date and facility (variable measured at 06.00am). ED census was linked by date, hour and facility. Remaining tables were linked on unique personal identifiers. The final selection of independent variables comprised patient age, gender, the Quan-Deyo modification of the Charlson Comorbidity Index [1], being referred to the ED by a physician, being transported to the ED in ambulance, perceived urgent medical condition (ED triage system 'RETTS' level 1-2 upon ED arrival [2]), radiology order occurring during the ED visit, leaving the ED against medical advice (LAMA), being discharged during on-call hours (10pm – 7am), during a holiday (including weekends), winter (Dec-Feb, roughly coherent with the influenza season), or summer (week 26-32, corresponding to Swedish vacation period). The co-morbidity score was calculated by linking all individual unique patient identifiers in the study population to all diagnosis data (ICD-10) registered in the healthcare analytics platform. The start of the diagnosis assessment period was set at 365.25 days before the first possible visit (i.e. before 00:00 Jan 1, 2015) and assessment continued throughout the study period. Hence, each individual visit was linked to any diagnoses for the patient registered throughout the region, from the start of the assessment period

up until the individual visit discharge timestamp. Diagnoses were mapped to the relevant co-morbidities in the R package 'icd' [3] (version 3.4.0). The LAMA variable was defined using mandatory input fields that are filled by ED nurses at patient departure.

Construction of the study endpoint

The outcome was assessed by linking records to the Swedish population register. Registering a 'notification of death' (dödsbevis) is a legal obligation in Sweden and must be completed before burial can be authorized. The notification of death is filled in and submitted by the diagnosing physician. As deaths are registered with a resolution of date, any deaths occurring on the date of the ED visit were considered inpatient deaths and therefore excluded. Although the registry should capture deaths in Swedish citizens, some loss to follow-up could result from non-Swedish residents (particularly common during summer).

Algorithm hyperparameter tuning

LR [4]

class sklearn.linear_model.**LogisticRegression**(*penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='warn', max_iter=100, multi_class='warn', verbose=0, warm_start=False, n_jobs=None*)

Optimized for C:[1e-6 – 0.25]

Optimal C: 0.015

Class_weight=Balanced

RF [5]

class sklearn.ensemble.**RandomForestClassifier**(*n_estimators='warn', criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None*)

Optimized for n_estimators: [40 – 200]

Optimized for max_depth: [5 – 25]

Optimal n_estimators: 120

Optimal max_depth: 5

Class_weight=balanced

AB [6]

class sklearn.ensemble.**AdaBoostClassifier**(*base_estimator=None, n_estimators=50, learning_rate=1.0, algorithm='SAMME.R', random_state=None*)

Optimized for base_estimator: [gini, entropy]

Optimized for learninig_rate: [0.1 – 2]

Optimized for n_estimators: [5 – 100]

Optimal base_estimators: gini

Optimal n_estimators: 65

Optimal learning_rate: 0.7

Class_weight=balanced

SVM [7]

class sklearn.svm.**SVC**(*C=1.0*, *kernel='rbf'*, *degree=3*, *gamma='auto_deprecated'*, *coef0=0.0*, *shrinking=True*, *probability=False*, *tol=0.001*, *cache_size=200*, *class_weight=None*, *verbose=False*, *max_iter=-1*, *decision_function_shape='ovr'*, *random_state=None*)

Optimized for C: [0.001 – 1]

Optimized for kernel: [rbf, poly]

Optimal C: 0.01

Optimal kernel: rbf

Class_weight=balanced

KNN [8]

Class sklearn.neighbors.**KNeighborsClassifier**(*n_neighbors=5*, *weights='uniform'*, *algorithm='auto'*, *leaf_size=30*, *p=2*, *metric='minkowski'*, *metric_params=None*, *n_jobs=None*, ***kwargs*)

Optimized for n_neighbors: [1 – 31]

Optimized for metric: [eucledian, minkowski]

Optimal neighbors: 11

Optimal metric: Euclidean

MLP [9]

Epochs = 200

Batch size = 500

Optimizer = rmsprop

Loss = binary cross entropy

Learning rate = 0.01

Activation functions = sigmoid

Optimized for Number of nodes in hidden layer: [5 – 15]

Optimal nodes: 9

References

1 – Quan H, Sundararajan V, Halfon P, et al. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data. Med Care Care 2005;43:1130-1139.

2 – Widgren B. RETTS: Akutsjukvård direkt. 1 ed. Lund, Sweden: Studentlitteratur, 2012.

3 – R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018: URL http://www.R-project.org/.

4 – sklearn.linear_model.LogisticRegression in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

5 – sklearn.ensemble.RandomForestClassifier in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

6 – sklearn.ensemble.AdaBoostClassifier in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html)

7 – sklearn.svm.SVC in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC)

8 – sklearn.neighbors.KNeighborsClassifier in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html)

9 – Keras: Deep Learning for humans. Chollet, F. 2015. Keras, GitHub. (https://github.com/fchollet/keras)

Reporting checklist for prediction model development and validation study.

Based on the TRIPOD guidelines.

Instructions to authors

Complete this checklist by entering the page numbers from your manuscript where readers will find each of the items listed below.

Your article may not currently address all the items on the checklist. Please modify your text to include the missing information. If you are certain that an item does not apply, please write "n/a" and provide a short explanation.

Upload your completed checklist as an extra file when you submit to a journal.

In your methods section, say that you used the TRIPOD reporting guidelines, and cite them as:

Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement.

		Reporting Item	Page Number
	#1	Identify the study as developing and / or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
	#2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
	#3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	3
	#3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	3
Source of data	#4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	4

	#4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	4
Participants	#5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5
	#5b	Describe eligibility criteria for participants.	4
	#5c	Give details of treatments received, if relevant	n/a No treatments administered
Outcome	#6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	4
	#6b	Report any actions to blind assessment of the outcome to be predicted.	n/a Outcome assessed at aggregate-level only
Predictors	#7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured	5
	#7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	n/a Assessed at aggregate-level only
Sample size	#8	Explain how the study size was arrived at.	4
Missing data	#9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	5
Statistical analysis methods	#10a	If you are developing a prediction model describe how predictors were handled in the analyses.	5
	#10b	If you are developing a prediction model, specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	5

	#10c	If you are validating a prediction model, describe how the predictions were calculated.	6
	#10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	5-6
	#10e	If you are validating a prediction model, describe any model updating (e.g., recalibration) arising from the validation, if done	5
Risk groups	#11	Provide details on how risk groups were created, if done.	n/a No risk-groups were created
Development vs. validation	#12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	6
Participants	#13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	See note 1
	#13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	See note 2
	#13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	See note 3
Model development	#14a	If developing a model, specify the number of participants and outcome events in each analysis.	See note 4
	#14b	If developing a model, report the unadjusted association, if calculated between each candidate predictor and outcome.	See note 5
Model specification	#15a	If developing a model, present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	n/a Provided upon request
	#15b	If developing a prediction model, explain how to the use it.	7

Model performance	#16	Report performance measures (with CIs) for the prediction model.	See note 6
Model-updating	#17	If validating a model, report the results from any model updating, if done (i.e., model specification, model performance).	n/a Models not updated after validation
Limitations	#18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	8-9
Interpretation	#19a	For validation, discuss the results with reference to performance in the development data, and any other validation data	7
	#19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	7-10
Implications	#20	Discuss the potential clinical use of the model and implications for future research	8-10
Supplementary information	#21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	5,7,25
Funding	#22	Give the source of funding and the role of the funders for the present study.	25

Author notes

1. 6 (ref table 1, 2)
2. 6 (ref table 2)
3. 6 (ref table 2)
4. 6 (ref table 2)
5. 6 (ref table 2)
6. 7 (ref table 3)

This checklist was completed on 18. November 2018 using http://www.goodreports.org/, a tool made by the EQUATOR Network in collaboration with Penelope.ai

BMJ Open

Training learning algorithms to predict 30-day mortality in patients discharged from the Emergency Department

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-028015.R1
Article Type:	Research
Date Submitted by the Author:	17-Feb-2019
Complete List of Authors:	Blom, Mathias; Lund University Medical Faculty, Department of Clinical Sciences Lund, Medicine Ashfaq, Awais; Halmstad University, Center for Applied Intelligent Systems Research (CAISR); Halland Hospital, Region Halland Sant'Anna, Anita; Halmstad University, Center for Applied Intelligent Systems Research (CAISR) Anderson, Philip; Brigham & Women's Hospital, Department of Emergency Medicine; Harvard Medical School Lingman, Markus; Sahlgrenska Academy, University of Gothenburg, Department of molecular and clinical Medicine/Cardiology; Halland Hospital, Region Halland
Primary Subject Heading:	Emergency medicine
Secondary Subject Heading:	Epidemiology, Palliative care
Keywords:	ACCIDENT & EMERGENCY MEDICINE, BIOTECHNOLOGY & BIOINFORMATICS, EPIDEMIOLOGY, PALLIATIVE CARE, PUBLIC HEALTH

Training learning algorithms to predict 30-day mortality in patients discharged from the Emergency Department

Mathias C. Blom M.D. Ph.D.[1], Awais Ashfaq M.Sc.[2], Anita Sant'Anna Ph.D.[2], Philip D. Anderson M.D.[3], Markus Lingman M.D. Ph.D[4].

Corresponding author:

Lingman, Markus, M.D. Ph.D.

Markus.Lingman@regionhalland.se

[1] – Department of Clinical Sciences Lund, Medicine, Lund University, Lund, Sweden.

[2] – Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, Sweden. Halland Hospital, Region Halland, Sweden

[3] – Department of Emergency Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, US.

[4] – Institute of Medicine, Dept. of Molecular and Clinical Medicine/Cardiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. Halland Hospital, Region Halland, Sweden

Keywords: Emergency Medicine, Mortality, Machine Learning, Advance Care Planning,

Word count: 2,482 (excluding abstract, figures, tables, legends and references)

Abstract

Objectives: The aim of this work was to train predictive algorithms for identifying patients at end of life (EOL) 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 different supervised learning algorithms to predict all-cause mortality within 30 days in patients discharged from EDs in southern Sweden, Europe.

Participants: Algorithms were trained using 65,776 visits and validated on 55,164 visits from a separate ED to which the algorithms 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 algorithm with highest discrimination attained ROC-AUC 0.95 (95% CI 0.93 - 0.96), with sensitivity 0.87 (95% CI 0.80, 0.93) and specificity 0.86 (0.86, 0.86) on the validation set.

Conclusions: Multiple algorithms displayed excellent discrimination on the validation set and outperformed available indexes for short-term mortality prediction in terms of ROC-AUC (by indirect comparison). The practical utility of the algorithms increases as the required data were captured electronically as a by-product of routine care delivery and did not require *de novo* collection.

Article summary

Strengths and limitations of this study

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

Background

Research suggests increasing healthcare costs in the U.S. and across the globe [1-3], with increased ambitions of care being a proposed driver [3]. Although technological advancements may result in improved diagnostics and treatments, trends indicate that the marginal benefit of healthcare spending has decreased over time [4], which questions whether interventions are always used wisely. The value equation states 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 value erosion by incentivizing volume and price irrespective of quality [6] and although randomized trials are lacking, observational studies of variation in U.S. 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 EOL and poorer quality of life in cancer patients [11-12], as well as indications that aggressive treatment may not be in line with patient preferences [13-16] even suggest that patient autonomy may be jeopardized at end of life (EOL). While a case has been made in the popular press, we are still not aware of any firm evidence linking overtreatment to the recently observed decreases in U.S. life expectancy [17].

We argue that the first step in improving EOL care and reducing overtreatment is to identify patients who may benefit from a proactive discussion about EOL preferences, and therefore aimed to train supervised learning algorithms to identify patients at EOL. To make the approach more readily implementable in clinical practice, we set out to study a source population that is both relevant and accessible for screening, and settled on patients visiting the Emergency

Department (ED) because of the strategic position of the latter in the process of care and the heterogeneity of patients that visit the ED.

Methods

Study Design

The study was conducted as a retrospective, population-based registry study utilizing data from a comprehensive healthcare analysis platform in Region Halland, southern Sweden. A consecutive sample of ED visits in the region from Jan 01 2015 to Dec 31 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 as a proxy for EOL (primary outcome). Discharged patients were deliberately selected as they largely reflect situations where acute inpatient admission is of limited benefit. Visits resulting in admission to inpatient departments or referral to other hospitals upon 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 upon 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 supplementary appendix. The unit of analysis is one ED visit. Complete-case analysis was deployed as the proportion missing values was low.

	Complete dataset[1] 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[4]
Female	0 (0.0)	49.5	49.0	0.19	0.22	0.48
Arrived by ambulance	0 (0.0)[2]	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)[3]	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 nighttime	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	3,385 (2.73)	44.2	43.9	0.24	0.18	0.09
Junior physician	3,385 (2.73)	22.5	25.2	0.25	0.19	0.22
Non-physician provider	3,385 (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[5]
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
Co-morbidity score	3,035 (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

Table 1: Descriptive statistics

[1] N before excluding missing values
[2] Database-linkage between source table and ambulance dispatches for 14,918 (12.0%) subjects

[3] Database-linkage between source table and radiology orders for 18,435 (14.9%) subjects
[4] P-value for difference in outcome, exposed vs unexposed, non-adjusted, development set. Arrived by ambulance, referred by physician, triage priority 1 & 2, discharged summer, non-physician provider with P<0.05.
[5] P-value for difference in predictor distribution, subjects experiencing outcome vs subjects not experiencing outcome, non-adjusted, development set. Age, Co-morbidity score and ED census with P<0.05.

Statistical analysis

Six different algorithms were selected for training, based on their principally different approaches to prediction. These were L2 regularized logistic regression (LR) [18], support vector machine (SVM) [19], K-nearest neighbors classifier (KNN) [20], boosted gradient trees (AB) [21], Random Forests™ (RF) [22] and Neural Network (MLP) [23]. All selected predictors were fed into each of the algorithms. As prediction algorithms assume that training sets have evenly distributed classes of the outcome, skewed datasets pose risks of biasing the algorithm towards the majority class. To mitigate this, we over-sampled the minority class in the development set [24] for KNN to equal proportions. For the other algorithms, we used an embedded cost matrix in the model function that penalized misclassified samples from the minority more than from the majority [25] (proportional to the inverse probability of belonging to the minority). Despite acknowledging the ongoing debate on reporting standards for rare event classifiers, we chose to optimize algorithms for area under the ROC-curve (ROC-AUC) as it makes for a straightforward comparison to algorithms published by others and is recommended by the authorities for evaluating diagnostic tests [26]. Once the optimal set of hyper-parameters was identified through systematic search (using 5-fold cross validation to reduce variance), the performance of each algorithm was evaluated on the validation set. Performance on the development and validation set was used to assess whether models were over- or underfit. The development set consisted of visits to one ED in the region and the validation set consisted of visits to another. 95% CI:s were

obtained by identifying the 5th and 95th percentiles of a probability distribution of each relevant measure, obtained by re-fitting the final algorithms on bootstrapped samples of the validation set (drawn with replacement over 1000 iterations) [27]. For face-validity, the relative importance of each predictor was assessed using the internal estimates of variable importance inherent to the RandomForests™ algorithm [22]. Continuous variables were normalized before being fed into the algorithms. Observations were designated predicted positive if the predicted probability of the outcome was ≥50%. Performance was reported as sensitivity and specificity in accordance with STARD [28] and benchmarked across algorithms by comparing 95% CI:s. Univariate comparisons were conducted using the Wilcoxon rank sum test for continuous variables and the chi2 test for indicator variables. Multicollinearity was addressed using Spearman's rho. Statistical analyses were undertaken in Python™ 3.6, scikit-learn 20.0 [29] and Keras [30]. Data analysis was conducted by one author (A.A.) with supervision from MB and ASA. TRIPOD reporting guidelines were used [31].

Results

Descriptive statistics

The development set included 65,776 observations and the validation set 55,164 observations, after excluding 3,035 observations with missing information for co-morbidity score. 3,385 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, while fewer of them were cared for by a junior provider (see Table 1).

ED census and nighttime discharge, along with hospital bed occupancy and weekend discharge, displayed moderate correlations (coefficients -0.46 and -0.52) (see Figure S1). All algorithms converged and did not indicate multicollinearity.

	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"	+8,070	117,815
Including all ED visits with discharge destination "LAMA"	+6,644	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	-3,035	120,940
Final study cohort	N/A	120,940

Table 2: Exclusion analysis

Model performance

All algorithms performed excellent on the development set, ranging from ROC-AUC 0.92 (95% CI 0.91, 0.94) for KNN to 1.00 (1.00, 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 algorithms was driven by sensitivity, i.e. an inability to correctly identify cases, which is in line with expectations for imbalanced tasks (i.e. the low prevalence of cases incited the algorithms to predict both cases and non-cases as negative). However, ROC-AUC was excellent for the remaining algorithms on the validation set (LR, SVM, RF, 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 supplementary appendix. Final models, source code and instructions are made available upon request.

Patient age and co-morbidity 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 co-morbid patients are at increased risk of death as well as that arriving by ambulance may indicate a more serious condition. A post hoc sensitivity analysis that was undertaken on the final RF algorithm by retraining it on the top 5 features only (age, co-morbidity score, arrival by ambulance, ED census and hospital bed occupancy) suggested only a small drop in performance from limiting the number of features (ROC-AUC 0.937, 95% CI 0.922-0.949).

	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, 0.937)	0.856 (0.792, 0.910)	0.850 (0.827, 0.871)	0.925 (0.904, 0.941)	0.891 (0.815, 0.952)	0.844 (0.818, 0.865)

SVM	0.944 (0.931, 0.956)	0.921 (0.881, 0.956)	0.854 (0.851, 0.856)	0.945 (0.933, 0.956)	0.869 (0.802, 0.931)	0.858 (0.855, 0.860)
MLP	0.975 (0.967, 0.979)	1.00 (0.963, 1.000)	0.922 (0.896, 0.934)	0.867 (0.828, 0.905)	0.500 (0.366, 0.655)	0.925 (0.899, 0.937)
RF	0.962 (0.953, 0.970)	0.750 (0.684, 0.815)	0.954 (0.950, 0.958)	0.934 (0.920, 0.946)	0.737 (0.647, 0.824)	0.907 (0.902, 0.912)
AB	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	0.499 (0.499, 0.513)	0.000 (0.000, 0.027)	0.999 (0.998, 0.999)
LR	0.940 (0.926, 0.953)	0.714 (0.650, 0.774)	0.944 (0.943, 0.946)	0.942 (0.928, 0.954)	0.890 (0.835, 0.944)	0.861 (0.859, 0.863)

Table 3: Algorithm performance (development and validation set)

Discussion

Four of the learning algorithms predicted all-cause 30-day mortality with excellent discrimination on the validation set (ROC-AUC > 0.90). This exceeds several previously reported algorithms (by indirect comparison, as clinical datasets are not available), such as ROC-AUC 0.860 of a frequently cited algorithm for short-term mortality prediction proposed by *Gagne et al* [32] as well as ROC-AUC 0.930 of algorithms aimed at identifying patients who may benefit from palliative care proposed by *Avati et al* [33] and an array of algorithms trained on less heterogenous patient subgroups that exhibit lower class imbalance (i.e. higher baseline risk). A non-exhaustive sample of such algorithms include the contributions made by *Miro* (ROC-AUC 0.836) [34], *Makar* (ROC-AUC 0.828) [35] and *Elfiky* (ROC-AUC 0.940) [36]. Additionally, as the algorithms 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 algorithms 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 subsample from the training distribution [33-37].

Many clinicians recognize 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 [2,38-39]. We hope that our algorithms can aid physicians who face such challenges in systematically identifying patients at EOL to schedule for more timely planning.

While screening healthy populations traditionally demands tests with high specificity, its absolute level depends on the scheduled intervention. If the intervention scheduled for patients deemed high-risk by our algorithms is a 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 FPR (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 co-morbidities that constitute most of the high-risk patients. An evaluation in primary care could also benefit false positives that result from patients at high risk of death due to an acute condition, that have been discharged from the ED

erroneously. While the latter patient group is not the main focus of this work, the algorithms can be retrained on a refined population of younger patients with fewer comorbidities to learn identify such erroneous discharges. Using follow-up in primary care as the intervention would also address the proposed importance of involving primary care in advance care planning [39]. It is already not uncommon to arrange follow-up in primary care after an ED visit, which makes us believe that scheduling predicted positives for such follow-up after discharge from the ED 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 algorithms fit well with the ED workflow. Whilst 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 algorithms makes such use less straightforward. However, current methods for deploying predictive algorithms in hospital information systems would allow algorithms 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 e.g. 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 [2,37] and we do not expect that implementing our algorithms in clinical practice will prevent accelerating costs of care. Rather, we hope that the algorithms can promote value in healthcare by bringing patients, physicians and families closer to timely EOL discussions. Additionally, the scarcity of evidence supporting EOL interventions [40] poses a need for prospective trials, and the

algorithms 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 algorithms is that they may overfit to the characteristics of the development set and therefore not perform similarly across sites [41]. To mitigate this situation, we implemented cross-validation and assessed algorithm performance on data from a separate hospital, that the algorithms were previously unexposed to. Also, the use of standard data-elements makes our algorithms less susceptible to being overfit to the practices of a specific institution, as compared to algorithms that make predictions from a wider array of data elements that tend to be more institution specific (e.g. text in EHR notes that may reflect individual physicians' documentation style etc.). As variations in local processes or populations are expected to occur over time, our algorithms should be continuously monitored and periodically retrained to maintain performance if implemented in clinical practice. The inverse-probability weighting scheme maintained in this exercise makes it unlikely that algorithm performance is significantly impacted by re-training on datasets displaying different levels of class-imbalance.

Before deployment, we also suggest that the algorithms 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 customization of the algorithms is achievable by optimizing the decision threshold to produce the most favourable tradeoff 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.

Conclusions

In this paper we report performance of supervised learning algorithms, that predict 30-day mortality in patients discharged from the Emergency Department with excellent discrimination. The algorithms 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.

References

1 – Moses H 3rd, Matheson DHM, Dorsey ER, George BP, Sadoff D, Yoshimura S. The anatomy of health care in the United States. JAMA 2013;310(18):1947-1963.

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;142:932-937.

4 – Cutler DM, Rosen AB, Vijan S. The Value of Medical Spending in the United States, 1960-2000. N Engl J Med 2006;355(9):920-927

5 – Porter ME. What Is Value in Health Care? N Engl J Med 2010;363(26):2477-2481.

6 – Schroeder SA, Frist W, National Commission on Physician Payment Reform. Phasing Out Fee-for-Service Payment. N Engl J Med 2013;368(21):2029-2032.

7 – Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder ÉL. The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care. Ann Intern Med, 2003;138(4):288-299.

8 – Yasaitis L, Fisher ES, Skinner JS, Chandra A. Hospital quality and intensity of spending: is there an association?. Health Aff (Millwood) 2009;28(4):w566-w572.

9 – Mittler JN, Landon BE, Fisher ES, Cleary PD., Zaslavsky AM. Market variations in intensity of Medicare service use and beneficiary experiences with care. Health services research 2010;45(3): 647-669.

10 – Wennberg JE, Bronner K, Skinner JS, Fisher ES, Goodman DC. 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 (Millwood) 2009;28(1):103-112.

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(14):1665-1673.

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(5):480-488.

13 – Groff AC, Colla CH, Lee TH. Days spent at home—a patient-centered goal and outcome. N Engl J Med 2016;375(17):1610-1612.

14 – Silveira MJ, Kim SY, Langa KM. Advance directives and outcomes of surrogate decision making before death. N Engl J Med 2010;362(13):1211-1218.

15 – Teno JM, Fisher ES, Hamel MB, Coppola K, Dawson NV. Medical care inconsistent with patients' treatment goals: Association with 1-year Medicare resource use and survival. JAGS 2002;50(3):496-500.

16 – Pritchard RS, Fisher ES, Teno JM, et al, for the SUPPORT Investigators. Influence of patient preferences and local health system characteristics on the place of death. JAGS 1998;46(10):1242-1250.

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 – Linear Model Selection and Regularization. In: James G, Witten D, Hastie T, Tibshirani R. Editors. An introduction to Statistical Learning with applications in R. 1ed. New York. NY: Springer. 2013. 203-264.

19 – Support Vector Machines and Flexible Discriminants. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 417-458.

20 – Prototype Methods and Nearest-Neighbors. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 459-484.

21 – Boosting and Additive Trees. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 337-389.

22 – Breiman L. Random Forests. Machine learning 2001;45(1):5-32.

23 – Neural Networks. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 389-416.

24 – Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. Journal of Big Data. 2018 Dec 1;5(1):42.

25 – Ling C.X., Sheng V.S. (2011) Class Imbalance Problem. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA

26 – Statistical guidance on reporting results from studies evaluating diagnostic tests. Rockville, MD: Food and Drug Administration, Center for Devices and Radiological Health. 2007. (Docket No. 2003D-0044)

27 – Efron B. Better bootstrap confidence intervals. Journal of the American statistical Association 1987;82(397):171-185.

28 – Bossuyt PM, Reitsma JB, Bruns DE, et al, for the STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.

29 – Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research 2011;12: 2825-2830.

30 – Chollet F. Keras: Deep Learning for humans. GitHub, 2015. (https://github.com/fchollet/keras)

31 - Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement.

32 – Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss SA. combined comorbidity score predicted mortality in elderly patients better than existing scores. J Clin Epidemiol 2011;64(7), 749-759.

33 – Avati A, Jung K, Harman S, Downing L, Ng A, Shah N. Improving Palliative Care with Deep Learning. International Conference on Bioinformatics and Biomedicine (BIBM), 2017. pp. 311-316.

34 – Miró Ò, 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(10):698-705.

35 – Makar M, Ghassemi M, Cutler DM, Obermeyer Z. Short-term mortality prediction for elderly patients using Medicare claims data. Int J Mach Learn Comput 2015;*5*(3):192-197.

36 – Elfiky A, Pany M, Parikh R, Obermeyer Z. 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(3):e180926.

37 – Einav L, Finkelstein A, Mullainathan S, Obermeyer Z. Predictive modeling of US health care spending in late life. Science 2018;360:1462-1465.

38 – 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(20):1591-1598.

39 – Lynn J, DeVries KO, Arkes HR, et al. Ineffectiveness of the SUPPORT Intervention: Review of Explanations. JAGS 2000;48(5):S206-S213.

40 – Halpern SD. Toward evidence-based end-of-life care. N Engl J Med 2015;373(21):2001-2003.

41 – Obermeyer Z, Lee TH. Lost in thought—the limits of the human mind and the future of medicine. N Engl J Med 2017;377(13):1209-1211.

Acknowledgements

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

Funding

This work was partly funded by Region Halland, Sweden. The authors also wish to recognize 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 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.

Conflicts of interests

We have read and understood BMJ policy on declaration of interests and declare that we have no competing interests.

Author contributions

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

Patient involvement

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.

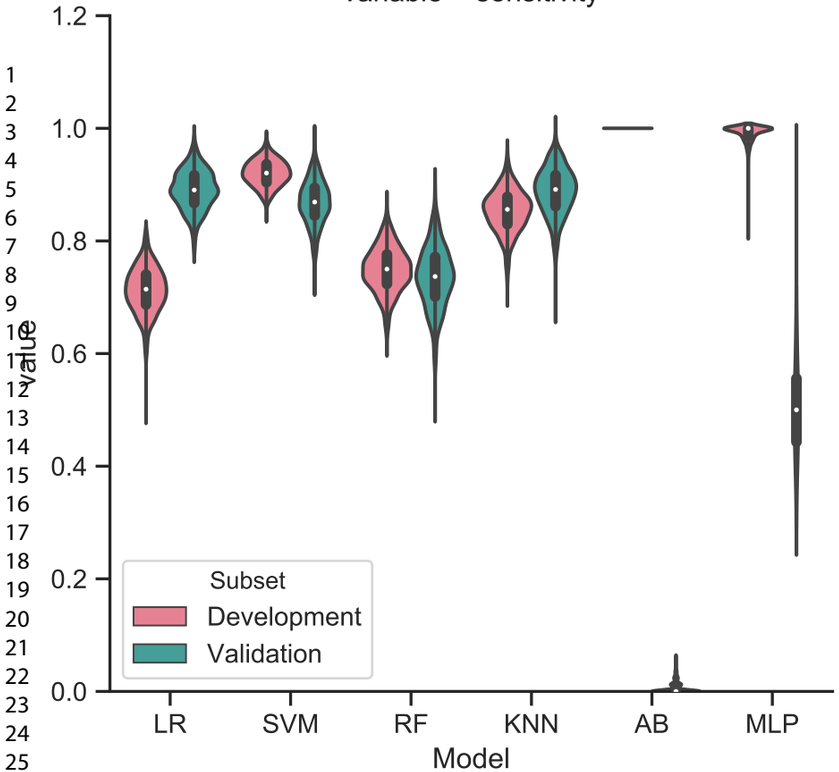
Data statement

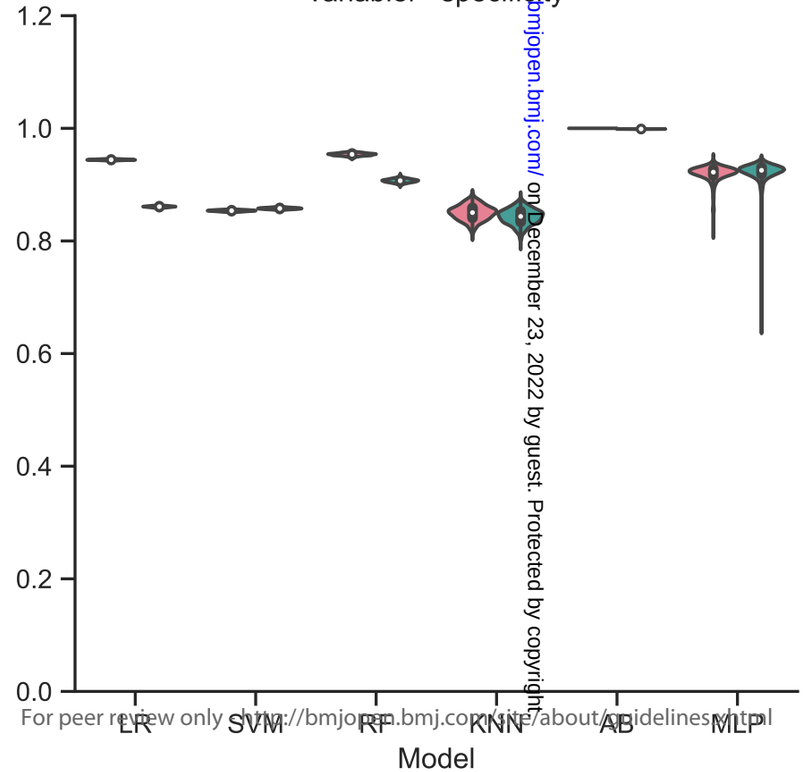
Technical appendix, statistical code and final models available upon request. Individual level patient data may not and therefore will not be shared.

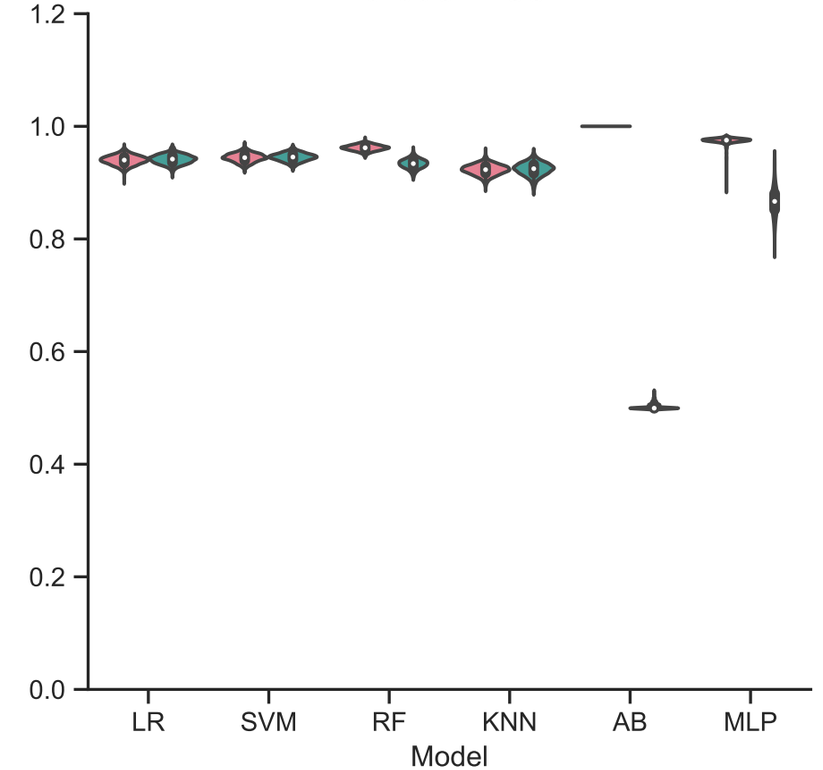
Figure captions

Figure 1: Algorithm performance (development and validation set)

Figure 2: Variable importance using the RF algorithm







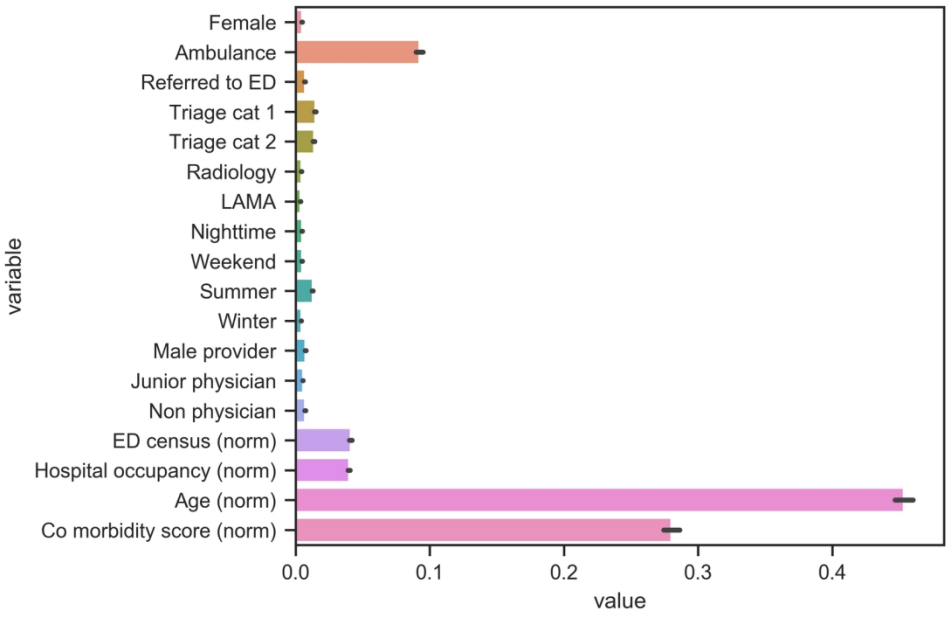
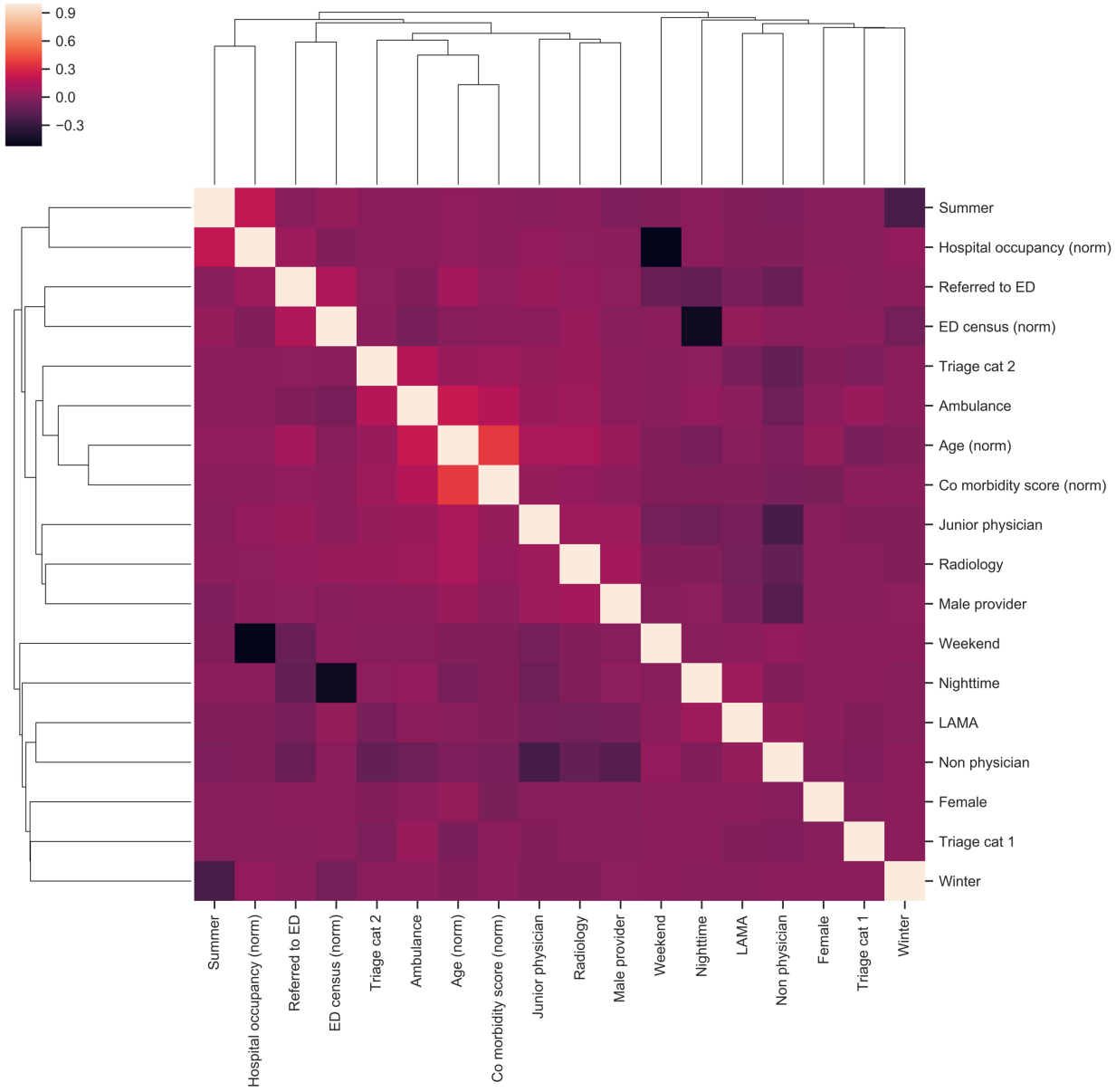


Figure 2: Variable importance using the RF algorithm

152x101mm (300 x 300 DPI)

Supplementary figures

Figure S1: Correlation matrix of predictors



Correlation coefficients (range -1, 1) for independent variables.

Supplementary Appendix

Construction of independent variables

Individual level Electronic Health Record (EHR) data from all ED visits in Region Halland during the period Jan 01 2015 to Dec 31 2016 were linked to records on inpatient visits, ambulance referrals and radiology orders. All tables were accessed through a recently constructed healthcare analytics platform, in Microsoft SQL Server 2014. Inpatient visits were linked to ED visits by unique personal identifiers derived from a subject's national Personal Identification Number (PIN) and a time criterion (inpatient registration +-3h of ED discharge), as were ambulance referrals (ambulance arrival +- 15min of ED arrival). Hospital bed occupancy was linked by date and facility (variable measured at 06.00am). ED census was linked by date, hour and facility. Remaining tables were linked on unique personal identifiers. The final selection of independent variables comprised patient age, gender, the Quan-Deyo modification of the Charlson Comorbidity Index [1], being referred to the ED by a physician, being transported to the ED in ambulance, perceived urgent medical condition (ED triage system 'RETTS' level 1-2 upon ED arrival [2]), radiology order occurring during the ED visit, leaving the ED against medical advice (LAMA), being discharged during on-call hours (10pm – 7am), during a holiday (including weekends), winter (Dec-Feb, roughly coherent with the influenza season), or summer (week 26-32, corresponding to Swedish vacation period). The co-morbidity score was calculated by linking all individual unique patient identifiers in the study population to all diagnosis data (ICD-10) registered in the healthcare analytics platform. The start of the diagnosis assessment period was set at 365.25 days before the first possible visit (i.e. before 00:00 Jan 1, 2015) and assessment continued throughout the study period. Hence, each individual visit was linked to any diagnoses for the patient registered throughout the region, from the start of the assessment period up until the individual visit discharge timestamp. Diagnoses were mapped to the relevant co-

morbidities in the R package 'icd' [3] (version 3.4.0). The LAMA variable was defined using mandatory input fields that are filled by ED nurses at patient departure.

Construction of the study endpoint

The outcome was assessed by linking records to the Swedish population register. Registering a 'notification of death' (dödsbevis) is a legal obligation in Sweden and must be completed before burial can be authorized. The notification of death is filled in and submitted by the diagnosing physician. As deaths are registered with a resolution of date, any deaths occurring on the date of the ED visit were considered inpatient deaths and therefore excluded. Although the registry should capture deaths in Swedish citizens, some loss to follow-up could result from non-Swedish residents (particularly common during summer).

Algorithm hyperparameter tuning

LR [4]

class sklearn.linear_model.**LogisticRegression**(*penalty='l2'*, *dual=False*, *tol=0.0001*, *C=1.0*, *fit_intercept=True*, *intercept_scaling=1*, *class_weight=None*, *random_state=None*, *solver='warn'*, *max_iter=100*, *multi_class='warn'*, *verbose=0*, *warm_start=False*, *n_jobs=None*)

Optimized for C:[1e-6 – 0.25]

Optimal C: 0.015

Class_weight=Balanced

RF [5]

class sklearn.ensemble.**RandomForestClassifier**(*n_estimators='warn', criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None*)

Optimized for n_estimators: [40 – 200]

Optimized for max_depth: [5 – 25]

Optimal n_estimators: 120

Optimal max_depth: 5

Class_weight=balanced

AB [6]

class sklearn.ensemble.**AdaBoostClassifier**(*base_estimator=None, n_estimators=50, learning_rate=1.0, algorithm='SAMME.R', random_state=None*)

Optimized for base_estimator: [gini, entropy]

Optimized for learninig_rate: [0.1 – 2]

Optimized for n_estimators: [5 – 100]

Optimal base_estimators: gini

Optimal n_estimators: 65

Optimal learning_rate: 0.7

Class_weight=balanced

SVM [7]

class sklearn.svm.**SVC**(*C=1.0*, *kernel='rbf'*, *degree=3*, *gamma='auto_deprecated'*, *coef0=0.0*, *shrinking=True*, *probability=False*, *tol=0.001*, *cache_size=200*, *class_weight=None*, *verbose=False*, *max_iter=-1*, *decision_function_shape='ovr'*, *random_state=None*)

Optimized for C: [0.001 – 1]

Optimized for kernel: [rbf, poly]

Optimal C: 0.01

Optimal kernel: rbf

Class_weight=balanced

KNN [8]

Class sklearn.neighbors.**KNeighborsClassifier**(*n_neighbors=5*, *weights='uniform'*, *algorithm='auto'*, *leaf_size=30*, *p=2*, *metric='minkowski'*, *metric_params=None*, *n_jobs=None*, ***kwargs*)

Optimized for n_neighbors: [1 – 31]

Optimized for metric: [eucledian, minkowski]

Optimal neighbors: 11

Optimal metric: Euclidean

MLP [9]

Epochs = 200

Batch size = 500

Optimizer = rmsprop

Loss = binary cross entropy

Learning rate = 0.01

Activation functions = sigmoid

Optimized for Number of nodes in hidden layer: [5 – 15]

Optimal nodes: 9

References

1 – Quan H, Sundararajan V, Halfon P, et al. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data. Med Care Care 2005;43:1130-1139.

2 – Widgren B. RETTS: Akutsjukvård direkt. 1 ed. Lund, Sweden: Studentlitteratur, 2012.

3 – R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018: URL http://www.R-project.org/.

4 – sklearn.linear_model.LogisticRegression in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

5 – sklearn.ensemble.RandomForestClassifier in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

6 – sklearn.ensemble.AdaBoostClassifier in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html)

7 – sklearn.svm.SVC in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC)

8 – sklearn.neighbors.KNeighborsClassifier in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html)

9 – Keras: Deep Learning for humans. Chollet, F. 2015. Keras, GitHub. (https://github.com/fchollet/keras)

Reporting checklist for prediction model development and validation study.

Based on the TRIPOD guidelines.

Instructions to authors

Complete this checklist by entering the page numbers from your manuscript where readers will find each of the items listed below.

Your article may not currently address all the items on the checklist. Please modify your text to include the missing information. If you are certain that an item does not apply, please write "n/a" and provide a short explanation.

Upload your completed checklist as an extra file when you submit to a journal.

In your methods section, say that you used the TRIPOD reporting guidelines, and cite them as:

Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement.

		Reporting Item	Page Number
	#1	Identify the study as developing and / or validating a multivariable prediction model, the target population, and the outcome to be predicted.	2
	#2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
	#3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	4-5, 8
	#3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	2
Source of data	#4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5

	#4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5
Participants	#5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5, 8
	#5b	Describe eligibility criteria for participants.	5
	#5c	Give details of treatments received, if relevant	n/a No treatments administered
Outcome	#6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	5
	#6b	Report any actions to blind assessment of the outcome to be predicted.	n/a Assessed at aggregate-level only
Predictors	#7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured	6-7, SA
	#7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	n/a Assessed at aggregate-level only
Sample size	#8	Explain how the study size was arrived at.	5
Missing data	#9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	6
Statistical analysis methods	#10a	If you are developing a prediction model describe how predictors were handled in the analyses.	9
	#10b	If you are developing a prediction model, specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	6-9

	#10c	If you are validating a prediction model, describe how the predictions were calculated.	9
	#10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	8-9
	#10e	If you are validating a prediction model, describe any model updating (e.g., recalibration) arising from the validation, if done	8, SA
Risk groups	#11	Provide details on how risk groups were created, if done.	n/a No risk-groups were created
Development vs. validation	#12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	6-7 (Table 1)
Participants	#13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	See note 1
	#13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	See note 2
	#13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	See note 3
Model development	#14a	If developing a model, specify the number of participants and outcome events in each analysis.	See note 4
	#14b	If developing a model, report the unadjusted association, if calculated between each candidate predictor and outcome.	See note 5
Model specification	#15a	If developing a model, present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	n/a Provided upon request
	#15b	If developing a prediction model, explain how to the use it.	8-9, 14-15

Model performance	#16	Report performance measures (with CIs) for the prediction model.	See note 6
Model-updating	#17	If validating a model, report the results from any model updating, if done (i.e., model specification, model performance).	n/a Models not updated after validation
Limitations	#18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	14-15
Interpretation	#19a	For validation, discuss the results with reference to performance in the development data, and any other validation data	10-11
	#19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	12-14
Implications	#20	Discuss the potential clinical use of the model and implications for future research	12-15
Supplementary information	#21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	11,24
Funding	#22	Give the source of funding and the role of the funders for the present study.	23

Author notes

1. 6-7, 10 (ref table 1, 2)
2. 6-7 (ref table 1)
3. 6-7 (ref table 1)
4. 6-7 (ref table 1)
5. 6-7 (ref table 1)
6. 11-12 (ref table 3)

This checklist was completed on 18. November 2018 using http://www.goodreports.org/, a tool made by the EQUATOR Network in collaboration with Penelope.ai

BMJ Open

Training machine learning models to predict 30-day mortality in patients discharged from the Emergency Department

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-028015.R2
Article Type:	Research
Date Submitted by the Author:	13-Apr-2019
Complete List of Authors:	Blom, Mathias; Lund University Medical Faculty, Department of Clinical Sciences Lund, Medicine Ashfaq, Awais; Halmstad University, Center for Applied Intelligent Systems Research (CAISR); Halland Hospital, Region Halland Sant'Anna, Anita; Halmstad University, Center for Applied Intelligent Systems Research (CAISR) Anderson, Philip; Brigham & Women's Hospital, Department of Emergency Medicine; Harvard Medical School Lingman, Markus; Sahlgrenska Academy, University of Gothenburg, Department of molecular and clinical Medicine/Cardiology; Halland Hospital, Region Halland
Primary Subject Heading:	Emergency medicine
Secondary Subject Heading:	Epidemiology, Palliative care
Keywords:	ACCIDENT & EMERGENCY MEDICINE, BIOTECHNOLOGY & BIOINFORMATICS, EPIDEMIOLOGY, PALLIATIVE CARE, PUBLIC HEALTH

Training machine learning models to predict 30-day mortality in patients discharged from the Emergency Department

Mathias C. Blom M.D. Ph.D.[1], Awais Ashfaq M.Sc.[2], Anita Sant'Anna Ph.D.[2], Philip D. Anderson M.D.[3], Markus Lingman M.D. Ph.D[4].

Corresponding author:

Lingman, Markus, M.D. Ph.D.

Markus.Lingman@regionhalland.se

[1] – Department of Clinical Sciences Lund, Medicine, Lund University, Lund, Sweden.

[2] – Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, Sweden. Halland Hospital, Region Halland, Sweden

[3] – Department of Emergency Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, US.

[4] – Institute of Medicine, Dept. of Molecular and Clinical Medicine/Cardiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. Halland Hospital, Region Halland, Sweden

Keywords: Emergency Medicine, Mortality, Machine Learning, Advance Care Planning,

Word count: 2,626 (excluding abstract, figures, tables, legends and references)

Abstract

Objectives: The aim of this work was to train machine learning models to identify patients at end of life (EOL) 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 - 0.96), with sensitivity 0.87 (95% CI 0.80, 0.93) and specificity 0.86 (0.86, 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 in terms 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.

Article summary

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 U.S. and across the globe [1-3], evidence suggests that advances in healthcare technologies and increased utilization 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 randomized trials on the topic are lacking, observational studies of variation in U.S. 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 jeopardized at EOL. We are not aware of firm evidence linking overtreatment to the recently observed fall in U.S. 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 utilizing data from a comprehensive healthcare analysis platform in Region Halland, southern Sweden. A consecutive sample of ED visits in the region from Jan 01 2015 to Dec 31 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 upon 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 upon 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 supplementary appendix. The unit of analysis is one ED visit. Complete-case analysis was deployed as the proportion missing values was low.

	Complete dataset[1] n=123,975	**Validation set** n=55,164	**Development set** n=65,776			
Variable	N missing (%)	% exposed[2]	% exposed	% experiencing outcome in exposed	% experiencing outcome in unexposed	P[3]
Female	0 (0.0)	49.5	49.0	0.19	0.22	0.48

Arrived by ambulance	0 (0.0)[4]	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)[5]	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 nighttime	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	3,385 (2.73)	44.2	43.9	0.24	0.18	0.09
Junior physician	3,385 (2.73)	22.5	25.2	0.25	0.19	0.22
Non-physician provider	3,385 (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[6]
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
Co-morbidity score	3,035 (2.45)	0.0	0.0	2.0	0.0	<0.001

		(0.0, 0.0)	(0.0, 0.0)	(1.0, 6.0)	(0.0, 0.0)	
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

Table 1: Descriptive statistics

[1] N before excluding missing values
[2] proportion of subjects sharing characteristic indicated in 'variable' column
[3] P-value for difference in outcome, exposed vs unexposed, non-adjusted, development set. Arrived by ambulance, referred by physician, triage priority 1 & 2, discharged summer, non-physician provider with $P<0.05$.
[4] Database-linkage between source table and ambulance dispatches for 14,918 (12.0%) subjects
[5] Database-linkage between source table and radiology orders for 18,435 (14.9%) subjects.
[6] P-value for difference in predictor distribution, subjects experiencing outcome vs subjects not experiencing outcome, non-adjusted, development set. Age, Co-morbidity score and ED census with $P<0.05$.

Statistical analysis

Six different algorithms were selected for model training, based on their principally different approaches to prediction. These were L2 regularized logistic regression (LR) [20], support vector machine (SVM) [21], K-nearest neighbours classifier (KNN) [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 datasets pose risks of biasing the algorithm towards the majority class. To mitigate this, we over-sampled 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 penalized 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 optimize models for area under the ROC-curve (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 hyper-parameters was identified through systematic grid-search (using 5-fold 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 over- or underfit. The development set consisted of visits to one ED in the region and the validation set consisted of visits to another. 95% CI:s were obtained by identifying the 5th and 95th percentiles of a probability distribution of each relevant measure, obtained by re-fitting 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 RandomForests™ algorithm [24]. Continuous variables were normalized before being fed into the models. Observations were designated predicted positive if the predicted probability of the outcome was ≥50%. Performance was reported as sensitivity and specificity in accordance with STARD [30] and benchmarked across models by comparing 95% CI:s. Univariate comparisons were conducted using the Wilcoxon rank sum test for continuous variables and the chi2 test for indicator variables. Multicollinearity was addressed using Spearman's rho. Statistical analyses were undertaken in Python™ 3.6, scikit-learn 20.0 [31] and Keras [32]. Data analysis was conducted by one author (A.A.) with supervision from M.B. and A.SA. TRIPOD reporting guidelines were used [33].

Results

Descriptive statistics

The development set included 65,776 observations and the validation set 55,164 observations, after excluding 3,035 observations with missing information for co-morbidity score. 3,385 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, 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 Figure S1). All models converged and did not indicate multicollinearity.

	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”	+8,070	117,815
Including all ED visits with discharge destination “LAMA”	+6,644	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	-3,035	120,940
Final sample	N/A	120,940

Table 2: Exclusion analysis

Model performance

All models performed excellently on the development set, ranging from ROC-AUC 0.92 (95% CI 0.91, 0.94) for KNN to 1.00 (1.00, 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, i.e. an inability to correctly identify cases, which is in line with expectations for imbalanced tasks (i.e. 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, 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 supplementary appendix. Final models, source code and instructions are made available upon request.

Patient age and co-morbidity 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 co-morbid patients are at increased risk of death as well as that arriving by ambulance may indicate a more serious condition. A post hoc sensitivity analysis that was undertaken on the final RF algorithm by retraining it on the top 5

features only (age, co-morbidity 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 0.922-0.949).

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

Table 3: Algorithm performance (development and validation set)

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 datasets 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 heterogenous patient subgroups that exhibit lower class imbalance (i.e. higher baseline risk). A non-exhaustive sample of such models include the contributions made by *Miro* (ROC-AUC 0.836) [36], *Makar* (ROC-AUC 0.828) [37] and *Elfiky* (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 subsample from the training distribution [35-39].

Many clinicians recognize 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 (FPR) (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 co-morbidities that constitute most of the high-risk patients. An evaluation in primary care could also benefit patients that 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. Whilst 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 e.g. 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 to models that make predictions from a wider array of data elements that tend to be more institution specific (e.g. 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 re-training on datasets 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 customization of the models is achievable by optimizing 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 Emergency Department 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.

References

1 – Moses H 3rd, Matheson DHM, Dorsey ER, George BP, Sadoff D, Yoshimura S. The anatomy of health care in the United States. JAMA 2013;310(18):1947-1963.

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;142:932-937.

4 – Cutler DM, Rosen AB, Vijan S. The Value of Medical Spending in the United States, 1960-2000. N Engl J Med 2006;355(9):920-927

5 – Porter ME. What Is Value in Health Care? N Engl J Med 2010;363(26):2477-2481.

6 – Schroeder SA, Frist W, National Commission on Physician Payment Reform. Phasing Out Fee-for-Service Payment. N Engl J Med 2013;368(21):2029-2032.

7 – Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder ÉL. The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care. Ann Intern Med, 2003;138(4):288-299.

8 – Yasaitis L, Fisher ES, Skinner JS, Chandra A. Hospital quality and intensity of spending: is there an association?. Health Aff (Millwood) 2009;28(4):w566-w572.

9 – Mittler JN, Landon BE, Fisher ES, Cleary PD., Zaslavsky AM. Market variations in intensity of Medicare service use and beneficiary experiences with care. Health services research 2010;45(3): 647-669.

10 – Wennberg JE, Bronner K, Skinner JS, Fisher ES, Goodman DC. 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 (Millwood) 2009;28(1):103-112.

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(14):1665-1673.

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(5):480-488.

13 – Groff AC, Colla CH, Lee TH. Days spent at home—a patient-centered goal and outcome. N Engl J Med 2016;375(17):1610-1612.

14 – Silveira MJ, Kim SY, Langa KM. Advance directives and outcomes of surrogate decision making before death. N Engl J Med 2010;362(13):1211-1218.

15 – Teno JM, Fisher ES, Hamel MB, Coppola K, Dawson NV. Medical care inconsistent with patients' treatment goals: Association with 1-year Medicare resource use and survival. JAGS 2002;50(3):496-500.

16 – Pritchard RS, Fisher ES, Teno JM, et al, for the SUPPORT Investigators. Influence of patient preferences and local health system characteristics on the place of death. JAGS 1998;46(10):1242-1250.

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(7):791–797.

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. *J Clin Oncol*. 2010;28(7):1203–1208.

20 – Linear Model Selection and Regularization. In: James G, Witten D, Hastie T, Tibshirani R. Editors. An introduction to Statistical Learning with applications in R. 1ed. New York. NY: Springer. 2013. 203-264.

21 – Support Vector Machines and Flexible Discriminants. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 417-458.

22 – Prototype Methods and Nearest-Neighbors. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 459-484.

23 – Boosting and Additive Trees. In: Hastie T, Tibshirani R, Friedman J. Editors. The elements of statistical learning 2ed. New York. NY: Springer. 2009. 337-389.

24 – Breiman L. Random Forests. Machine learning 2001;45(1):5-32.

25 – Neural Networks. 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, Seliya N. A survey on addressing high-class imbalance in big data. Journal of Big Data. 2018 Dec 1;5(1):42.

27 – Ling C.X., Sheng V.S. (2011) Class Imbalance Problem. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA

28 – Statistical guidance on reporting results from studies evaluating diagnostic tests. Rockville, MD: Food and Drug Administration, Center for Devices and Radiological Health. 2007. (Docket No. 2003D-0044)

29 – Efron B. Better bootstrap confidence intervals. Journal of the American statistical Association 1987;82(397):171-185.

30 – Bossuyt PM, Reitsma JB, Bruns DE, et al, for the STARD Group. 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: 2825-2830.

32 – Chollet F. Keras: Deep Learning for humans. GitHub, 2015. (https://github.com/fchollet/keras)

33 - Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement.

34 – Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss SA. combined comorbidity score predicted mortality in elderly patients better than existing scores. J Clin Epidemiol 2011;64(7), 749-759.

35 – Avati A, Jung K, Harman S, Downing L, Ng A, Shah N. Improving Palliative Care with Deep Learning. International Conference on Bioinformatics and Biomedicine (BIBM), 2017. pp. 311-316.

36 – Miró Ò, 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(10):698-705.

37 – Makar M, Ghassemi M, Cutler DM, Obermeyer Z. Short-term mortality prediction for elderly patients using Medicare claims data. Int J Mach Learn Comput 2015;*5*(3):192-197.

38 – Elfiky A, Pany M, Parikh R, Obermeyer Z. 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(3):e180926.

39 – Einav L, Finkelstein A, Mullainathan S, Obermeyer Z. Predictive modeling of US health care spending in late life. Science 2018;360:1462-1465.

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(20):1591-1598.

41 – Lynn J, DeVries KO, Arkes HR, et al. Ineffectiveness of the SUPPORT Intervention: Review of Explanations. JAGS 2000;48(5):S206-S213.

42 – Halpern SD. Toward evidence-based end-of-life care. N Engl J Med 2015;373(21):2001-2003.

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(13):1209-1211.

Acknowledgements

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

Funding

This work was partly funded by Region Halland, Sweden. The authors also wish to recognize 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 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.

Conflicts of interests

We have read and understood BMJ policy on declaration of interests and declare that we have no competing interests.

Author contributions

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

Patient involvement

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.

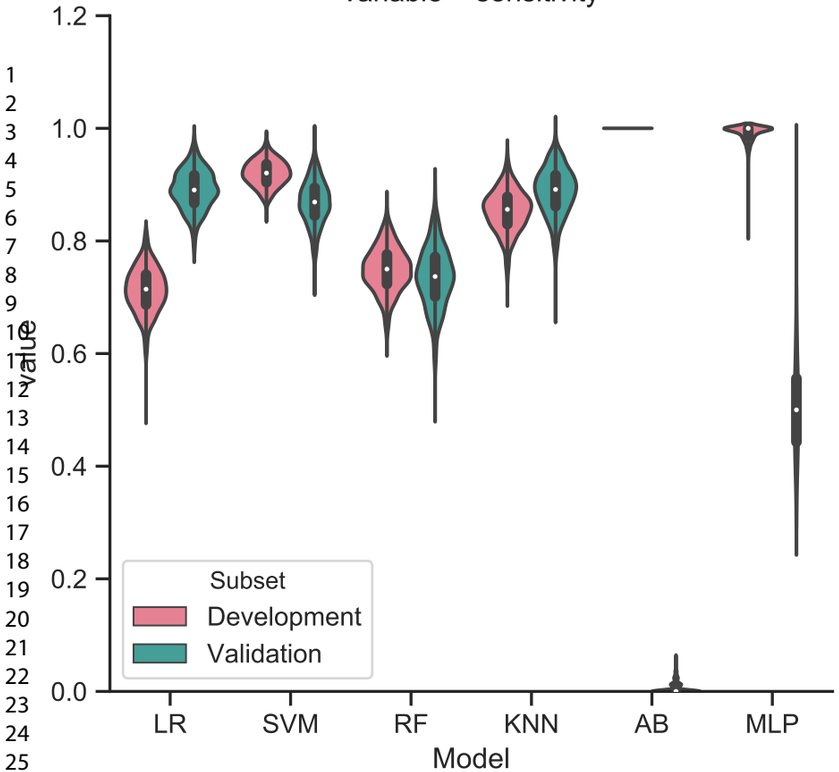
Data statement

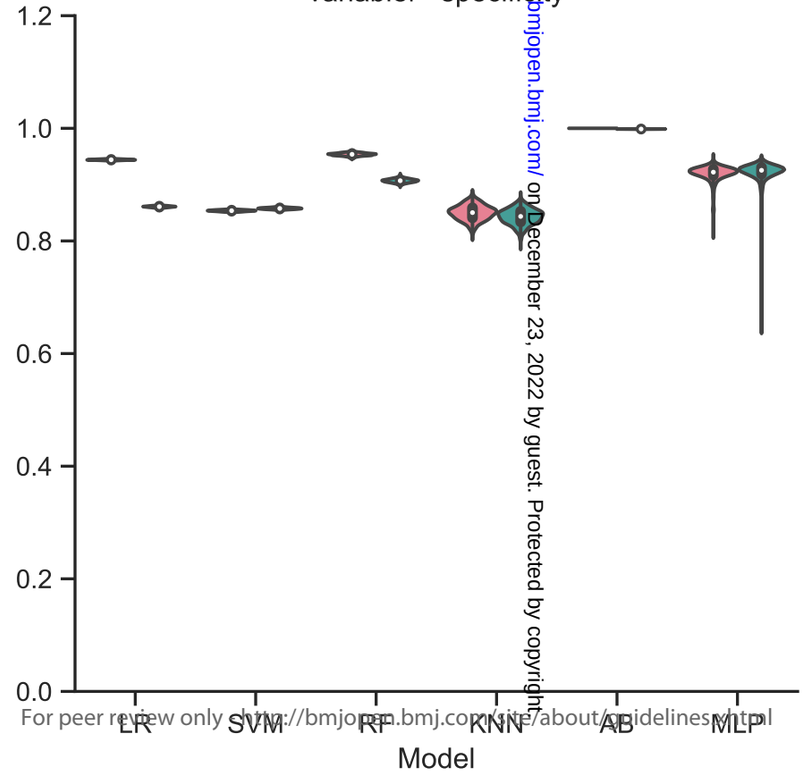
Technical appendix, statistical code and final models available upon request. Individual level patient data may not and therefore will not be shared.

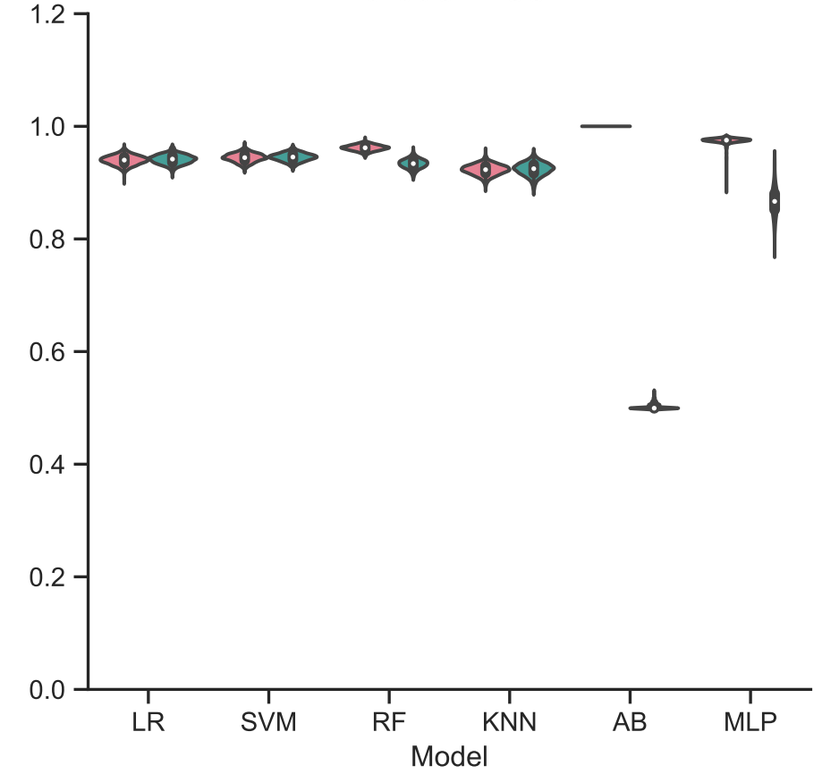
Figure captions

Figure 1: Algorithm performance (development and validation set)

Figure 2: Variable importance using the RF algorithm







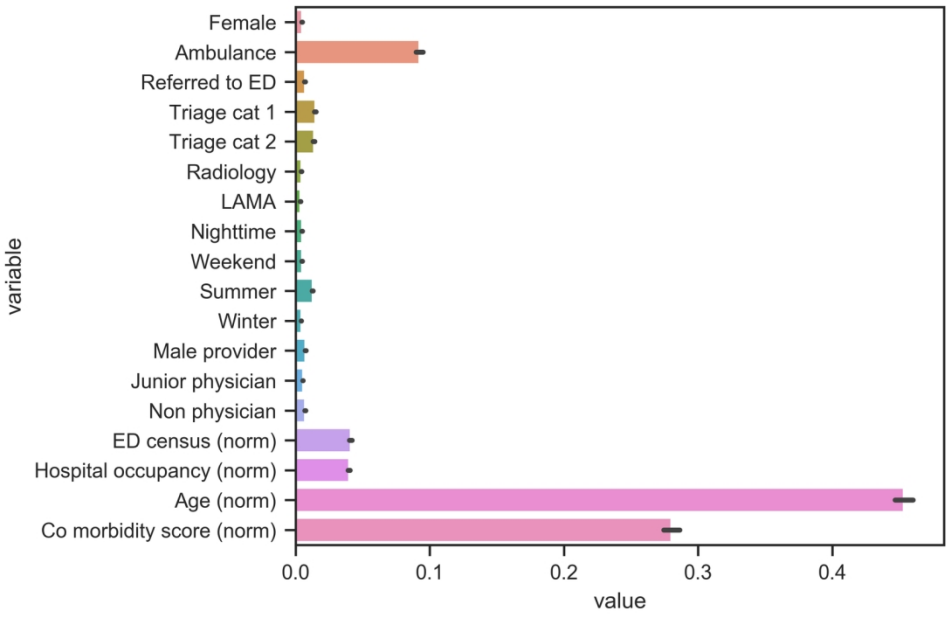
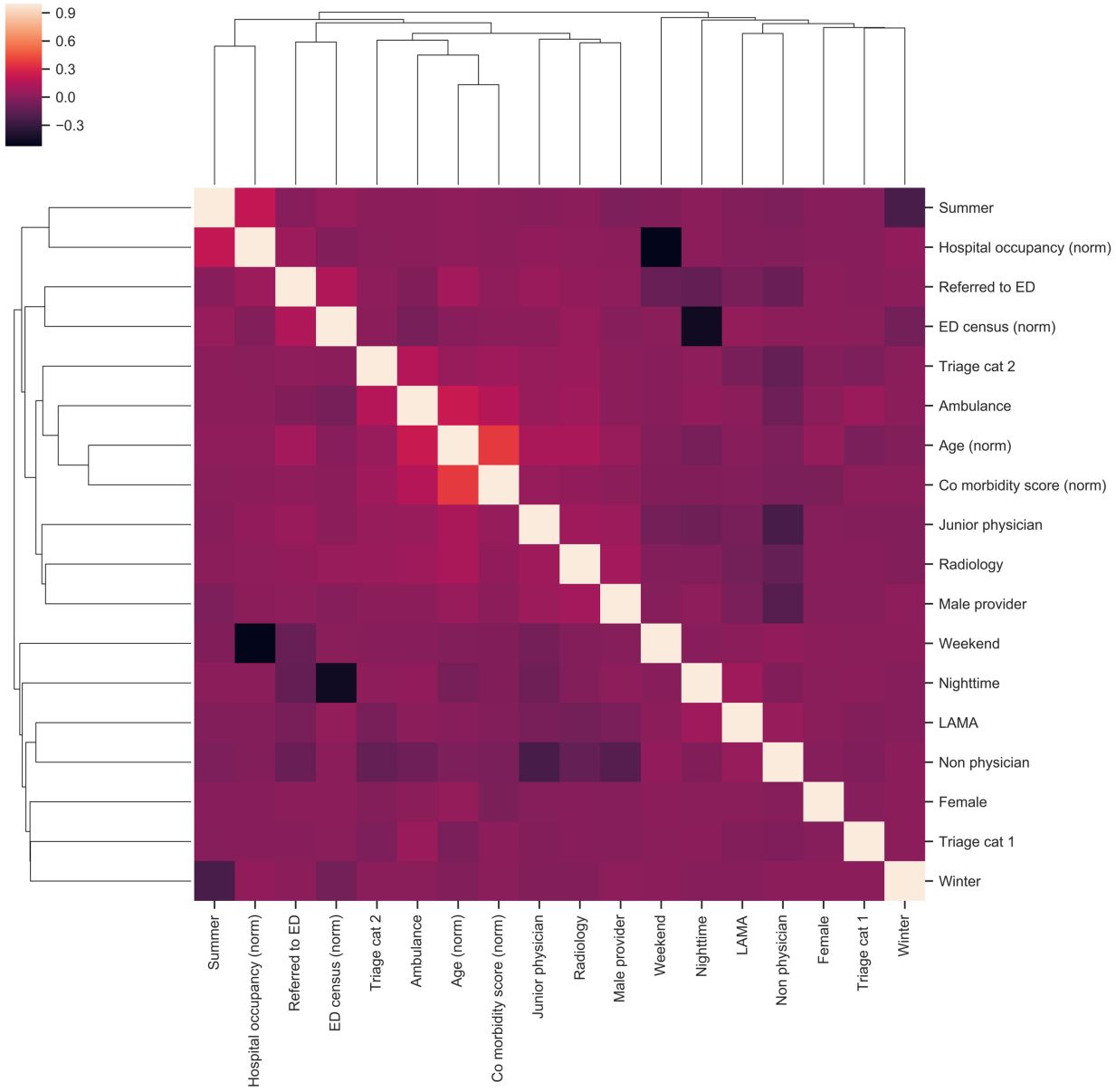


Figure 2: Variable importance using the RF algorithm

152x101mm (300 x 300 DPI)

Supplementary figures

Figure S1: Correlation matrix of predictors



Correlation coefficients (range -1, 1) for independent variables.

Supplementary Appendix

Construction of independent variables

Individual level Electronic Health Record (EHR) data from all ED visits in Region Halland during the period Jan 01 2015 to Dec 31 2016 were linked to records on inpatient visits, ambulance referrals and radiology orders. All tables were accessed through a recently constructed healthcare analytics platform, in Microsoft SQL Server 2014. Inpatient visits were linked to ED visits by unique personal identifiers derived from a subject's national Personal Identification Number (PIN) and a time criterion (inpatient registration +-3h of ED discharge), as were ambulance referrals (ambulance arrival +- 15min of ED arrival). Hospital bed occupancy was linked by date and facility (variable measured at 06.00am). ED census was linked by date, hour and facility. Remaining tables were linked on unique personal identifiers. The final selection of independent variables comprised patient age, gender, the Quan-Deyo modification of the Charlson Comorbidity Index [1], being referred to the ED by a physician, being transported to the ED in ambulance, perceived urgent medical condition (ED triage system 'RETTS' level 1-2 upon ED arrival [2]), radiology order occurring during the ED visit, leaving the ED against medical advice (LAMA), being discharged during on-call hours (10pm – 7am), during a holiday (including weekends), winter (Dec-Feb, roughly coherent with the influenza season), or summer (week 26-32, corresponding to Swedish vacation period). The co-morbidity score was calculated by linking all individual unique patient identifiers in the study population to all diagnosis data (ICD-10) registered in the healthcare analytics platform. The start of the diagnosis assessment period was set at 365.25 days before the first possible visit (i.e. before 00:00 Jan 1, 2015) and assessment continued throughout the study period. Hence, each individual visit was linked to any diagnoses for the patient registered throughout the region, from the start of the assessment period up until the individual visit discharge timestamp. Diagnoses were mapped to the relevant co-

morbidities in the R package 'icd' [3] (version 3.4.0). The LAMA variable was defined using mandatory input fields that are filled by ED nurses at patient departure.

Construction of the study endpoint

The outcome was assessed by linking records to the Swedish population register. Registering a 'notification of death' (dödsbevis) is a legal obligation in Sweden and must be completed before burial can be authorized. The notification of death is filled in and submitted by the diagnosing physician. As deaths are registered with a resolution of date, any deaths occurring on the date of the ED visit were considered inpatient deaths and therefore excluded. Although the registry should capture deaths in Swedish citizens, some loss to follow-up could result from non-Swedish residents (particularly common during summer).

Algorithm hyperparameter tuning

LR [4]

class sklearn.linear_model.**LogisticRegression**(*penalty='l2'*, *dual=False*, *tol=0.0001*, *C=1.0*, *fit_intercept=True*, *intercept_scaling=1*, *class_weight=None*, *random_state=None*, *solver='warn'*, *max_iter=100*, *multi_class='warn'*, *verbose=0*, *warm_start=False*, *n_jobs=None*)

Optimized for C:[1e-6 – 0.25]

Optimal C: 0.015

Class_weight=Balanced

RF [5]

class sklearn.ensemble.**RandomForestClassifier**(*n_estimators='warn', criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None*)

Optimized for n_estimators: [40 – 200]

Optimized for max_depth: [5 – 25]

Optimal n_estimators: 120

Optimal max_depth: 5

Class_weight=balanced

AB [6]

class sklearn.ensemble.**AdaBoostClassifier**(*base_estimator=None, n_estimators=50, learning_rate=1.0, algorithm='SAMME.R', random_state=None*)

Optimized for base_estimator: [gini, entropy]

Optimized for learninig_rate: [0.1 – 2]

Optimized for n_estimators: [5 – 100]

Optimal base_estimators: gini

Optimal n_estimators: 65

Optimal learning_rate: 0.7

Class_weight=balanced

SVM [7]

class sklearn.svm.**SVC**(*C=1.0*, *kernel='rbf'*, *degree=3*, *gamma='auto_deprecated'*, *coef0=0.0*, *shrinking=True*, *probability=False*, *tol=0.001*, *cache_size=200*, *class_weight=None*, *verbose=False*, *max_iter=-1*, *decision_function_shape='ovr'*, *random_state=None*)

Optimized for C: [0.001 – 1]

Optimized for kernel: [rbf, poly]

Optimal C: 0.01

Optimal kernel: rbf

Class_weight=balanced

KNN [8]

Class sklearn.neighbors.**KNeighborsClassifier**(*n_neighbors=5*, *weights='uniform'*, *algorithm='auto'*, *leaf_size=30*, *p=2*, *metric='minkowski'*, *metric_params=None*, *n_jobs=None*, ***kwargs*)

Optimized for n_neighbors: [1 – 31]

Optimized for metric: [eucledian, minkowski]

Optimal neighbors: 11

Optimal metric: Euclidean

MLP [9]

Epochs = 200

Batch size = 500

Optimizer = rmsprop

Loss = binary cross entropy

Learning rate = 0.01

Activation functions = sigmoid

Optimized for Number of nodes in hidden layer: [5 – 15]

Optimal nodes: 9

References

1 – Quan H, Sundararajan V, Halfon P, et al. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data. Med Care Care 2005;43:1130-1139.

2 – Widgren B. RETTS: Akutsjukvård direkt. 1 ed. Lund, Sweden: Studentlitteratur, 2012.

3 – R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018: URL http://www.R-project.org/.

4 – sklearn.linear_model.LogisticRegression in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

5 – sklearn.ensemble.RandomForestClassifier in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

6 – sklearn.ensemble.AdaBoostClassifier in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html)

7 – sklearn.svm.SVC in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC)

8 – sklearn.neighbors.KNeighborsClassifier in Scikit-learn: Machine learning in Python. [Cited 2018 November 5]. (http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html)

9 – Keras: Deep Learning for humans. Chollet, F. 2015. Keras, GitHub. (https://github.com/fchollet/keras)

Reporting checklist for prediction model development and validation study.

Based on the TRIPOD guidelines.

Instructions to authors

Complete this checklist by entering the page numbers from your manuscript where readers will find each of the items listed below.

Your article may not currently address all the items on the checklist. Please modify your text to include the missing information. If you are certain that an item does not apply, please write "n/a" and provide a short explanation.

Upload your completed checklist as an extra file when you submit to a journal.

In your methods section, say that you used the TRIPOD reporting guidelines, and cite them as:

Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement.

		Reporting Item	Page Number
	#1	Identify the study as developing and / or validating a multivariable prediction model, the target population, and the outcome to be predicted.	2
	#2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
	#3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	4-5, 8
	#3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	2
Source of data	#4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5

	#4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5
Participants	#5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5, 8
	#5b	Describe eligibility criteria for participants.	5
	#5c	Give details of treatments received, if relevant	n/a No treatments administered
Outcome	#6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	5
	#6b	Report any actions to blind assessment of the outcome to be predicted.	n/a Assessed at aggregate-level only
Predictors	#7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured	6-7, SA
	#7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	n/a Assessed at aggregate-level only
Sample size	#8	Explain how the study size was arrived at.	5
Missing data	#9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	6
Statistical analysis methods	#10a	If you are developing a prediction model describe how predictors were handled in the analyses.	9
	#10b	If you are developing a prediction model, specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	6-9

	#10c	If you are validating a prediction model, describe how the predictions were calculated.	9
	#10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	8-9
	#10e	If you are validating a prediction model, describe any model updating (e.g., recalibration) arising from the validation, if done	8, SA
Risk groups	#11	Provide details on how risk groups were created, if done.	n/a No risk-groups were created
Development vs. validation	#12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	6-7 (Table 1)
Participants	#13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	See note 1
	#13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	See note 2
	#13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	See note 3
Model development	#14a	If developing a model, specify the number of participants and outcome events in each analysis.	See note 4
	#14b	If developing a model, report the unadjusted association, if calculated between each candidate predictor and outcome.	See note 5
Model specification	#15a	If developing a model, present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	n/a Provided upon request
	#15b	If developing a prediction model, explain how to the use it.	8-9, 14-15

Model performance	#16	Report performance measures (with CIs) for the prediction model.	See note 6
Model-updating	#17	If validating a model, report the results from any model updating, if done (i.e., model specification, model performance).	n/a Models not updated after validation
Limitations	#18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	14-15
Interpretation	#19a	For validation, discuss the results with reference to performance in the development data, and any other validation data	10-11
	#19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	12-14
Implications	#20	Discuss the potential clinical use of the model and implications for future research	12-15
Supplementary information	#21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	11,24
Funding	#22	Give the source of funding and the role of the funders for the present study.	23

Author notes

1. 6-7, 10 (ref table 1, 2)
2. 6-7 (ref table 1)
3. 6-7 (ref table 1)
4. 6-7 (ref table 1)
5. 6-7 (ref table 1)
6. 11-12 (ref table 3)

This checklist was completed on 18. November 2018 using http://www.goodreports.org/, a tool made by the EQUATOR Network in collaboration with Penelope.ai