Published in final edited form as:

Am J Crit Care. 2018 September ; 27(5): 381–391. doi:10.4037/ajcc2018957.

ADVANCING IN-HOSPITAL CLINICAL DETERIORATION PREDICTION MODELS

Alvin D. Jeffery, PhD, RN [medical informatics fellow, postdoctoral research fellow],
US Department of Veterans Affairs, Tennessee Valley Health-care System, Nashville, Tennessee,
Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee.

Mary S. Dietrich, PhD, MS [professor of statistics and measurement],
Schools of Medicine (Biostatistics, Vanderbilt-Ingram Cancer Center, Psychiatry) and Nursing,
Vanderbilt University.

Daniel Fabbri, PhD [assistant professor],
Department of Biomedical Informatics, Vanderbilt University.

Betsy Kennedy, PhD, RN [professor],
School of Nursing, Vanderbilt University.

Laurie L. Novak, PhD [assistant professor],
Department of Biomedical Informatics, Vanderbilt University.

Joseph Coco, MS [senior application developer], and
Department of Biomedical Informatics, Vanderbilt University.

Lorraine C. Mion, PhD, RN [professor]
College of Nursing, The Ohio State University, Columbus, Ohio.

Abstract

Background—Early warning systems lack robust evidence that they improve patients' outcomes, possibly because of their limitation of predicting binary rather than time-to-event outcomes.

Objectives—To compare the prediction accuracy of 2 statistical modeling strategies (logistic regression and Cox proportional hazards regression) and 2 machine learning strategies (random forest and random survival forest) for in-hospital cardiopulmonary arrest.

Methods—Retrospective cohort study with prediction model development from deidentified electronic health records at an urban academic medical center.

Results—The classification models (logistic regression and random forest) had statistical recall and precision similar to or greater than those of the time-to-event models (Cox proportional hazards regression and random survival forest). However, the time-to-event models provided

To purchase electronic or print reprints, contact American Association of Critical-Care Nurses, 101 Columbia, Aliso Viejo, CA 92656. Phone, (800) 899-1712 or (949) 362-2050 (ext 532); fax, (949) 362-2049; reprints@aacn.org.

Corresponding author: Alvin D. Jeffery, 2525 West End Ave, #1475, Nashville, TN 37203 (alvinjeffery@gmail.com).

predictions that could potentially better indicate to clinicians whether and when a patient is likely to experience cardiopulmonary arrest.

Conclusions—As early warning scoring systems are refined, they must use the best analytical methods that both model the underlying phenomenon and provide an understandable prediction. (*American Journal of Critical Care*. 2018; 27:381–391)

widespread implementation of rapid response teams and early warning scoring systems throughout hospitals has resulted in debatable improvements in clinical deterioration outcomes.^{1,2} Even if early warning systems and rapid response teams improve patients' outcomes, the incidence of in-hospital clinical deterioration remains high and is associated with low survival rates.^{3–5} Given that the prevention of adverse outcomes will depend on early recognition followed by appropriate management, tools to aid these processes are needed. Clinical prediction models, especially those incorporated into decision support tools that automatically retrieve data from electronic health records, are becoming increasingly popular and might be able to assist in the early identification of clinical deterioration.^{2,6–15}

Optimal statistical approaches to embed within decision support tools and assist clinicians with recognition are still being identified. Most statistical approaches are simply classification models that attempt to identify the likelihood of an event. Researchers have focused on increasingly accurate models, but accuracy is not the only important feature of a statistical method's performance. For example, a model resulting in a single probability as opposed to probability trends over time might yield weaker models for implementation into the clinical environment. For nurses, especially those in a hospital, identifying when an event is likely to occur (or at least monitoring trends over time) might be equally as important as the classification outcome of whether an event will occur at any point.

In this study, we compared the accuracy of 2 traditional statistical modeling strategies (logistic regression and Cox proportional hazards regression) and 2 related machine learning strategies (random forest and random survival forest) for in-hospital cardiopulmonary arrest (CPA). We selected these 4 strategies on the basis of their common use in the scientific literature and because 2 of the strategies (logistic regression and random forest) predict a binary outcome, whereas the other 2 strategies (Cox proportional hazards regression and random survival forest) predict a time-to-event outcome (Table 1). The traditional statistical strategies leverage regression methods for classification and survival analyses, and the machine learning strategies average the results of many decision trees created by splitting a random selection of predictor variables in each tree.¹⁶ We evaluated each of the approaches for accuracy and discrimination, the expected number of alarms at select thresholds, and the differences in model outputs with respect to what was being predicted. We hypothesized that the machine learning strategies would provide improved accuracy and discrimination and that the time-to-event models would provide outputs more amenable to human interpretation for evaluation in future work.

Methods

Design and Setting

For this retrospective cohort study, we collected data from deidentified copies of the electronic health records of adults admitted to a large urban academic medical center from 2006 through 2015. A start date of 2006 accounted for changes in the rapid response team's organizational policy, which could have influenced the outcome of interest given that these changes placed increased emphasis on early recognition and management of clinical deterioration outside the intensive care unit. The institutional review board approved the study.

Variables

We defined the outcome of interest (dependent variable) by using *Current Procedural Terminology (CPT)* code 92950 (cardiopulmonary resuscitation). A review of the literature guided our selection of candidate predictor variables, which comprised demographics, vital signs, laboratory values, and *International Classification of Diseases, Ninth Revision (ICD-9)* codes upon hospital admission. For time-to-event outcomes (event day for cases and length of stay for controls), *CPT* codes were the most accurate method for identifying an exact date of care provided in this data source. To identify an initial hospitalization day, we searched patients' records for any of approximately 50 hospitalization *CPT* codes or documentation of a Braden assessment. We identified subsequent hospital days either by these same criteria or by the presence of a complete blood count or basic metabolic profile specimen collection *CPT* code. We constructed a hospitalization stay by combining all sequential dates in which 1 of the aforementioned criteria was met. For patients with an emergency department visit on the day before hospitalization, the emergency department visit date served as the first hospitalization day. Supplemental Figure 1 (available online only at www.ajconline.org) illustrates the distribution of the time-to-event variable for all patients, for cases (event day), and for controls (length of stay).

Sample

We excluded patients who received cardiopulmonary resuscitation on the same day as an emergency department visit or on the first day of hospitalization (Figure 1) to address a more homogeneous population. Among eligible patients with multiple cardio-pulmonary resuscitation events, we retained the index encounter. We defined control patients as hospitalized patients who never experienced a cardiopulmonary arrest (ie, lacked *CPT* code 92950). For control patients with multiple hospitalizations during the study period, we retained the encounter with the least amount of missing data.

Data Analysis: Preprocessing

We began preprocessing by exploring extreme values, patterns of missingness, and collinear associations. We recoded physiologically implausible values (eg, serum sodium < 100 mEq/L, pulse rate > 240 beats per minute) as missing. We removed 10 of the 60 candidate predictor variables because they (a) were missing in more than 80% of patients (eg, blood gas values), (b) were highly collinear with another variable (on the basis of Spearman $\rho > 0.4$

or because they could be predicted by other values in a regression model with > 90% of the variance explained), or (c) had indeterminate time stamps in which the first value was indistinguishable from latter values (eg, blood pressure). A full list of candidate predictor variables with rationales for exclusion is provided in Supplemental Table 1 (available online only). Characteristics of patients for final predictor variables are presented in Table 2.

Because of the large amount of unexplainable missing data (approximately 40% for laboratory values and 60% for vital signs) and the lack of definitive guidelines for handling that magnitude of missing data,¹⁷ we separately performed a statistical simulation study using 10 million patients. We replicated distributions and associations from the empirical data to create a population, imposed several missing data causes (ie, completely at random, at random, and not at random), and tested 3 imputation approaches to identify which method was most accurate under the missing data assumptions. Imputation approaches included missing but assumed normal (similar to a median imputation), multiple imputation without the outcome, and multiple imputation with the outcome. The best approach under most assumptions was multiple imputation with the outcome using chained equations with predicted mean matching; therefore, we used that approach for our study.

Data Analysis: Model Development

We included no longitudinal values (eg, length of stay) or multiple assessments in our model development. This initial work was cross-sectional in nature. We included each predictor variable's first available measure on the first hospitalization day in the following 4 statistical approaches: logistic regression, Cox proportional hazards regression, random forest, and random survival forest. We used logistic multivariable regression to model receipt of cardiopulmonary resuscitation as a dichotomous variable. We used Cox proportional hazards multivariable regression to model the same outcome as a time-to-event variable and allow for censoring.¹⁶ We used machine learning approaches for both binary outcome classification (random forest) and the time-to-event outcome (random survival forest).^{16,18} Both random forest approaches build classification (or time-to-event) trees, each comprising a random sample of predictor variables. Trees are split into branches on the basis of cut points that optimize differences between the 2 new branches. After multiple trees are built, predictions are averaged to develop a forest.

We flexibly fit the logistic and Cox regression models by using restricted cubic splines and no interaction effects between variables. Consistent with multiple imputation, we fit these models to multiple imputed data sets and pooled coefficients and performance metrics across model fits. We performed post hoc analyses of residuals and influential observations and found that reducing the number of knots in the restricted cubic splines from 5 to 3 helped models meet assumptions. We assessed calibration and performed internal validation by using the bootstrap of the last imputed data set from the multiple imputation process. The random forest and random survival forest methods were trained by using 50% of the data from the last imputed data set from the multiple imputation process. We reserved another 25% of the data for testing and the final 25% of the data for validation.

We compared statistical model performance via area under the receiver operating characteristic curve (AUROC) scores and maximum F_1 scores, along with graphic reviews

of receiver operating characteristic curves and recall-precision curves, respectively. We developed hypotheses related to clinical impact and interpretability by comparing positive prediction rate (number of patients triggering an alarm), recall (sensitivity or true-positive rate), and graphical representations of model predictions from pooled and individual patients. We also explored variable importance rankings using a partial Wald χ^2 test minus the predictor's degrees of freedom for the regression models and mean decrease in accuracy (when variable is absent from a tree) for the random forest models. The validation data set held out during the machine learning approaches (25% of the original data) served as the data for direct comparison of the models' expected future performance. Rather than using imputed values from the multiple imputation process, we performed median imputation for missing values to create a data set with greater similarity to the clinical environment, where multiple imputation is not easily feasible. Supplemental Figure 2 (available online only) contains a visual representation of data used for imputation, development, and validation. We performed all analyses with statistics software (R version 3.3.1, the R Foundation).¹⁹ The specific R packages used along with the mathematical formulas used to compare the models are available in the supplementary material (Supplemental Table 2 and Supplemental Figure 3, available online only).

Results

Statistical Performance

From a statistical perspective, all models performed similarly on the basis of AUROC but differed with respect to harmonic mean of recall and precision (F_1 score). AUROC values of the 4 models ranged from 0.847 to 0.861, suggesting good, consistent performance, yet maximum F_1 scores ranged from 0.170 to 0.325, suggesting poor and variable performance (Table 3 and Figure 2). The order of most important variables changed with each model, but *ICD-9* codes associated with respiratory, circulatory, genitourinary, endocrine, and symptom-based diagnoses and diagnostic and therapeutic procedures were among the 10 most important variables in all 4 of the models. Variable importance rankings were similar for logistic regression and the random forest. However, between the survival approaches, the *ICD-9* codes were more influential in the Cox regression model, whereas the clinical variables were more influential in the random survival forest. Additional details of the variable importance differences between models can be found in Supplemental Figure 4 (available online only).

Clinical Impact Performance

From a clinical impact perspective, the random forest and random survival forest models identified more patients than logistic and Cox regression models at the same thresholds for CPA event probabilities ranging from 0.006 (actual event rate) to 0.12 (20 times the event rate) (Figure 3). Similarly, the random forest and random survival forest models had higher sensitivity rates at these same thresholds. With respect to the display of predictions that can be provided to clinicians, logistic regression and random forest models can provide a point estimate probability, whereas the Cox regression and random survival forest models can provide event probabilities that vary across time (even though their coefficients remain fixed).²⁰ Figure 4 illustrates the estimated probability of a CPA event produced by all 4

models for 2 different patients: the average patient obtained by median values for all variables and an “ill” patient with several abnormal values. The random survival forest curve for the ill patient illustrates the most drastic change in predicted probability.

Because the random survival forest predictions showed the largest variability across time for the ill patient, we explored whether the random survival forest model demonstrated a similar degree of variability in predicted probabilities among all patients in our available data set. We averaged the random survival forest prediction curves for all individuals in the data set and compared these against the average Cox regression model predictions for the same individuals. Figure 5 shows that the day-to-day changes in probabilities predicted by the random survival forest curves for patients with and without CPA were much larger than those predicted by the Cox regression model.

Discussion

Using a large data set, we directly compared regression modeling and machine learning techniques for predicting in-hospital CPA. The approaches produced similar AUROC values ranging from 0.847 to 0.861, which are comparable to the findings of other researchers. A recent systematic review of early warning system scores for in-hospital clinical deterioration found most AUROCs in the range of 0.74 to 0.86 for CPA.¹ These moderately large AUROCs should not be surprising given the low event rate of CPA. Another study directly comparing classification modeling strategies for CPA (ie, logistic regression vs machine learning methods) has recently been published,²¹ and the findings differed slightly from ours in that the random forest approach outperformed logistic regression with respect to AUROC (0.801 vs 0.770). The investigators also found that respiratory rate, heart rate, and age were the 3 most important predictor variables, whereas we found several laboratory values to be the most important clinical variables in our models. Of note, they used a composite outcome of non-intensive care unit CPA, unexpected intensive care unit transfer, and death rather than a single end point of CPA.

Conversely, the statistical performance of all modeling approaches was more dissimilar for recall and precision, with F_1 scores of 0.170 to 0.325. The 2 regression models (Cox proportional hazards for time-to-event outcomes and logistic for classification outcomes) performed similarly, with F_1 scores of 0.284 and 0.273, respectively. In contrast, the time-to-event machine learning approach (random survival forest) performed worse than the classification machine learning approach (random forest), with F_1 scores of 0.170 and 0.325, respectively. Unfortunately, we were not able to compare our F_1 scores with those of other studies because these metrics are not frequently reported in CPA prediction literature. With rare events, comparing precision (ie, positive predictive value) is preferable to specificity because of precision's sensitivity to event rate, which can provide insight into the clinical burden of false alarms.²²

The potential clinical influence of the models with respect to number of alarms varied as well. At all thresholds, machine learning approaches produced more clinical alarms than regression approaches (Figure 3). This finding was accompanied by the benefit of increased sensitivity, but too many alarms could contribute to clinicians' alert fatigue. Increased

thresholds decrease the positive prediction rate and recall (sensitivity) while increasing precision (positive predictive value). In our study, increases in precision occurred at increasingly higher thresholds but eventually returned to zero in 3 of the 4 approaches (Figure 2). The random forest model did not exhibit the same behavior, and in fact, precision reached 1 at the most extreme threshold before returning to values similar to those generated by other approaches. For clinical environments where precision is valued more than recall (ie, where certainty in a positive prediction is more important than a false-negative result), the random forest approach could be more appropriate.

In terms of clinical interpretability, we used this study to generate the hypothesis that prediction trends of time-to-event models might be more likely to influence clinicians' decisions. Time-to-event models produce trajectory curves that align more closely with the underlying deterioration phenomenon than does a single probability that is expressed as a straight line on a graph (Figure 4). The display of graphical probability trends offers a potential solution to alarm fatigue that might result from simple numerical cutoffs. Although there does not appear to be a single superior approach at this time, given that the random forest machine learning methods have several advantages (ie, fewer assumptions and increased variability in prediction trends) over the traditional statistical regression models and the time-to-event models allow prediction trends, the random survival forest model might provide the best option for further model development work for in-hospital CPA. Future research to determine what is most likely to influence clinicians' decisions would be helpful.

Strengths and Limitations

We leveraged robust prediction model methods, including flexible regression models and newer machine learning methods. Random forest models have the benefit of fewer predictor variable assumptions than traditional modeling strategies (eg, linearity, interaction effects) and minimal overfitting compared with simple classification and regression trees. A benefit of using survival models is the generation of probability estimate curves that could provide clinicians with a better idea of whether an event might happen earlier or later in a patient's stay. Providing more precise CPA probability estimates is of more value to in-hospital nurses than are measurements of noncritical events, such as 30-day readmission rates or pressure ulcers.

As with all studies dependent on electronic medical records, limitations include missing values and potential variation in accuracy of data (eg, *CPT* codes). Even though some evidence suggests that using a composite measure (eg, CPA, intensive care unit transfer, and mortality) increases statistical power for prediction of clinical deterioration,²³ we used a single outcome of CPA defined by administrative *CPT* codes because this variable had the most accurate time stamp in our data set. We included only data available upon admission even though we expect that adding more values as they become available would increase the predictive accuracy of the model. Several additional approaches exist for repeated-measures data (eg, mixed-effects regression, time-varying covariate survival models, and discrete-time survival models), but starting with a more straightforward approach was a beneficial first step and sets the stage for more robust methods in the near future. These repeated-measures

methods should continue to be explored despite our finding that a single-time model performed similarly to multi-time models with respect to AUROC. In an attempt to increase the signal-to-noise ratio in the current analysis, we included *ICD-9* codes even though these data would not be available for a real-time clinical decision support tool. Future studies could compare the performance of approximated models that are developed with fewer variables (eg, only those that are most commonly available in real time) following the development of a saturated model with many predictor variables.

The amount of missing data further limits the trustworthiness and clinical applicability of our models. We found no evidence that patient characteristics influenced missing data patterns, and thus we assumed data were missing completely at random. On the basis of manual chart review within the research database by subject matter experts, we determined that missing data most likely resulted from data loss during transfer from electronic health records when the organization created the research database. Data loss might be attributed to inadequate data queries or misspecification of data sources, among other causes. The use of multiple imputation with chained equations and predicted mean matching, especially in such a large sample, produced results that were very similar to population/true values in our statistical simulations. Although having more nonmissing values would have been preferred, missing data within clinical records are common, and we cannot simply ignore data that are present by discarding variables with excessive missing data.

Future Directions

Future work should focus on obtaining data sets with fewer missing data and including additional variables that might predict CPA (eg, mental status scales).²⁴ The field of predictive analytics for in-hospital CPA continues to expand, as noted by people publishing prospective protocols²⁵ and testing additional statistical methods, such as discrete-time survival frameworks and generalized linear dynamic models.^{26–28} We excluded patients who experienced CPA on their first day of care because we anticipated that different statistical strategies and model variables would be necessary to represent the phenomenon occurring earlier in a patient's hospitalization (eg, using only emergency room triage data). Although we reported the heuristic advantage of noting trend line displays, we should investigate whether trends or point estimates are more likely to influence nurses' behavior. Nurses' responses to predictive information could be explored within the larger context of design and usability studies.²⁹ We provided information on variable importance; however, these findings could be due to the amount of missing data, and the importance ordering should be revisited in future studies.

Conclusions

As we continue to develop probability-based clinical decision support tools for recognizing clinical deterioration, we must use the most appropriate statistical methods to model the underlying phenomenon. Improvement in accuracy is only one aspect of building decision support tools that are beneficial to clinicians. Potential clinical impact (eg, prediction format or number of alarms) is also an important consideration as we consider usefulness for bedside nurses. If we expect clinicians to incorporate these tools into their clinical

workflows, we must be cognizant of both of these issues. Finally, given the potential impact of decision support interventions on workflow, nurses' roles, and patients' outcomes, we advocate for increased collaboration between nurse scientists and biomedical informatics researchers to develop decision support tools that influence nursing work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

FINANCIAL DISCLOSURES

This research was supported by Clinical and Translational Science Award No. UL1TR000445 from the National Center for Advancing Translational Sciences and by the resources and the use of facilities at the Department of Veterans Affairs, Tennessee Valley Healthcare System. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences, the National Institutes of Health, the Department of Veterans Affairs, or the United States government.

REFERENCES

1. Smith ME, Chiovaro JC, O'Neil M, et al. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. *Ann Am Thorac Soc* 2014; 11(9): 1454–1465. [PubMed: 25296111]
2. Makam AN, Nguyen OK, Auerbach AD. Diagnostic accuracy and effectiveness of automated electronic sepsis alert systems: a systematic review. *J Hosp Med* 2015;10(6):396–402. [PubMed: 25758641]
3. Andersen LW, Berg KM, Chase M, Cocchi MN, Massaro J, Donnino MW; American Heart Association's Get With The Guidelines®-Resuscitation Investigators. Acute respiratory compromise on inpatient wards in the United States: incidence, outcomes, and factors associated with in-hospital mortality. *Resuscitation* 2016;105:123–129. [PubMed: 27255952]
4. Merchant RM, Yang L, Becker LB, et al.; American Heart Association's Get With The Guidelines-Resuscitation Investigators. Incidence of treated cardiac arrest in hospitalized patients in the United States. *Crit Care Med* 2011;39(11):2401–2406. [PubMed: 21705896]
5. Go AS, Mozaffarian D, Roger VL, et al.; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2014 update: a report from the American Heart Association. *Circulation* 2014;129(3):e28–e292. [PubMed: 24352519]
6. Jarvis S, Kovacs C, Briggs J, et al. Can binary early warning scores perform as well as standard early warning scores for discriminating a patient's risk of cardiac arrest, death or unanticipated intensive care unit admission? *Resuscitation* 2015;93:46–52. [PubMed: 26051812]
7. Jarvis S, Kovacs C, Briggs J, et al. Aggregate National Early Warning Score (NEWS) values are more important than high scores for a single vital signs parameter for discriminating the risk of adverse outcomes. *Resuscitation* 2015;87:75–80. [PubMed: 25433295]
8. Kang MA, Churpek MM, Zdravetz FJ, Adhikari R, Twu NM, Edelson DP. Real-time risk prediction on the wards: a feasibility study. *Crit Care Med* 2016;44(8):1468–1473. [PubMed: 27075140]
9. Moss TJ, Lake DE, Calland JF, et al. Signatures of subacute potentially catastrophic illness in the ICU: model development and validation. *Crit Care Med* 2016;44(9):1639–1648. [PubMed: 27452809]
10. Finlay GD, Rothman MJ, Smith RA. Measuring the modified early warning score and the Rothman index: advantages of utilizing the electronic medical record in an early warning system. *J Hosp Med* 2014;9(2):116–119. [PubMed: 24357519]

11. Alvarez CA, Clark CA, Zhang S, et al. Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC Med Inform Decis Mak* 2013;13:28. [PubMed: 23442316]
12. Churpek MM, Yuen TC, Park SY, Gibbons R, Edelson DP. Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards. *Crit Care Med* 2014;42(4):841–848. [PubMed: 24247472]
13. Escobar GJ, LaGuardia JC, Turk BJ, Ragins A, Kipnis P, Draper D. Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. *J Hosp Med* 2012; 7(5): 388–395. [PubMed: 22447632]
14. Evans RS, Kuttler KG, Simpson KJ, et al. Automated detection of physiologic deterioration in hospitalized patients. *J Am Med Inform Assoc* 2015;22(2):350–360. [PubMed: 25164256]
15. Kirkland LL, Malinchoc M, O’Byrne M, et al. A clinical deterioration prediction tool for internal medicine patients. *Am J Med Qual* 2013;28(2):135–142. [PubMed: 22822159]
16. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd ed New York, NY: Springer; 2009.
17. Steyerberg E *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* New York, NY: Springer; 2009.
18. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Statistics* 2008;2(3):841–860.
19. R: A Language and Environment for Statistical Computing [computer program] Version 3.3.1 Vienna, Austria: R Foundation for Statistical Computing; 2016.
20. Harrell F *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* 2nd ed New York, NY: Springer; 2015.
21. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016;44(2):368–374. [PubMed: 26771782]
22. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Crit Care* 2015;19:285. [PubMed: 26268570]
23. Churpek MM, Yuen TC, Edelson DP. Predicting clinical deterioration in the hospital: the impact of outcome selection. *Resuscitation* 2013;84(5):564–568. [PubMed: 23022075]
24. Zadavec FJ, Tien L, Robertson-Dick BJ, et al. Comparison of mental-status scales for predicting mortality on the general wards. *J Hosp Med* 2015;10(10):658–663. [PubMed: 26374471]
25. Xu M, Tam B, Thabane L, Fox-Robichaud A. A protocol for developing early warning score models from vital signs data in hospitals using ensembles of decision trees. *BMJ Open* 2015;5(9):e008699.
26. Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation* 2016;102:1–5. [PubMed: 26898412]
27. Caballero Barajas KL, Akella R. Dynamically modeling patient’s health state from electronic medical records: a time series approach. In: *KDD ‘15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* New York, NY: Association for Computing Machinery; 2015.
28. Kipnis P, Turk BJ, Wulf DA, et al. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *J Biomed Inform* 2016;64:10–19. [PubMed: 27658885]
29. Dunn Lopez K, Gephart SM, Raszewski R, Sousa V, Shehorn LE, Abraham J. Integrative review of clinical decision support for registered nurses in acute care settings. *J Am Med Inform Assoc* 2017;24(2):441–450. [PubMed: 27330074]

Effectively predicting clinical deterioration requires both mathematical accuracy and a consideration of clinicians' needs.

With about 170 000 patients' actual clinical data, we used predictive analytics to identify high-risk individuals.

Evaluation of the 4 prediction models emphasized potential impact on false alarms, in addition to accuracy.

Figure 1.
Selection of patients who received cardiopulmonary resuscitation (CPR).

Figure 2.

Receiver operating characteristic curves (left) and recall-precision curves (right) for logistic regression, Cox proportional hazards regression, random forest, and random survival forest approaches. Evaluation of survival approaches is provided at the median time point, day 2.

Figure 3.

Comparison of positive prediction rate and sensitivity among all models at thresholds comprising the event rate in this data set (0.006) and several of its multiples.

Figure 4.

Comparison of estimated probability of a cardiopulmonary arrest (CPA) event from 2 fictitious patients. Top: average patient defined as all model variables' values set at the median value. Bottom: ill patient characterized by several abnormal values (ie, creatinine = 2 mg/dL [177 μ mol/L], glucose = 300 mg/dL [16.6 mmol/L], potassium = 5 mEq/L [5 mmol/L], sodium = 150 mEq/L [150 mmol/L], hemoglobin = 7 g/dL [70 g/L], red cell distribution width = 20%, respiratory rate = 24/min, pulse = 115/min, and age = 80 years). The y-axis scales are different in the 2 graphs.

Figure 5. Summary curves for all predicted patients, stratified by those with cardiopulmonary arrest (CPA) versus those without CPA. Dashed lines indicate first and third quartiles of the random survival forest.

Table 1Comparison of analytical approaches to predicting in-hospital cardiopulmonary arrest^a

Purpose	Approach	
	Statistical	Machine learning
Classification: predicts whether	Logistic regression	Random forest
Survival/time to event: predicts how likely an event is at each time point	Cox proportional hazards regression	Random survival forest

^aOur chosen statistical approaches leverage regression methods. Our chosen machine learning approaches average the results of many decision trees that have been created by splitting a random selection of predictor variables in each tree.

Table 2Descriptive statistics comparing patients who did and did not receive cardiopulmonary resuscitation^a

Variable	Controls (n = 168177)		Cases (n = 980)		<i>p</i> ^b
	Median (lower-upper quartile]	Mean (SD)	Median (lower-upper quartile]	Mean (SD)	
Age, y	55 (39–69)	54(19)	61 (48–71)	59 (17)	<.001
Respirations, breaths per minute	18 (16–20)	18.6 (4.2)	18 (16–22)	20.2 (6.2)	<.001
Pulse, beats per minute	90 (77–104)	92 (21)	96 (80–113)	97 (22)	<.001
Body mass index ^c	27.7 (23.7–32.9)	29.2 (8.6)	27.4 (23.4–33.4)	29.6 (9.1)	.92
Calcium, mg/dL	8.90 (8.40–9.30)	8.83 (0.81)	8.70 (8.10–9.20)	8.62 (0.96)	<.001
Anion gap, mEq/L	9.0 (7.0–11.0)	9.1 (3.5)	10.0 (8.0–13.0)	10.5 (4.4)	<.001
Glucose, mg/dL	111 (95–139)	129 (69)	123 (102–166)	148 (87)	<.001
Creatinine, mg/dL	0.92 (0.75–1.20)	1.25 (1.36)	1.23 (0.89–1.94)	1.96 (2.15)	<.001
Serum carbon dioxide, mEq/L	25.0 (22.0–27.0)	24.4 (3.9)	24.0 (21.0–27.0)	23.8 (5.2)	<.001
Sodium, mEq/L	138 (136–140)	137.8 (4.0)	137 (134–140)	136.9 (5.6)	<.001
Potassium, mEq/L	3.90 (3.60–4.30)	3.96 (0.62)	4.10 (3.60–4.60)	4.15 (0.81)	<.001
Platelets, x10 ³ /μL	223 (172–283)	235 (107)	211 (149–275)	226 (131)	<.001
White blood cell count, x10 ³ /μL	9.3 (6.9–12.9)	10.7 (8.4)	11.0 (7.6–16.2)	13.7 (15.8)	<.001
Red cell distribution width, %	13.8 (13.1–15.1)	14.4(2.1)	15.2 (13.9–16.9)	15.7 (2.5)	<.001
Hemoglobin, g/dL	12.4 (10.6–14.0)	12.3 (2.4)	11.4 (9.7–13.3)	11.5 (2.6)	<.001
Procedural code					
Urinary system	0.00 (0.00–0.00)	0.01 (0.12)	0.00 (0.00–0.00)	0.02 (0.23)	.01
Integumentary system	0.00 (0.00–0.00)	0.03 (0.18)	0.00 (0.00–0.00)	0.09 (0.67)	.08
Respiratory system	0.00 (0.00–0.00)	0.02 (0.18)	0.00 (0.00–0.00)	0.12 (0.52)	<.001
Nose, mouth, and pharynx	0.00 (0.00–0.00)	0.04 (0.23)	0.00 (0.00–0.00)	0.10 (0.55)	<.001
Nervous system	0.00 (0.00–0.00)	0.06 (0.35)	0.00 (0.00–0.00)	0.18 (0.73)	<.001
Musculoskeletal system	0.00 (0.00–0.00)	0.23 (0.55)	0.00 (0.00–1.00)	0.37 (0.75)	<.001
Male genital system	0.00 (0.00–0.00)	0.00 (0.04)	0.00 (0.00–0.00)	0.00 (0.00)	.36
Hemic and lymphatic system	0.00 (0.00–0.00)	0.00 (0.07)	0.00 (0.00–0.00)	0.03 (0.25)	<.001
Female genital system	0.00 (0.00–0.00)	0.01 (0.09)	0.00 (0.00–0.00)	0.00 (0.10)	.41
Eye	0.00 (0.00–0.00)	0.00 (0.09)	0.00 (0.00–0.00)	0.01 (0.12)	.68
Endocrine system	0.00 (0.00–0.00)	0.00 (0.05)	0.00 (0.00–0.00)	0.00 (0.07)	.86
Ear	0.00 (0.00–0.00)	0.00 (0.04)	0.00 (0.00–0.00)	0.00 (0.00)	.31
Digestive system	0.00 (0.00–0.00)	0.04 (0.30)	0.00 (0.00–0.00)	0.12 (0.75)	.001
Diagnostic and therapeutic	0.00 (0.00–0.00)	0.21 (0.63)	0.00 (0.00–2.00)	1.04(1.84)	<.001
Cardiovascular system	0.00 (0.00–0.00)	0.12 (0.51)	0.00 (0.00–0.00)	0.61 (1.42)	<.001
Diagnostic code					
Blood and blood-forming organs	0.00 (0.00–0.00)	0.29 (0.81)	0.00 (0.00–2.00)	1.12 (1.68)	<.001
Circulatory system	1.00 (0.00–2.00)	1.80 (3.10)	4.00 (0.00–9.00)	5.70 (6.20)	<.001
Congenital anomalies	0.00 (0.00–0.00)	0.04 (0.34)	0.00 (0.00–0.00)	0.11 (0.68)	<.001
Digestive system	0.00 (0.00–0.00)	0.49 (1.36)	0.00 (0.00–2.00)	1.12 (2.14)	<.001

Variable	Controls (n = 168177)		Cases (n = 980)		<i>p</i> ^b
	Median (lower-upper quartile]	Mean (SD)	Median (lower-upper quartile)	Mean (SD)	
Endocrine, nutritional, metabolic, immunity	0.00(0.00–1.00)	0.83 (1.54)	2.00 (0.00–5.00)	2.69 (2.83)	<.001
Genitourinary system	0.00 (0.00–0.00)	0.37 (1.04)	1.00 (0.00–3.00)	1.55 (2.01)	<.001
Infectious and parasitic diseases	0.00 (0.00–0.00)	0.16 (0.59)	0.00 (0.00–1.00)	0.84(1.46)	<.001
Injury and poisoning	0.00 (0.00–1.00)	1.50 (4.40)	1.00 (0.00–3.00)	3.70 (8.20)	<.001
Mental disorders	0.00 (0.00–0.00)	0.53 (1.46)	0.00 (0.00–1.00)	0.53 (1.15)	.04
Musculoskeletal system and connective tissue	0.00 (0.00–0.00)	0.35 (0.98)	0.00 (0.00–0.00)	0.54(1.43)	.002
Neoplasms	0.00 (0.00–0.00)	0.29 (1.41)	0.00 (0.00–0.00)	0.68 (1.94)	<.001
Nervous systems and sense organs	0.00 (0.00–0.00)	0.36 (0.95)	0.00 (0.00–1.00)	0.78 (1.62)	<.001
Nonspecific abnormal findings	0.00 (0.00–0.00)	0.15 (0.47)	0.00 (0.00–0.00)	0.27 (0.65)	<.001
Pregnancy, childbirth, and the puerperium	0.00 (0.00–0.00)	0.22 (1.42)	0.00 (0.00–0.00)	0.03 (0.54)	<.001
Respiratory system	0.00 (0.00–1.00)	0.61 (1.32)	2.00 (0.00–6.00)	3.54 (3.57)	<.001
Skin and subcutaneous tissue	0.00 (0.00–0.00)	0.12 (0.61)	0.00 (0.00–0.00)	0.29 (1.02)	<.001
Symptoms	1.00 (0.00–2.00)	1.50 (2.10)	3.00 (1.00–5.00)	3.30 (3.20)	<.001
Ill-defined or unknown causes of morbidity or mortality	0.00 (0.00–0.00)	0.05 (0.25)	0.00 (0.00–0.00)	0.19 (0.55)	<.001
Supplemental V codes	1.00 (0.00–2.00)	1.40 (2.00)	2.00 (1.00–5.00)	3.30 (3.20)	<.001

^aControl group was 50% female (n = 84 148) and 50% male (n = 83 626); case group was 40% female (n = 393) and 60% male (n = 586); Pearson test, *P* < .001.

^bWilcoxon test.

^cCalculated as weight in kilograms divided by height in meters squared.

Table 3

Performance of statistical modeling approaches

Strategy	AUROC	F_1 score
Logistic regression	0.851	0.273
Cox proportional hazards	0.854	0.284
Random forest	0.861	0.325
Random survival forest	0.847	0.170

Abbreviations: AUROC, area under the receiver operating characteristic curve; F_1 score, harmonic mean of recall and precision.