min-SIA: a Lightweight Algorithm to Predict the Risk of 6-Month Mortality at the Time of Hospital Admission

Nishant Sahni, MD, MS[1] (iD), Roshan Tourani, PhD[2], Donald Sullivan, MD[3], and Gyorgy Simon, PhD[2]

[1]Division of General Internal Medicine, University of Minnesota, Minneapolis, MN, USA; [2]Institute of Health Informatics, University of Minnesota, Minneapolis, MN, USA; [3]Oregon Health Sciences University, Portland, OR, USA.

BACKGROUND: Predicting death in a cohort of clinically diverse, multi-condition hospitalized patients is difficult. This frequently hinders timely serious illness care conversations. Prognostic models that can determine 6-month death risk at the time of hospital admission can improve access to serious illness care conversations.

OBJECTIVE: The objective is to determine if the demographic, vital sign, and laboratory data from the first 48 h of a hospitalization can be used to accurately quantify 6-month mortality risk.

DESIGN: This is a retrospective study using electronic medical record data linked with the state death registry.

PARTICIPANTS: Participants were 158,323 hospitalized patients within a 6-hospital network over a 6-year period.

MAIN MEASURES: Main measures are the following: the first set of vital signs, complete blood count, basic and complete metabolic panel, serum lactate, pro-BNP, troponin-I, INR, aPTT, demographic information, and associated ICD codes. The outcome of interest was death within 6 months.

KEY RESULTS: Model performance was measured on the validation dataset. A random forest model—mini serious illness algorithm—used 8 variables from the initial 48 h of hospitalization and predicted death within 6 months with an AUC of 0.92 (0.91–0.93). Red cell distribution width was the most important prognostic variable. min-SIA (mini serious illness algorithm) was very well calibrated and estimated the probability of death to within 10% of the actual value. The discriminative ability of the min-SIA was significantly better than historical estimates of clinician performance.

CONCLUSION: min-SIA algorithm can identify patients at high risk of 6-month mortality at the time of hospital admission. It can be used to improved access to timely, serious illness care conversations in high-risk patients.

KEY WORDS: palliative care; hospital outcomes; predictive models; data mining.

Abbreviations

aPTT	Activated partial thromboplastin time
AST	Aspartate amino transferase
ALT	Alanine amino transferase
AHRQ	Agency for Health Care Research and Quality
EMR	Electronic medical record
MCV	Mean corpuscular volume
WBC	White blood cell count
CMP	Complete metabolic panel
CBC	Complete blood count
BMP	Basic metabolic panel
BNP	Brain natriuretic peptide
RF	Random forest
Lab	Laboratory
LR	Likelihood ratio
LR	Logistic regression
SD	Standard deviation
AUC	Area under the curve
ROC	Receiver operator curve
OOB	Out of bag
ICD9-CM	International Classification of Diseases 9-Clinical Modification
ICD10	International Classification of Diseases 10
INR	International normalized ratio
MD-Gini	Mean Decrease in Gini Index
MI	Machine learning

Key Statements

- *There is a lack of accurate prognostic models that predict 6-month mortality at the time of hospital admission in multi-condition patients.*
- *We describe the development and validation of a user-friendly, lean, machine learning model (min-SIA) that accurately risk-stratifies patients at the time of hospital admission. min-SIA has a remarkable AUC of 0.92 for death within 6 months of admission. This discriminative ability exceeds AUCs reported for clinician performance in the literature (0.60–0.85)*
- *The min-SIA model is an innovative advancement that can potentially improve the delivery of timely serious illness conversations in appropriate situations.*

INTRODUCTION

Accurate and timely prognostication is essential for ensuring that seriously ill patients receive care that is concordant with their goals and values—a critical component of high-quality care. Early conversations about advance care planning (ACP) with seriously ill patients can improve outcomes for patients and caregivers.[1, 2] However, serious illness care conversations often occur too late, when patients are in crisis or unable to

make decisions for themselves.[3, 4] Of those admitted to hospitals, less than half needing palliative care actually receive it.[5, 6] One of the major barriers to timely serious illness care conversations is the poor prognostic performance of clinicians in predicting longer term mortality.[7] This is especially true in patients who have a number of serious, chronic medical conditions—such patients constitute the largest proportion of hospice utilizers in the USA. Timely prognostication and referral remains the Achilles heel for timely serious illness discussions.[8] Despite prior efforts, a relatively simple predictive model to accurately prognosticate 6-month mortality in diverse, multi-condition patients at the time of hospital admission remains elusive.[3, 6] Existing prognostic models in this area frequently rely extensively on variables that may not be available to clinicians at the time of hospital admission, or use a very large number of variables—making them unwieldy to use.[9] In this study, we aim to develop a user-friendly predictive model that estimates the probability of 6-month mortality at the time of hospital admission. Such a model will facilitate the objective and timely identification of high-risk hospitalized, "multi-condition" patients.

METHODS

Inclusion Criterion

After obtaining institutional review board approval, we used our institution's clinical data warehouse to create an EMR-derived dataset of hospital admissions for 158,323 patients within a 6-hospital network in the Twin Cities area, Minnesota. The encounters spanned a 6-year period ranging from 2012 to 2018. The hospital system consists of one 450-bed university tertiary care center and 5 community hospitals ranging from 100 to 450 beds in capacity. Patients who were less than 18 years of age who did not consent to their medical record being used for research purposes or had less than a year of follow-up mortality data were excluded. We included hospitalizations to all units and services as long as they met the above criterion. Vital status and death dates were obtained from the state death registry. Our database had the complete death record issued from 2011 onwards for deceased individuals who were born in Minnesota, had died in Minnesota, or have ever had a permanent address in the state.

Model Variables

Our dataset included 4 broad classes of variables (features) that were very commonly available in the EMR from most hospitalizations and were clinically relevant: (1) *demographic variables*: age, sex, race; (2) *physiologic variables*: systolic blood pressure, diastolic blood pressure, pulse, respiratory rate, temperature, pulse-oximetry readings, and body mass index; (3) *biochemical variables*: serum sodium, potassium, chloride, bicarbonate, creatinine, urea nitrogen, ALT, AST, alkaline phosphatase, total bilirubin, albumin, white blood cell count, hematocrit, hemoglobin, platelet count, mean corpuscular volume, red cell distribution width, troponin, pro-BNP, INR, aPTT, arterial blood gas results; (4) *clinical co-morbidity variables*: We created a co-morbidity profile for each patient across the 30 classes of diseases in the AHRQ co-morbidity category index from ICD codes billed during an encounter.[10]

All laboratory and physiologic data was time-stamped and was obtained within 48 h of the admission. For each data element, the first available measurement within 48 h of hospital admission was used in the model. The primary outcome of interest was predicting whether death occurred within 6 months of hospital admission.

Missing Data

We tested two imputation strategies to deal with missing data. The first was the *k*-nearest neighbor approach, which replaced missing data in an encounter with the values of its nearest neighbor based on a distance measure.[9] The second was the median imputation approach where missing values for a variable were replaced with median values for the variable.[9] The two approaches did not change model performance significantly. Due to its simplicity and fast computation time, the median imputation was used.

Dataset Partitioning

The dataset was partitioned into a derivation dataset and a validation dataset with encounters selected randomly at a ratio of 0.6/0.4.

Modeling

We compared the performance of logistic regression (LR) to a class of ML models known as random forest (RF) models.[9, 11] Due to their higher discriminative performance, we focused on developing the RF models. A more detailed explanation of RFs can be obtained by referring to existing reviews on this methodology.[11] RFs are known for their superior "out of box" performance and are able to handle non-linear data and are less prone to over-fitting.[11–13] RFs are based on decision trees. Decision tree algorithms formulate decision rules to fit the underlying data. However, decision trees are frequently "unstable" and are sensitive to minor alterations in the data.[11] RFs aggregate the results of many different decision trees, in order to eliminate this instability. RFs utilize two basic strategies to achieve this objective: (1) The algorithm utilizes a random subset of the training data to build each new tree in the ensemble; (2) A random subset of features is utilized for constructing each decision rule in a tree. This approach avoids introducing an in-ordinate degree of bias in the classification, stemming from a few influential observations.[13, 19] Variable importance is interpreted in RFs, by using an importance measure known as "Mean Decrease in Gini index" (MD-Gini). MD-Gini measures a variable's performance by randomly permuting it and measuring the resultant change in

classification error. [14] For each RF classifier, 501 trees were used in the ensemble in our analysis. The mtry parameter which is the number of variables randomly sampled as candidates at each split was sqrt(p) where p is number of variables in the model. We used the RF implementation from the "ranger" package in R for our analysis.

Statistical Tests

For non-normal variables, median values with interquartile range (IQR) were reported. Mean with standard deviation (SD) was reported for normal variables. The significance of comparisons between two non-normal continuous variables was tested using the Wilcoxon test. For comparisons between two categorical variables, the Fisher test was used.

Model Validation and Testing

The discriminative performance of the models was measured by constructing receiver operator curves (ROC) and calculating the area under the curve (AUC) on the validation dataset. In clinical studies, the AUC gives the probability that a randomly selected patient who experienced an event (e.g., a disease or condition) had a higher risk score than a patient who had not experienced the event. It is equal to the area under the receiver operating characteristic (ROC) curve and ranges from 0.5 to 1. The 95% confidence intervals around the AUC estimates were estimated using the DeLong method, which is implemented in the pROC package in R.[15] In order to evaluate whether the predicted probability of 6-month mortality from the random forest model reflected the observed probabilities,

we constructed model calibration curves. In a perfectly calibrated model, all the points would fall along the diagonal straight line.

RESULTS

Characteristics of the Cohort

The demographic, physiological, and laboratory characteristics of the encounters are shown in Table 1. In 8.1% of the hospitalizations, death occurred within 6 months of hospital admission (Table 1). The median age, creatinine, blood urea nitrogen (BUN), mean corpuscular volume (MCV), white blood count (WBC), and red cell distribution width (RDW) were higher in hospitalizations that were followed by death within 6 months (Table 1). The albumin and hemoglobin readings were significantly lower for patients who died within 6 months of hospital admission (Table 1).

Variable Importance in the Model

We called our algorithm the serious illness algorithm (SIA). The highest-ranking 25 features of SIA are shown in Figure 1. We also constructed min-SIA8, an algorithm that had only 8 variables: red cell distribution width, age, presence of metastatic disease or tumor, serum albumin, platelet count, creatinine, and total bilirubin. We also tested a 10-variable algorithm called min-SIA10. min-SIA10 included serum alkaline phosphatase and hemoglobin in addition to the variables in min-SIA8.

Table 1 Cohort Characteristics

	Survival greater than 6 months	Survival less than 6 months	p value
Number of patients	145,478	12,845	
	0 (0.0)	12,845 (100.0)	< 0.001
Age (years, median [IQR])	50.16 (20.02)	72.32 (15.79)	< 0.001
Sex (=female, number %)	92,871 (63.8)	6423 (50.0)	< 0.001
Race (=white, number %)	119,615 (82.2)	11,299 (88.0)	< 0.001
Metastatic disease (number, %)	4530 (3.1)	3479 (27.1)	< 0.001
Tumor (number, %)	13,283 (9.1)	4722 (36.8)	< 0.001
Serum albumin (g/dL, median [IQR])	3.80 [3.30, 4.20]	3.10 [2.60, 3.60]	< 0.001
Total bilirubin (mg/dL, median [IQR])	0.60 [0.40, 0.90]	0.70 [0.50, 1.20]	< 0.001
Blood urea nitrogen (mg/dL, median [IQR])	15.00 [11.00, 20.00]	24.00 [16.00, 37.00]	< 0.001
Serum chloride (mEq/L, median [IQR])	104.00 [101.00, 106.00]	102.00 [97.00, 105.00]	< 0.001
Serum bicarbonate (mEq/L, median [IQR])	26.00 [24.00, 28.00]	26.00 [23.00, 29.00]	0.939
Serum creatinine (mg/dL, median [IQR])	0.84 [0.70, 1.04]	1.05 [0.77, 1.59]	< 0.001
Blood glucose (mg/dL, median [IQR])	109.00 [94.00, 133.00]	118.00 [99.00, 150.00]	< 0.001
Hemoglobin (mmol/L, median [IQR])	12.70 [11.20, 14.10]	11.40 [9.70, 13.10]	< 0.001
International normalized ratio (IU, median [IQR])	1.06 [0.99, 1.24]	1.25 [1.09, 1.77]	< 0.001
Serum lactate (mmol/L, median [IQR])	1.50 [1.00, 2.20]	2.00 [1.30, 3.50]	< 0.001
Mean corpuscular volume (femtoliters/cell, median [IQR])	90.00 [86.00, 93.00]	92.00 [87.00, 97.00]	< 0.001
Serum sodium (mEq/L, median [IQR])	139.00 [136.00, 141.00]	137.00 [134.00, 140.00]	< 0.001
Platelet count (X 10^9/L, median [IQR])	218.00 [175.00, 267.00]	203.00 [138.00, 275.00]	< 0.001
Pro-brain natriuretic peptide (pg/mL, median [IQR])	1267.50 [300.00, 4120.00]	4130.00 [1267.50, 11,200.00]	< 0.001
Red cell distribution width (μm, median [IQR])	13.50 [12.90, 14.40]	15.20 [13.90, 17.10]	< 0.001
White blood cell count (X 10^9/L, median [IQR])	8.90 [6.70, 12.00]	9.60 [6.80, 13.70]	< 0.001
Body mass index (kg/m^2)	28.13 [24.37, 32.77]	25.10 [21.59, 29.66]	< 0.001

The table is stratified by whether death occurred 6 months of hospital admission. Interquartile ranges are listed in parentheses. Median values are reported for non-normal variables. The Wilcox test is used for comparisons between the non-normal continuous variables, and the Fisher test is used for comparisons between categorical variables. IQR, interquartile range; SD, standard deviation
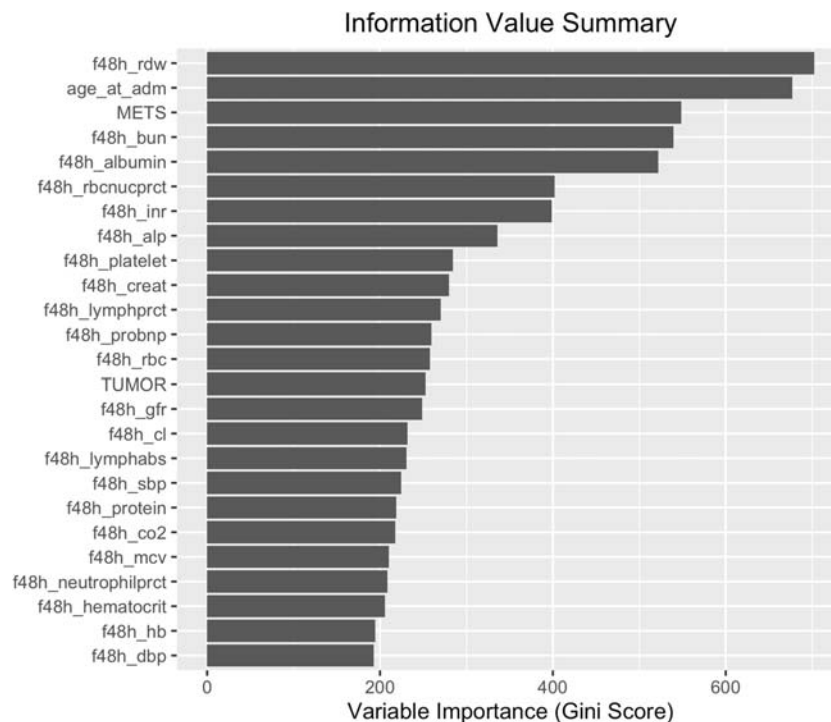
Fig. 1 Feature importance in the random forest models. The 25 highest-ranking features in the SIA model ranked by importance (as measured by the Mean Decrease in Gini Index). f48_ prefix refers to values obtained within the first 48 h of a hospitalization.

Discriminative Performance of Models for Death Within 6-Months of a Hospital Admission

The SIA model with all available predictors in the dataset (54 predictors) had an AUC of 0.94 (0.93–0.95). The leaner models with 8 and 10 variables—min-SIA8 and min-SIA10—had AUCs of 0.92 (0.91–0.92) and 0.93 (0.91–0.93), respectively. The ROC curve for min-SIA8 is shown in Figure 2.

Model Calibration

The calibration of a model is a measure of how well the probabilities estimated by the model reflect the observed prob-
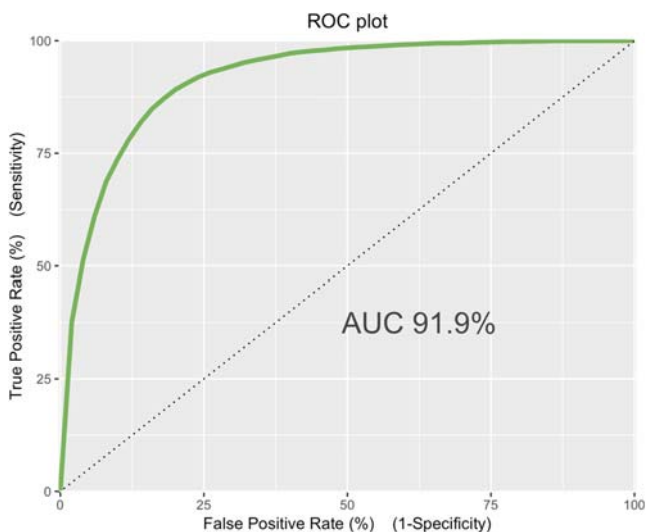

Fig. 2 The receiver operator curve for the minSIA8 model.

abilities. min-SIA8 and min-SIA10 had excellent calibration across the whole probability range. Even though that SIA had a higher AUC, the min-SIA models were better calibrated.

Models' Cumulative Gains, Recall, and Accuracy

The cumulative gains or recall plot (Fig. 3, left panel) visualizes the percentage of targets selected at a certain threshold of predicted probability (k%). For example, if patients within the top 20% of predicted probability range were selected, we would "capture" 83% of the patients that died within 6 months with the min-SIA8 model (Fig. 3, left panel). For the full SIA model, this number would be 88% (data not shown). At a threshold of the top 20% of predicted probability (k = 20%), the accuracy of the min-SIA8 is 85.3% (Fig. 3, right panel).

Addressing Potential Selection Bias

We repeated the model development and testing using one of two approaches: approach (1): In this approach, each distinct hospitalization was treated as a unit of analysis. We used the last set of data from each available hospitalization for each patient; approach (2): In this approach, each unique patient was treated as a unit of analysis. The dataset was sampled and one hospitalization for each patient was randomly selected for inclusion analysis (i.e., random admission model). This was done to test the effect of potential selection bias that could be theoretically introduced by using multiple data-points from the same patient. Both these strategies yielded models with nearly identical AUCs and predictive performance.[9]
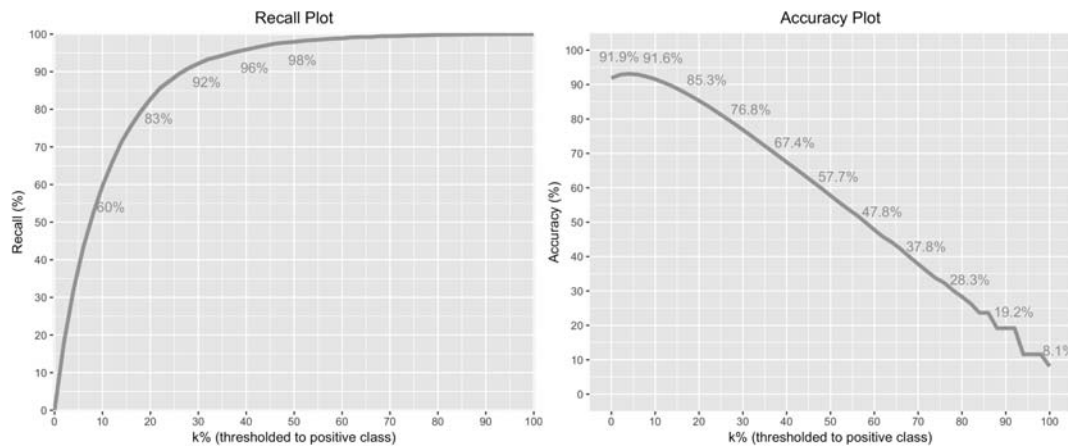
Fig. 3 Left panel: The recall plot shows the percentage of the overall number of cases in a given category "gained" (*y*-axis) when we apply the min-SIA8 and select the highest *k*-deciles (*x*-axis). For example, if the positivity threshold is set to be the highest-ranking 20% cases (by predicted probability), then 83% of true positives would be selected. Right panel: The accuracy plot plots the accuracy (*y*-axis) of the model at each decile threshold of predicted probability (*x*-axis).

DISCUSSION

We demonstrate that it is possible to accurately identify patients who have a high risk of 6-month mortality at the time of hospital admission. We construct and validate min-SIA8, which is a high-performing and lightweight model. min-SIA uses data that is typically available during the first 48 h of admission to clinicians and delivers remarkable discriminative performance. To our knowledge, SIA and min-SIA have the one of the highest AUCs described for predictive models in multi-condition, hospitalized patients.[9, 16] The min-SIA relies on 8 predictors and is fairly easy to use for clinicians. The probability estimates produced by the model closely mirror the observed rates of mortality as demonstrated in the calibration curve.

Clinicians are very poor at estimating the probability of survival beyond a few days—even in intensive care unit patients.[17] Clinician discriminative ability has ranged from an AUC of 0.5 to 0.79 for 6-month survival, in prior studies.[8, 17] Compared with this, the min-SIA has an excellent AUC at 0.92. Other studies looking at longer term mortality estimation (3–12 months) in multi-condition hospitalized patients have achieved an AUC of around 0.94 with a deep-learning approach and 0.91 with random forests.[9, 16] However, both these studies used a much larger number of predictors than min-SIA and relied on data beyond what was available at the time of hospital admission, thereby limiting their use at the beginning of a hospitalization.[9, 16]

Even though the SIA has higher AUC (0.94) than the min-SIA, the min-SIA was better calibrated and used a much smaller number of variables. The difference in discriminative performance was not large enough to be clinically meaningful. Given that min-SIA retains the excellent performance of the SIA—but achieves this by using a much lower number of variables, we selected the min-SIA for web-deployment (https://niceguy.shinyapps.io/sia8/) for purposes of demonstration.

When these models are deployed at a systems wide level (such as with automatic EMR interfacing), it is possible to capture 83–88% of patients that die within 6 months of admission, by screening patients in the top 2 decile of predicted probability (Fig. 3, left panel). This would facilitate automated flagging of high-risk patients for clinical review. Such a strategy would ensure that the majority of patients that could benefit from a serious illness consultation would be identified in a timely manner.

It is notable that red cell distribution width emerged as the single most important variable in our prognostic model. Strikingly, it outperformed age as a prognostic factor. While previous studies have shown that red cell distribution width (RDW) is linked with mortality,[18] our data highlights that RDW is central to prognostication. The mechanistic underpinnings of the link between mortality and RDW are not entirely clear. Some studies have suggested that RDW may be correlated with several different biological pathways such as chronic inflammation, malnutrition, and blood disease.[19] It may therefore be viewed as an unspecific and general "chronic disease prognostic marker."[19]

Our models were developed and validated on a clinically diverse cohort (Supplementary Table 1). Our dataset includes data from a large multi-hospital health system. The system encompasses a university tertiary care center and urban, suburban, and semi-rural hospitals. Ultimately, our model needs to be validated in other settings in order to demonstrate geographic and temporal portability. We used state death registry data for ascertaining the date of death (for out-of-hospital deaths). If a death were not reported to the Minnesota state registry, then it would not be captured in our dataset.

We demonstrate that it is possible to develop high-performing, parsimonious, predictive models—such as min-SIA8—to accurately identify patients at high risk for 6-month mortality at the time of hospital admission. This could potentially be used in areas where accurate risk-

stratification is crucial, such as institutional implementations of serious illness care programs and outcomes research.[20] Future work is needed to test how to incorporate this model into clinical workflow, in order to improve the delivery of timely serious illness care conversations in appropriate situations. Care will have to be taken that any such model implementation is part of a comprehensive serious illness care program designed around the bedrock principles of autonomy, beneficence, non-maleficence, justice, privacy, and confidentiality.

Corresponding Author: *Nishant Sahni, MD, MS; Division of General Internal Medicine University of Minnesota, Delaware Street SE, MMC 741, Minneapolis, MN, USA (e-mail: nishant.sahni@gmail.com).*

Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

Disclaimer: *The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.*

REFERENCES

1. **Curtis JR**, **Downey L**, **Back AL**, **et al.** Effect of a patient and clinician communication-priming intervention on patient-reported goals-of-care discussions between patients with serious illness and clinicians: a randomized clinical trial. *JAMA Intern Med* 2018; 178: 930–940.
2. **Bernacki R**, **Hutchings M**, **Vick J**, **et al.** Development of the Serious Illness Care Program: a randomised controlled trial of a palliative care communication intervention. *BMJ Open*; 5:e009032. 2015. https://doi.org/10.1136/bmjopen-2015-009032.
3. **You JJ**, **Downar J**, **Fowler RA**, **et al.** Barriers to goals of care discussions with seriously ill hospitalized patients and their families: a multicenter survey of clinicians. *JAMA Intern Med* 2015; 175: 549–56.
4. **Anderson WG**, **Chase R**, **Pantilat SZ**, **et al.** Code status discussions between attending hospitalist physicians and medical patients at hospital admission. *J Gen Intern Med* 2011; 26: 359–366.
5. **Hanson LC**, **Rowe C**, **Wessell K**, **et al.** Measuring palliative care quality for seriously ill hospitalized patients. *J Palliat Med* 2012; 15: 798–804.
6. **Bernacki RE**, **Block SD**. Communication about serious illness care goals. *JAMA Intern Med* 2014; 174: 1994.
7. **Fine PG**. Hospice underutilization in the U.S.: the misalignment of regulatory policy and clinical reality. *J Pain Symptom Manage* 2018; 56: 808–815.
8. **White N**, **Reid F**, **Harris A**, **et al.** A systematic review of predictions of survival in palliative care: how accurate are clinicians and who are the experts? *PLoS ONE*; 11:e0161407. 2016. DOI: https://doi.org/10.1371/journal.pone.0161407.
9. **Sahni N**, **Simon G**, **Arora R**. Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multi-condition patients: a proof-of-concept study. *J Gen Intern Med*. 2018. https://doi.org/10.1007/s11606-018-4316-y.
10. **Tabak YP**, **Sun X**, **Nunez CM**, **et al.** Using electronic health record data to develop inpatient mortality predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS). *J Am Med Informatics Assoc* 2014; 21: 455–463.
11. **Breiman L**. Random forests. *Mach Learn* 2001; 45: 5–32.
12. **Shi T**, **Horvath S**. Unsupervised learning with random forest predictors. *J Comput Graph Stat* 2006; 15: 118–138.
13. **Diaz A**, **Bourassa MG**, **Guertin M-C**, et al. Long-term prognostic value of resting heart rate in patients with suspected or proven coronary artery disease. *Eur Heart J* 2005; 26: 967–74.
14. **Breiman L**. Random forest. *Mach Learn* 1999; 45: 1–35.
15. **DeLong ER**, **DeLong DM**, **Clarke-Pearson DL**. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44: 837–845.
16. **Avati A**, **Jung K**, **Harman S**, **et al.** Improving palliative care with deep learning. In: *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*. 2017, pp. 311–316.
17. **Detsky ME**, **Harhay MO**, **Bayard DF**, **et al.** Discriminative accuracy of physician and nurse predictions for survival and functional outcomes 6 months after an icu admission. *JAMA - J Am Med Assoc* 2017; 317: 2187–2195.
18. **Hunziker S**, **Stevens J**, **Howell MD**. Red cell distribution width and mortality in newly hospitalized patients. *Am J Med* 2012; 125: 283–291.
19. **Zurauskaite G**, **Meier M**, **Voegeli A**, **et al.** Biological pathways underlying the association of red cell distribution width and adverse clinical outcome: results of a prospective cohort study. *PLoS One* 2018. https://doi.org/10.1371/journal.pone.0191280.
20. **Bernacki R**, **Hutchings M**, **Vick J**, **et al.** Development of the Serious Illness Care Program: a randomised controlled trial of a palliative care communication intervention. *BMJ Open*. 2015. https://doi.org/10.1136/bmjopen-2015-009032.