Predicting In-Hospital Mortality in the SUPPORT2 Critical Care Dataset

Xinxin Zhou
Brown University (DATA 1030)
GitHub: https://github.com/NoraZhouXX/1030-Final-Project-SUPPORT2

1. Introduction

Accurate assessment of in-hospital mortality risk in critically ill patients supports clinical decision-making and resource planning. However, mortality prediction is sensitive: models may reflect clinician behavior (such as treatment-limiting decisions) and demographic variations, necessitating cautious interpretation of strong predictive performance. This project utilizes the SUPPORT2 dataset, which includes 9,105 critically ill adult patients admitted to five U.S. teaching hospitals between 1989 and 1994. The dataset contains demographic characteristics (age, gender, race, income), diagnostic categories, comorbidities, functional status (including measures of activities of daily living), and physiological/laboratory indicators recorded on the third day of the study. The prediction target is hospdead, a binary indicator for in-hospital mortality, with 0 = non-survival and 1 = survival. With an outcome incidence of 25.9%, a simple "always-alive" baseline model achieves 74.1% accuracy but lacks clinical discriminative value. Previous studies using the SUPPORT cohort primarily focused on long-term prognosis. For example, Knaus et al. (1995) developed and validated an 180-day survival model for critically ill hospitalized adults based on diagnosis, age, pre-admission length of stay, cancer status, neurological function, and 11 physiological indicators collected on day 3; This model achieved an area under the ROC curve (AUC) of approximately 0.79 in both the development and validation cohorts, with further improved discriminatory ability when combined with physician predictions (AUC≈0.82). This study inspired multiple variable designs in the SUPPORT2 model, including derived severity scores (e.g., APS, SPS) and physician predictions. This project focused on in-hospital mortality and explicitly excluded outcome-related variables and potentially sensitive variables that might cause information leakage (e.g., cost/fee data and model-predicted survival) prior to fitting machine learning models.

2. Explanatory Data Analysis

To gain insights into the dataset and the target variable, exploratory data analysis was performed to visualize the distribution and correlation of some key features. The EDA revealed three practical issues. First, the entire dataset contained substantial missing data, with highly uneven distribution across variables: Figure 1 shows approximately 62% of patients had missing ADL patient (adlp) data, while multiple laboratory indicators on Day 3 exhibited missing rates ranging from 25% to 55% (urine output, blood glucose, BUN, albumin, bilirubin, PaO_2/FiO_2, pH).

Second, the target variable exhibits moderate imbalance (mortality rate: 25.9%), potentially rendering threshold-dependent metrics like accuracy misleading. Therefore, area under the ROC curve (AUC) was prioritized for evaluation. Third, clinical severity variables clearly distinguish outcomes. For example, both age and APS are higher on average among patients who die in the hospital (Figure 2). These patterns are consistent with clinical intuition and suggest that a model should capture both baseline vulnerability (age, comorbidities, functional status) and acute physiologic derangement (APS/SPS, labs, coma score).

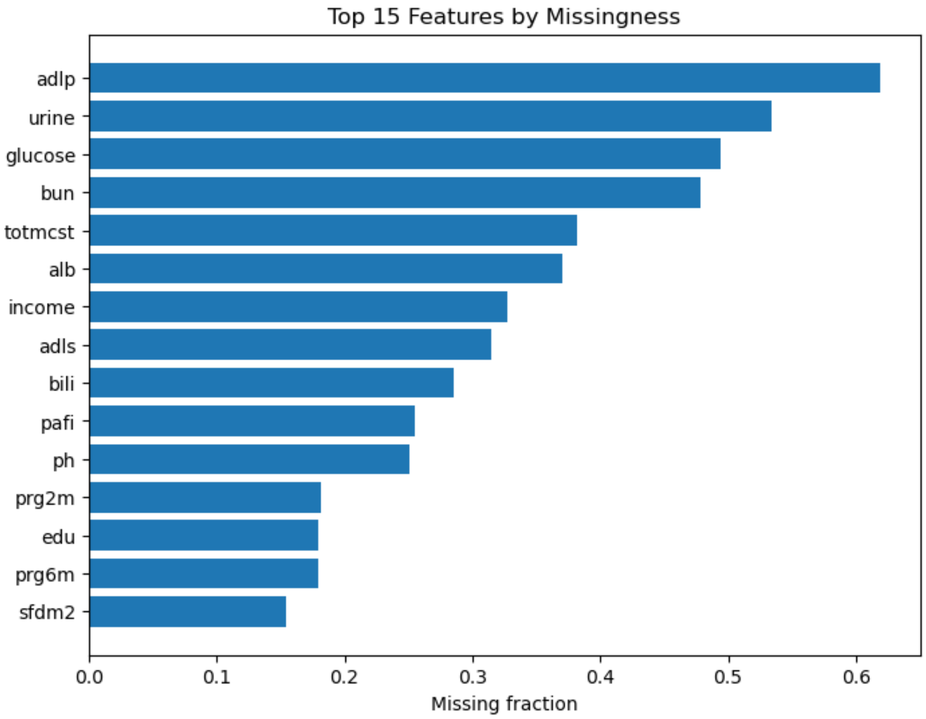

Figure 1. *Top 15 variables by missingness in SUPPORT2. Several lab/physiology variables (e.g., urine output, glucose, BUN) and ADL measures contain substantial missingness, motivating careful handling.*

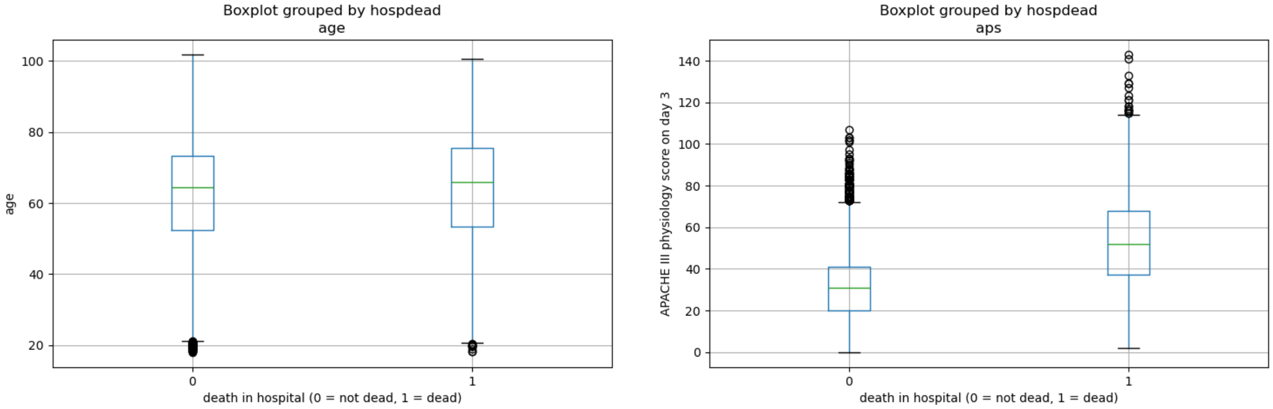

Figure 2. *Age and APS distributions by in-hospital outcome. Patients who die tend to be older and have higher APS.*

3. Methods

3.1 Evaluation Metrics

Given that this is a binary classification problem with moderate class imbalance in the dataset, the area under the ROC curve (AUC) was selected as the primary evaluation metric. This approach avoids the need for thresholding and demonstrates robustness to class imbalance. For completeness, the accuracy and F1 score of the positive class were also reported.

3.2 Data Splitting

To quantify the uncertainty introduced by data splitting, a stratified three-way splitting scheme was employed, with each iteration repeated five times using different random seeds. Each iteration utilized a 60%/20%/20% split ratio for the training-validation-test sets (first splitting 80%/20% for other/test, then splitting the other set 75%/25% for train/val). Report the mean ± standard deviation of evaluation metrics across the five random seed iterations. For random models (e.g., Random Forest, XGBoost), I aligned the model's random state with the splitting seed, so the reported variance reflects both data splitting variability and algorithmic randomness.

3.3 Data Cleaning & Preprocessing

The original dataset contained 48 columns. I first excluded the index column and columns that could potentially lead to data leakage (features measured after day 3), including hospital charges/costs, length of stay variables, follow-up time, and pre-calculated survival predictions. The final feature set comprised 31 predictor variables covering: demographic characteristics; diagnostic groups/categories; comorbidities; Do Not Resuscitate (DNR) status; functional status (ADL); and Day 3 physiological/laboratory variables. Categorical features were padded with a constant "missing" category and one-hot encoded. Ordinal variables (e.g., income, cancer stage, ADL scale) were explicitly sorted and encoded with OrdinalEncoder. Continuous variables underwent standardization; missing values were not imputed during preprocessing and remained in NaN format.

3.4 Handling missingness for continuous features

Logistic regression, linear SVM, and random forests cannot directly handle NaN values. Reduced-feature analysis was applied: data were grouped based on the distribution patterns of observed/missing values in continuous variables, and a "feature-reduced" model was trained separately for each pattern—using only the continuous columns observed in that pattern (plus the always-available coded categorical/ordinal predictor variables). During prediction, patients are routed to the corresponding model based on their missing value pattern. In contrast, XGBoost naturally handles missing values by learning a default split direction for NaNs in each tree, so it can be trained on the full feature set without reduced-feature patterns.

3.5 Hyperparameter Tuning and Cross Validation

Four algorithms were compared: logistic regression (linear), linear support vector machines (linear), random forests (nonlinear), and XGBoost gradient boosting trees (nonlinear). Hyperparameters were tuned using fine-grid search (see Table 1). For logistic regression/support vector machines/random forests, optimal hyperparameters were selected based on validation set AUC values within each split. For XGBoost, 5-fold cross-validation was performed on the combined training + validation set with early stopping enabled in each fold; the parameter combination maximizing the cross-validated mean AUC was ultimately selected and evaluated on the held-out test set.

Model	Hyperparameter Tuned	Best Values
Logistic Regression	C ∈ {0.01, 0.1, 1, 10, 100}	C: 0.1
Linear SVM	C ∈ {0.0001, 0.001, 0.01, 0.1, 1}	C: 0.01

Random Forest	max_depth = {1, 3, 10, 30},
	max_features = {0.05, 0.1, 0.25, 0.5}
		max_depth = 10,
		max_features = 0.1

XGBoost	max_depth ∈ {1, 3, 10, 30},
	reg_alpha ∈ {0.01, 0.1, 1},
	reg_lambda ∈ {0.1, 1, 10},
	n_estimators = 10000 with early stopping
		max_depth = 3, reg_alpha = 0.1,
		reg_lambda = 1

Table 1. *Models and hyperparameter tuning ranges used for training and evaluation.*

4. Results

4.1 Predictive performance

Table 2 and Figure 3 summarize the test performance across five random splits. The majority class baseline AUC value is 0.5. All four machine learning models demonstrate strong discrimination capabilities (ROC AUC≈0.92–0.93), significantly outperforming the majority class baseline. XGBoost delivered the best overall performance, achieving a test mean AUC of 0.925±0.005 and an F1 score of 0.755±0.011. Compared to the baseline AUC of 0.50, this represents an improvement of (0.925 - 0.50)/0.005≈79 standard deviations. Logistic regression and linear support vector machines performed nearly identically, both achieving mean AUC ≈ 0.922 and F1 scores around 0.74. This indicates that well-regularized linear decision boundaries suffice to capture most predictive signals within the SUPPORT2 features. Random Forest also achieved competitive discrimination (AUC≈0.921), but its F1 score was lower (≈0.69) and exhibited greater variability across splits. This indicates that the added nonlinearity did not translate into improved classification performance for the minority class (death) in this scenario. For interpretability analysis (feature importance and SHAP), I refitted the XGBoost model using a representative dataset (random_state=42), achieving AUC=0.933, accuracy=0.891, and F1 score=0.785 on the retained test set.

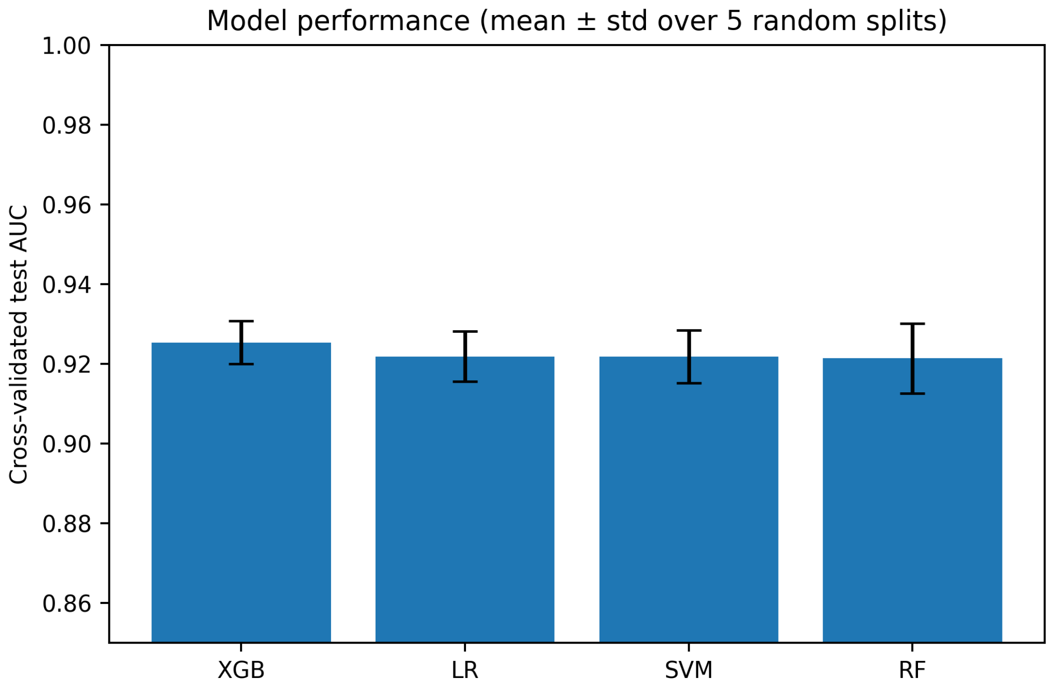

Figure 3. *Test ROC AUC (mean ± std) over five stratified random splits. XGBoost achieves the highest average AUC.*

Model	ROC AUC (mean ± std)	Accuracy (mean ± std)	F1 (positive class) (mean ± std)
Baseline (majority)	0.500 ± 0.000	0.741 ± 0.000	0.000 ± 0.000
Logistic Regression	0.922 ± 0.006	0.872 ± 0.007	0.737 ± 0.016
Linear SVM	0.922 ± 0.007	0.873 ± 0.008	0.739 ± 0.017
Random Forest	0.921 ± 0.009	0.861 ± 0.013	0.688 ± 0.043
XGBoost	0.925 ± 0.005	0.877 ± 0.006	0.755 ± 0.011

Table 2. *Model performance on held-out test sets (mean ± std across 5 random splits). Baseline is the majority-class classifier.*

4.2 Global feature importance

Three global importance metrics were calculated for the XGBoost model: (i) XGBoost internal importance (gain; Figure 4), (ii) permutation importance (Figure 5), and (iii) SHAP global importance (Figure 6). Across the three metrics, the feature DNR (Do Not Resuscitate) emerged as the single most influential predictor, followed by physiological severity (SPS, APS), neurological function (coma score), and functional status (ADL). Several laboratory indicators (BUN, albumin, WBC, bilirubin) and respiratory rate also contributed, while demographic variables such as race and income were relatively minor in this model. Notably is the dominant role of DNR status. DNR is not a biological indicator but reflects treatment preferences and clinical decisions—decisions themselves influenced by prognosis. Thus, high predictive power does not imply causation. Models incorporating DNR should not be used for automated triage without careful consideration of feedback mechanisms and fairness.

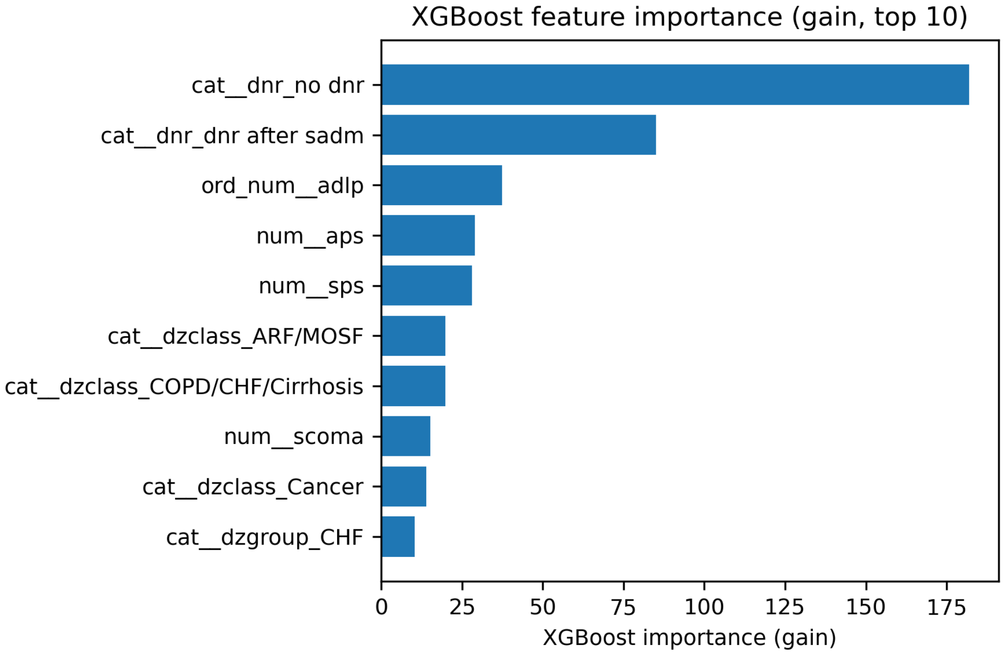

Figure 4. *XGBoost built-in global feature importance (gain). DNR status and severity scores (SPS/APS) dominate.*

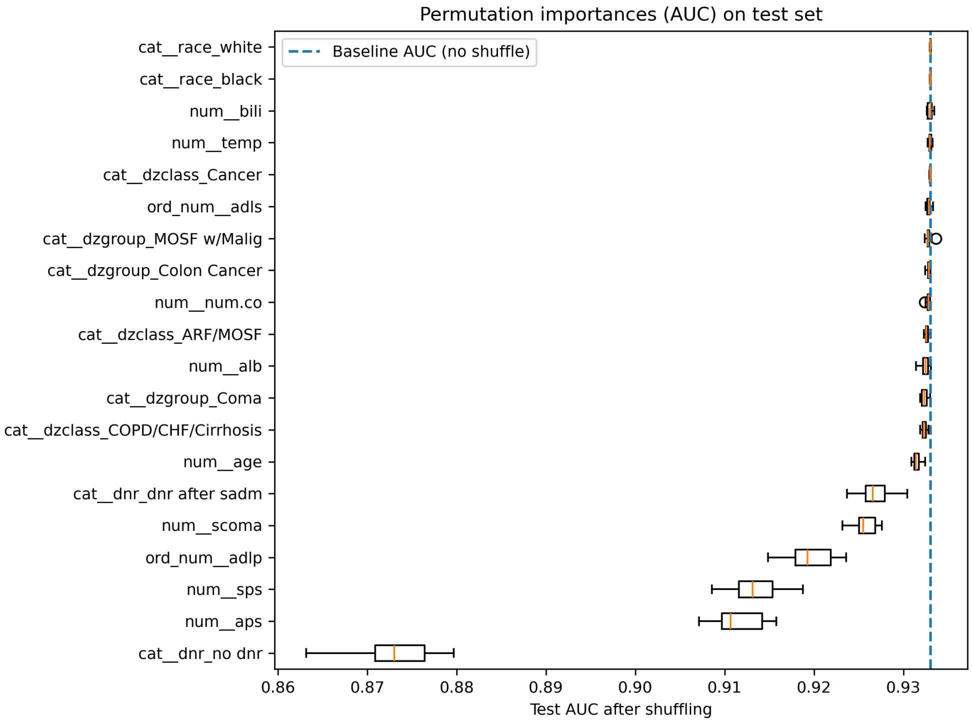

Figure 5. *Permutation importance measured by drop in test ROC AUC when a feature is shuffled. DNR, APS, SPS, ADL, and coma score are the strongest contributors.*

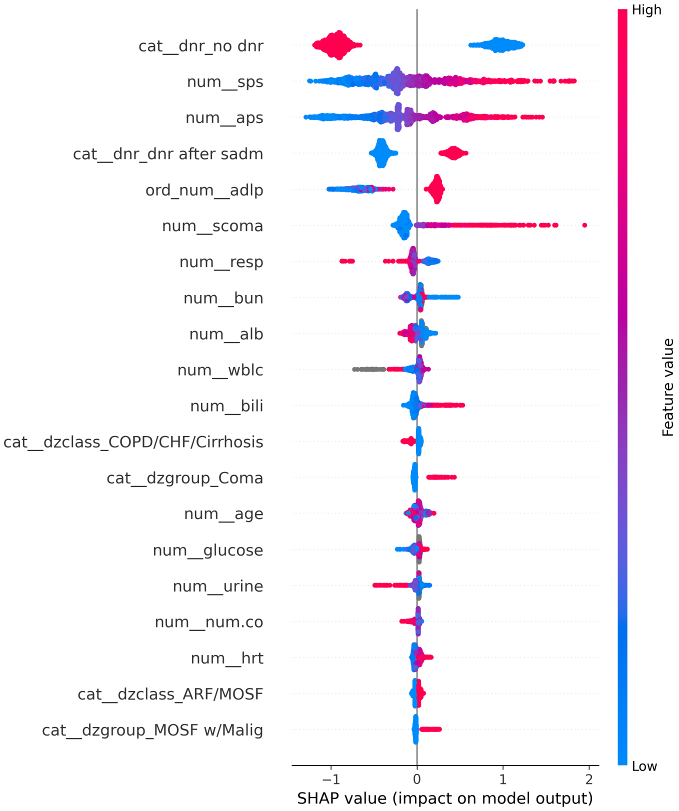

Figure 6. *SHAP summary plot for the XGBoost model. Features are ordered by mean |SHAP|, highlighting DNR and severity scores.*

4.3 Local Feature Importance

Figure 7 shows an example SHAP force plot for an individual test patient. In this low-risk example, the absence of a DNR order, relatively preserved functional status, and lower severity scores all push the prediction strongly toward survival. A mildly elevated bilirubin level slightly increases the risk but is not enough to overturn the overall low-risk prediction.

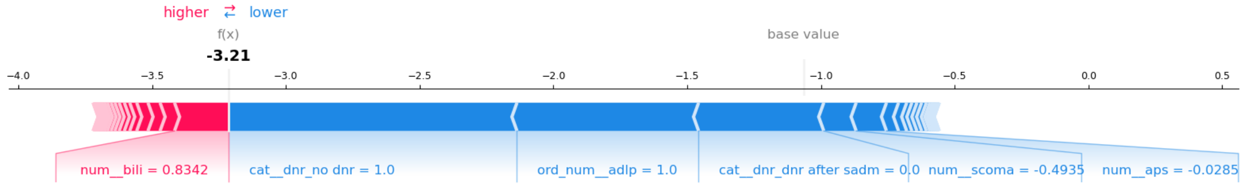

Figure 7. *Local SHAP force plot for an individual test patient, illustrating how DNR and severity variables increase predicted mortality risk.*

5. Outlook

To further refine this project, the following measures can be implemented:

First, the dataset can be improved or supplemented. SUPPORT2 provides only a “snapshot” of physiological indicators around Day 3. However, modern ICUs routinely collect high-frequency time-series data on vital signs, laboratory parameters, and interventions. Integrating longitudinal trends with treatment information can capture clinical instability not reflected in static features.

Second, for the interpretability part, the permutation importance results indicate that only a small number of variables (e.g., DNR status, APS score, SPS score) contribute the majority of marginal predictive signal. Once these core predictors are incorporated into the model, many lower-ranked features contribute almost nothing to the AUC value. This likely stems from redundancy and strong correlations among clinical variables. Further improvement lies in systematically exploring feature selection or integrating correlated predictors into composite scores.

6. References

[1] Knaus, W. A., Harrell, F. E., Jr, Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Jr, Dawson, N. V., Fulkerson, W. J., Jr, Califf, R. M., Desbiens, N., Layde, P., Oye, R. K., Bellamy, P. E., Hakim, R. B., & Wagner, D. P. (1995). The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. Study to understand prognoses and

preferences for outcomes and risks of treatments. Annals of internal medicine, 122(3), 191–203. https://doi.org/10.7326/0003-4819-122-3-199502010-00007

[2] The SUPPORT Principal Investigators. (1995). 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). Journal of the American Medical Association, 274(20), 1591–1598. https://doi.org/10.1001/jama.1995.03530200027032