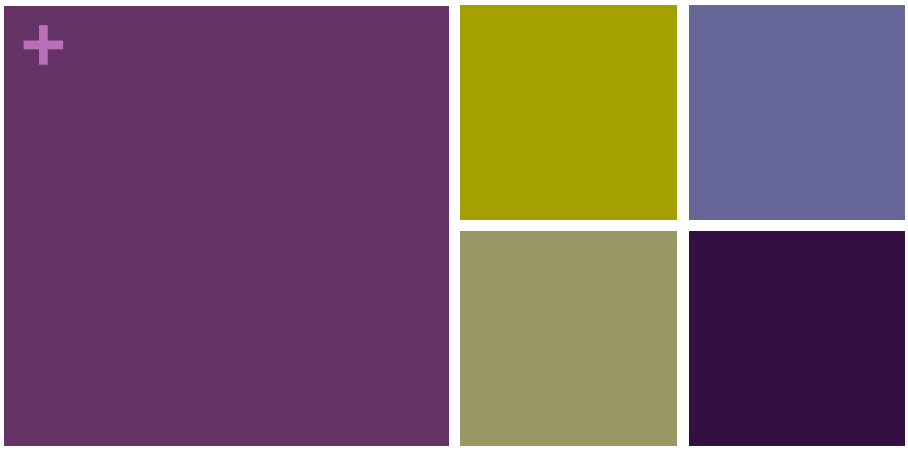


Statistical Modeling Approaches for CDS Tools Predicting In-Hospital Cardiopulmonary Arrest

Alvin D. Jeffery, PhD, RN
Vanderbilt University School of Nursing

Dissertation: Specific Aims

1. Compare the performance of logistic regression, survival analysis, random forests, and random survival forests in the development and validation of a dynamic predictive model for in-hospital cardiopulmonary arrest based on patient characteristics

AJCC AMERICAN JOURNAL OF CRITICAL CARE, September 2018, Volume 27, No. 5

Advancing In-Hospital Clinical Deterioration Prediction Models

By Alvin D. Jeffery, PhD, RN, Mary S. Dietrich, PhD, MS, Daniel Fabbri, PhD, Betsy Kennedy, PhD, RN, Laurie L. Novak, PhD, Joseph Coco, MS, and Lorraine C. Mion, PhD, RN

Aim #1: Statistical Model

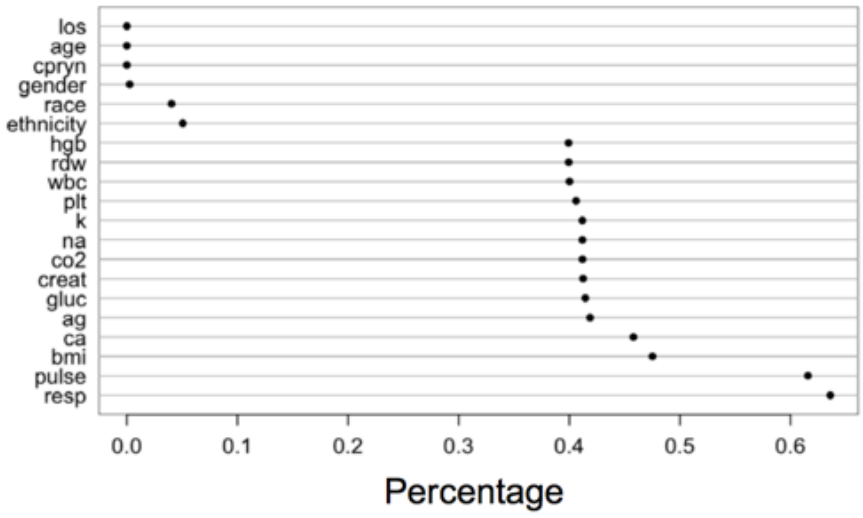
- Data Source:
  - VU Synthetic Derivative (~170,000 eligible patients, ~0.6% event rate)
- Dependent Variable
  - Cardiopulmonary Arrest (CPT Code 92950) after 1st Hospital Day
- Independent Variables (from 1st Hospital Day)
  - Patient Characteristics (demographics/labs/vital signs)
  - ICD-9 Codes
- Data Analysis
  - Classification vs. Time-to-Event
  - "Traditional" vs. Machine Learning

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

Purpose	Approach	
	Statistical	Machine learning
Classification: predicts whether an event will occur	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

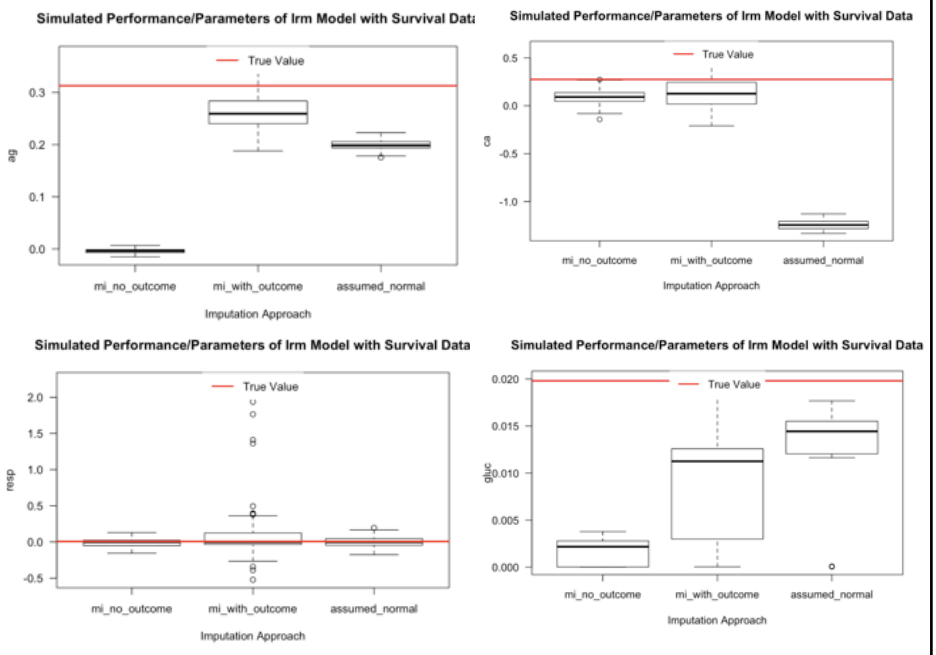
[a] Our chosen statistical approaches leverage regression methods. Our chosen machine learning approaches average the results of many decision trees that have been created by splitting a random selection of predictor variables in each tree.

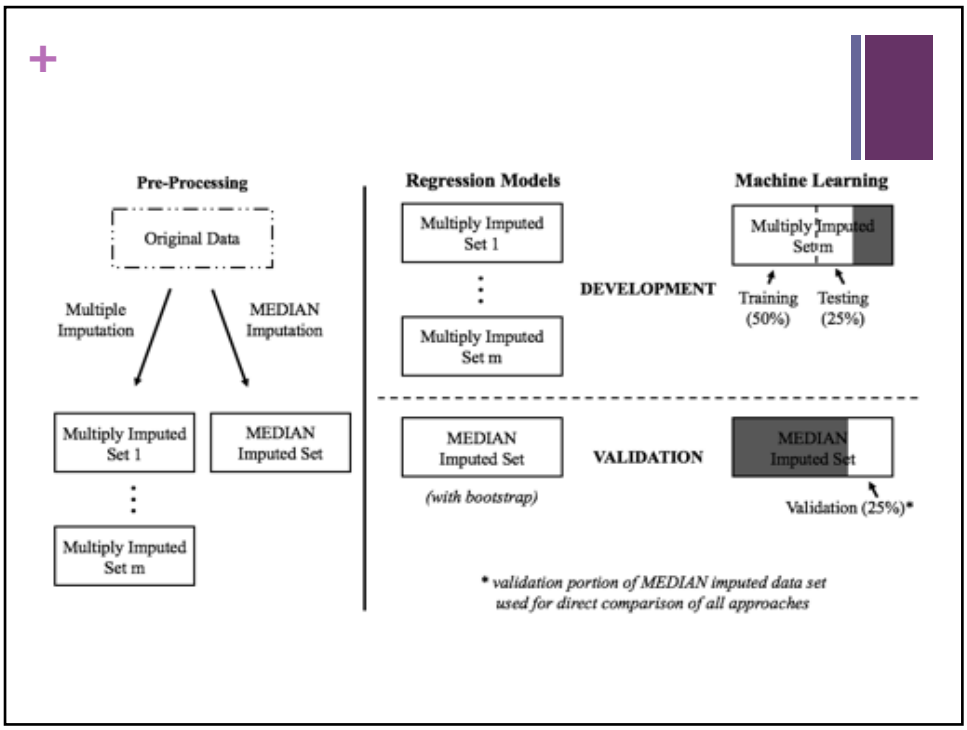
Figure 1. Percentage of missing values for each variable.

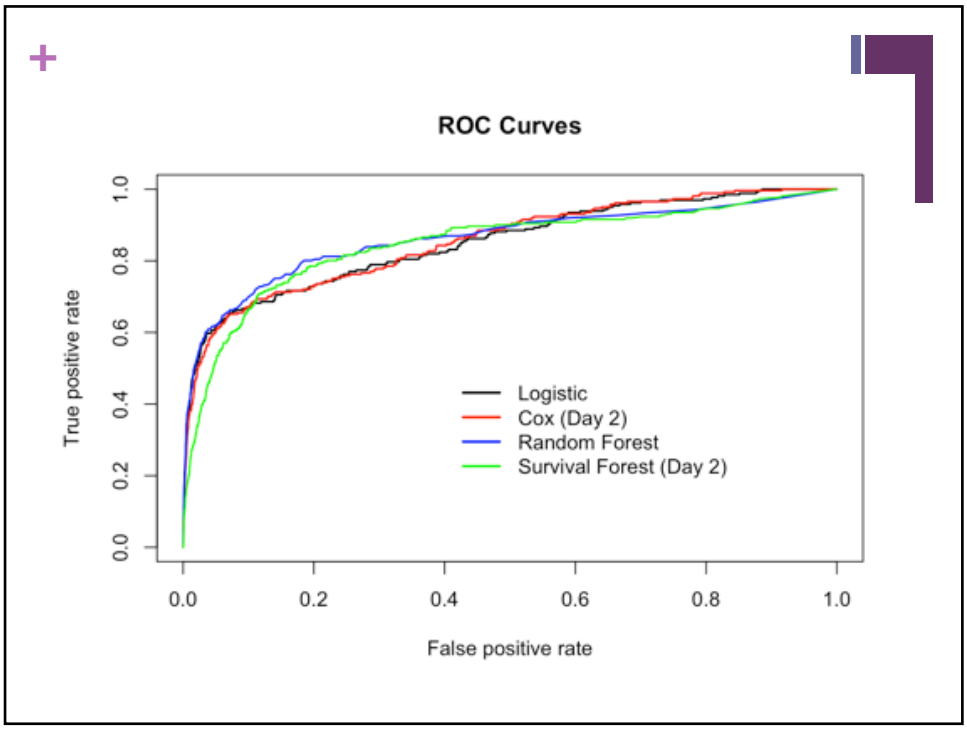


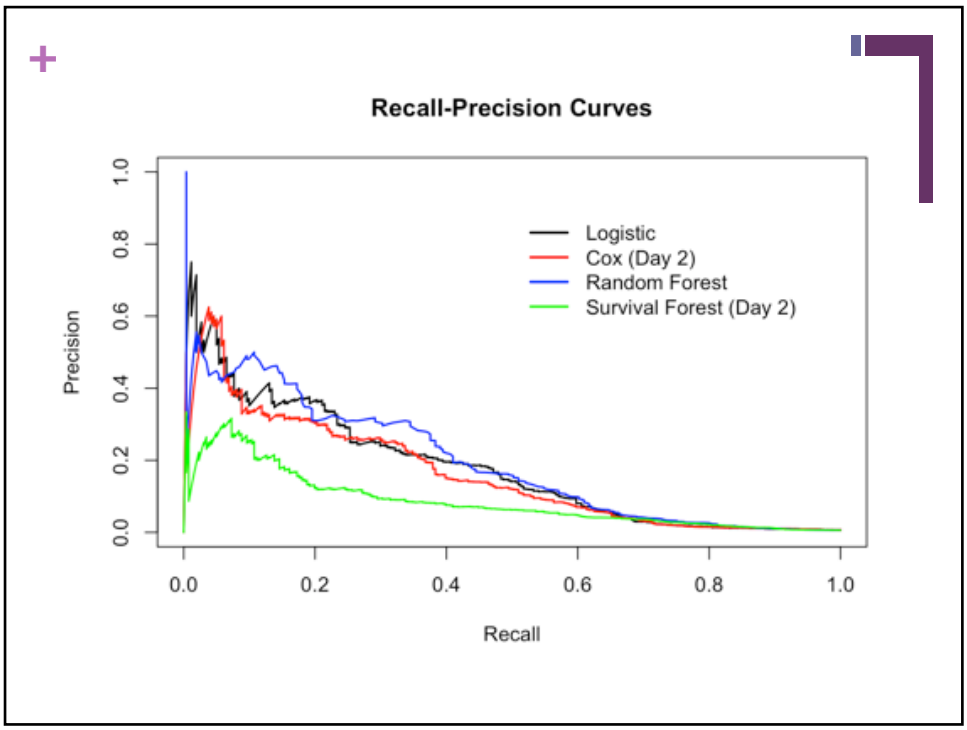
Side Track: Simulation

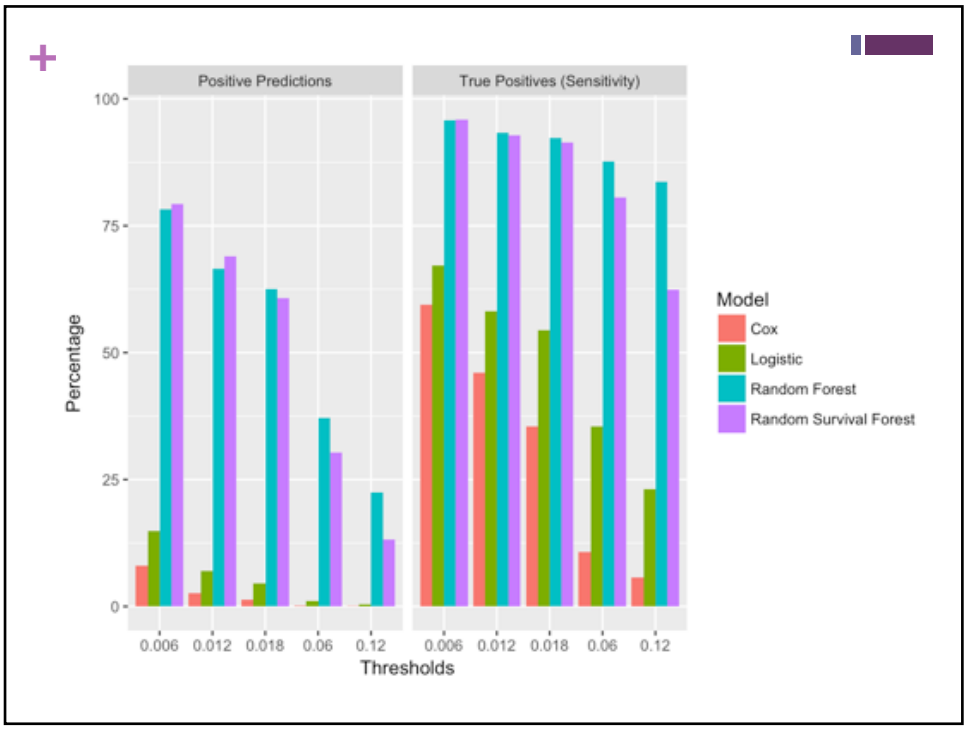
- 10,000,000 patients using 2 approaches:
  - Mimicking variance-covariance matrix of empirical data
  - Mimicking time-to-event distributions & predictors' association with that distribution
- 100 iterations of:
  - Sample 150,000 patients without replacement
  - Induce MCAR & MAR at different proportions
  - Impute with Predicted-Mean-Matching with Chained Equations (similar to Multiple Imputation) +/- Outcome and Missing-Assumed-Normal
  - Build Logistic & Cox PH Models on Imputed Samples
- Compared sample results with "True" population values

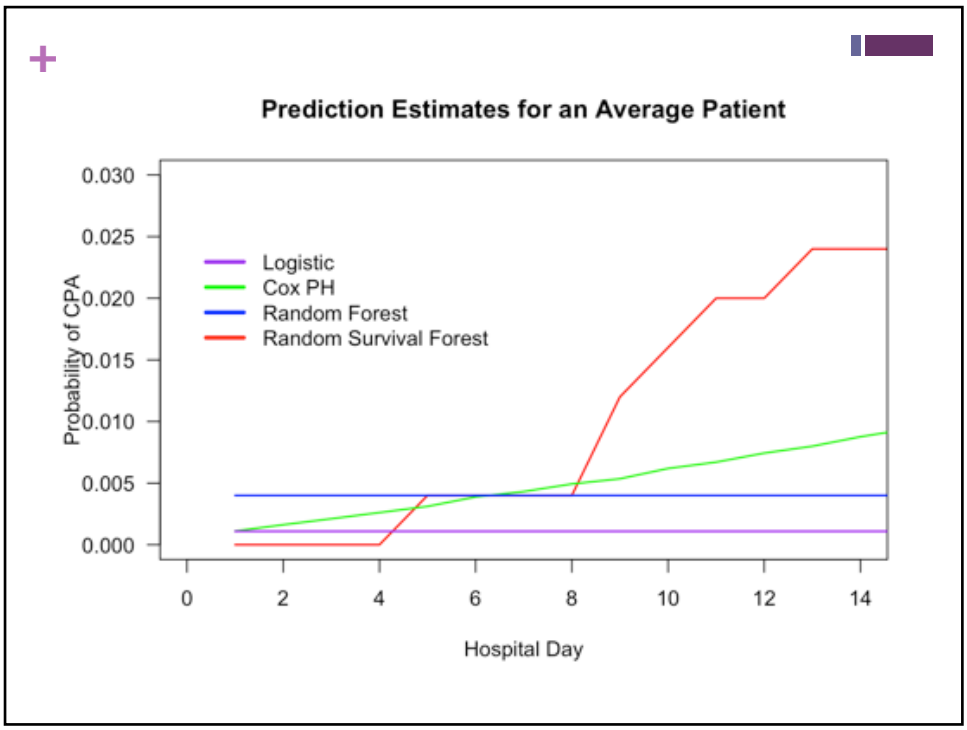


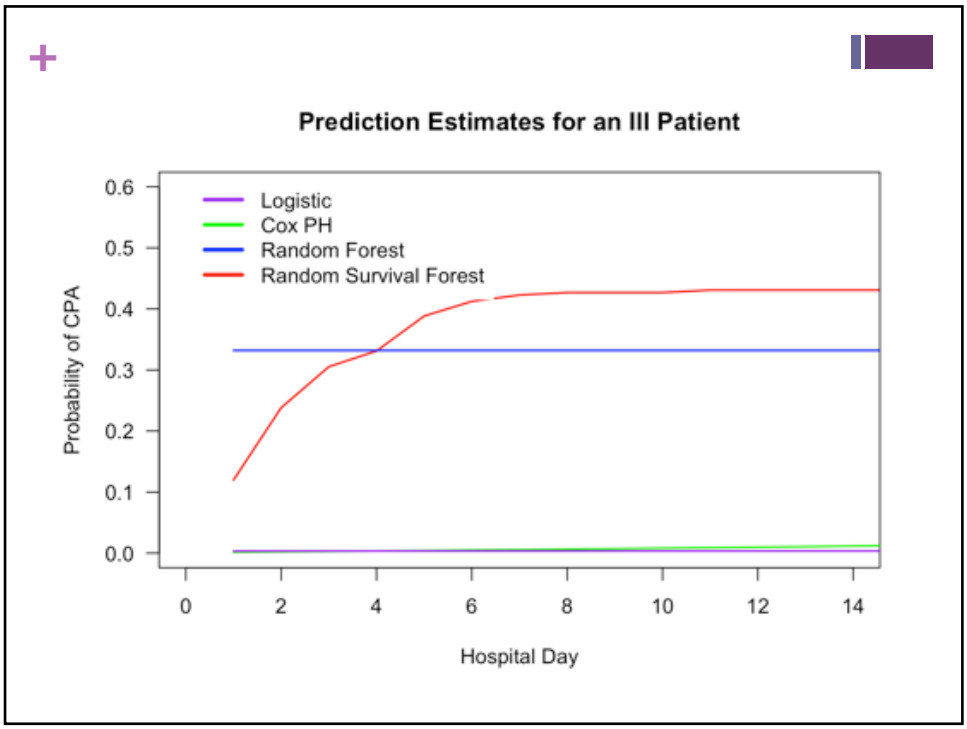

Pre-Processing
Original Data
Multiple Imputation
MEDIAN Imputation
Multiply Imputed Set 1
Multiply Imputed Set m
MEDIAN Imputed Set
Regression Models
Multiply Imputed Set 1
Multiply Imputed Set m
DEVELOPMENT
MEDIAN Imputed Set
(with bootstrap)
VALIDATION
Machine Learning
Multiply Imputed Set m
Training (50%)
Testing (25%)
MEDIAN Imputed Set
Validation (25%)*
* validation portion of MEDIAN imputed data set used for direct comparison of all approaches

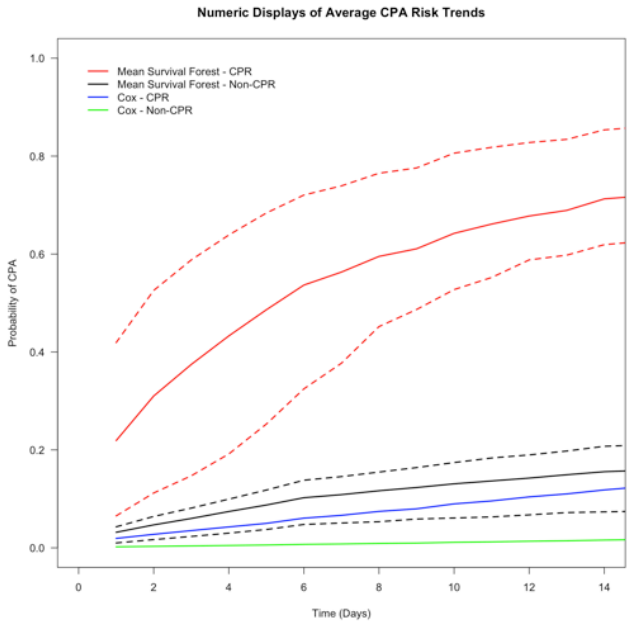

ROC Curves
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic
Cox (Day 2)
Random Forest
Survival Forest (Day 2)


Recall-Precision Curves
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Logistic
Cox (Day 2)
Random Forest
Survival Forest (Day 2)


Positive Predictions
True Positives (Sensitivity)
Percentage
100
75
50
25
0
0.006
0.012
0.018
0.06
0.12
Thresholds
Model
Cox
Logistic
Random Forest
Random Survival Forest


Prediction Estimates for an Average Patient
Probability of CPA
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Logistic
Cox PH
Random Forest
Random Survival Forest
0
2
4
6
8
10
12
14
Hospital Day


Prediction Estimates for an Ill Patient
Probability of CPA
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Logistic
Cox PH
Random Forest
Random Survival Forest
0
2
4
6
8
10
12
14
Hospital Day



Acknowledgements

U.S. Department of Veterans Affairs

VANDERBILT School of Nursing

VAQS QUALITY SCHOLARS
DEVELOPING LEADERS & SCHOLARS TO IMPROVE HEALTHCARE

CRISS CENTER FOR RESEARCH AND INNOVATION IN SYSTEMS SAFETY

NIH National Center for Advancing Translational Sciences

UL1TR000445

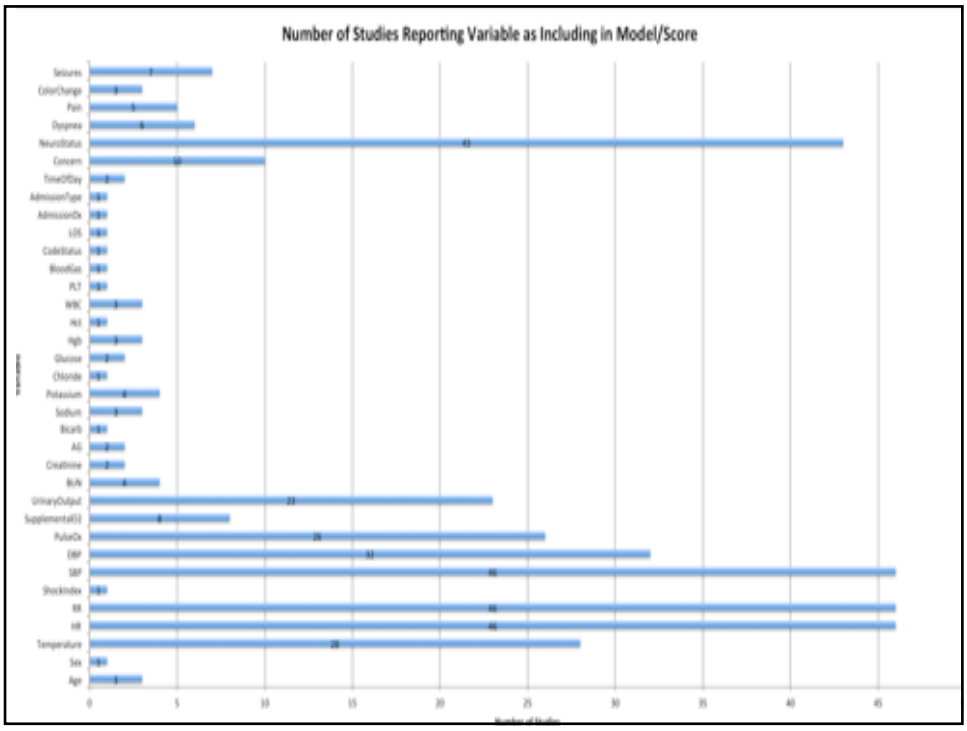
AHRQ Agency for Healthcare Research and Quality
Advancing Excellence in Health Care

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.

Alvin D. Jeffery

Contact

Email: alvinjeffery@gmail.com
Twitter: @Nurse_Alvin
LinkedIn: https://www.linkedin.com/in/alvinjeffery/
Website: http://www.quantitativenurse.com



Variable	Inclusion in Final Models	Reason for Exclusion
Age	Yes	
Gender	Yes	
Race	No	Small sample in some categories resulted in a singular matrix during model fits
Ethnicity	No	Small sample in some categories resulted in a singular matrix during model fits
Body Mass Index	Yes	
Heart Rate	Yes	
Respiratory Rate	Yes	
Blood Pressure	No	Data source listed all timestamps at 00:00, so we were unable to determine first value
Sodium	Yes	
Potassium	Yes	
Chloride	No	Could be predicted by other variables in a regression model with $R^2 > 0.9$
Glucose	Yes	
Blood Urea Nitrogen	No	Collinear with Creatinine (Spearman's rho ~ 0.4)
Creatinine	Yes	
Anion Gap	Yes	
Calcium	Yes	
Carbon Dioxide	Yes	
White Blood Cell Count	Yes	
Red Blood Cell Count	No	Collinear with Hemoglobin (Spearman's rho ~ 0.8)
Hemoglobin	Yes	
Platelet Count	Yes	
Red Cell Distribution Width	Yes	
Blood Gas Panel*	No	Missing in > 80% of patients
Braden Score	No	Missing in > 80% of patients
ICD-9 Codes	Most	The Obstetrical procedure category was removed because it resulted in a singular matrix during model fits
CPT Codes	No	Only used for outcome variables