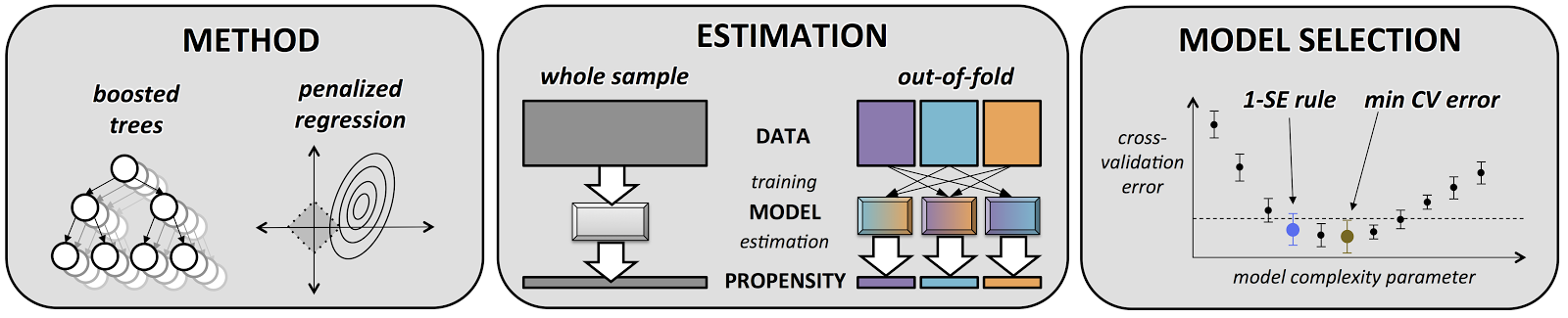
**APPENDIX: The impact of acute organ dysfunction on long-term survival in sepsis**

**Supplemental Methods.**

Our propensity score sensitivity analytic approach was based on gradient boosted tree models. Gradient boosted trees use additive ensembles of decision trees to predict the target value or exposure (acute organ dysfunction) for each observation (patient) based on all features (pre-sepsis ICD9 codes). Their accuracy results from an iterative fitting procedure used to generate the final model: the first tree is trained to predict the target value from the covariates, the second tree is trained to predict the difference between the first tree’s prediction and the target value, the third tree predicts the difference between the sum of the first two trees predictions’ and the target value, and so on. In this fashion, each tree corrects for the errors made by the previous trees.

We fixed the depth of each individual tree at an interaction depth of 3 levels and a learning rate at 0.085. We selected the number of trees based on 5-fold cross-validation and the ‘one standard error’ (1-SE) rule: instead of using the number of trees that empirically minimized the cross-validation error, we used the minimum number of trees that resulted in a cross-validation error that was statistically indistinguishable from the minimum observed cross-validation error.[1](#_ENREF_1) Using the same model fit on a dataset to make predictions for that same dataset can cause overfitting, which in this case results in inflated propensity scores assigned to patients who truly experienced organ dysfunction and deflated scores assigned to those who did not. While this effect is less notable when using inflexible linear models, it becomes pronounced when using non-linear approaches (e.g. gradient boosted trees) because the model can “memorize” what the exposure was for each patient. To prevent this overfitting, we used four-fifths of the data to fit a model with the selected number of trees and used that model to estimate the propensity scores for the held-out fifth. We repeated this process across each fold to estimate propensity scores for each patient. These strategies (e.g., the 1-SE rule and *out-of-fold prediction*) – termed as the ‘novel’ approach – were designed to ensure that information about a patient's actual exposures did not leak into the same patient’s estimated propensity scores (Appendix Figure 1). The ‘traditional’ modeling approach does not include the out-of-fold prediction or 1-SE rule.

**Appendix Figure 1.** Novel methods for fitting sophisticated propensity score models. Our propensity scoring approach was designed to use either gradient boosted trees or penalized linear regression (left panel). Because gradient boosted trees demonstrated improved matching performance, our subsequent analyses only used tree-based approaches. Using models to estimate propensities can be done by fitting a model on all the data, or by fitting several models on parts of the data (center panel). Our approach used ‘out of fold’ estimations. Finally, selection of model complexity is done by cross-validation to minimize prediction error or to minimize model complexity while achieving an error that is statistically comparable to the minimum based on the 1-standard error rule (right panel).



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Organ Type** | |  | **1** | **2** | **3** | **4** |
|  |  |  |  |  |  |  |
| **Liver** | Standard | Bilirubin, mg/dL | 1.2-2.0 | 2.0-6.0 | 6.0-12.0 | >12.0 |
|  | Modified | AST or ALT (U/L) | 200-1000 | >1000 |  |  |
| **Nervous** | Standard | Glasgow Coma Score | 13-14 | 10-12 | 6-9 | <6 |
|  | Modified | Clinical notation (e.g., agitation)[2](#_ENREF_2) |  | Present |  |  |
| **Respiratory** | Standard | PaO2/FIO2 mm Hg | 300-400 | 200-300 | 100-200 | <100 |
|  | Modified | O2sat/FIO2 mm Hg | 315-400 | 235-315 | 135-235 | <135 |
| **Coagulation** | Standard | Platelets, x1000/µL | 100-150 | 50-100 | 20-50 | <20 |
|  | Modified | INR | 3-6 | >6 |  |  |
| **Renal** | Standard | Creatinine, mg/dL | 1.2-2.0 | 2.0-3.5 | 3.5-5.0 | >5.0 |
| **Cardiac** | Standard |  | MAP mm Hg <70 | Dopamine <5 or  any Dobutamine | Dopamine 5-15 or Epinephrine <0.1 or Norepinephrine <0.1 | Dopamine >15 or Epinephrine >0.1 or Norepinephrine >0.1 |
|  |  |  |  |  |  |  |

**Appendix Table 1.** Criteria for quantifying standard and modified SOFA subscores. Normal subscores were calculated as suggested in the literature, without including respiratory support or urine output measurement. Modified subscore criteria are displayed in the lower row of each organ type.

1. Escobar GJ, Gardner MN, Greene JD, Draper D, Kipnis P. Risk-adjusting Hospital Mortality Using a Comprehensive Electronic Record in an Integrated Health Care Delivery System. *Med Care.* 2013;51(5):446-453.

**Appendix Table 2.** Number of sepsis patients with organ failure (black text in upper line) and their corresponding hospital mortality (red italic text in lower line) stratified by modified maximum SOFA subscore values during hospitalization. P-values for a test of trends between increasing SOFA subscores and hospital mortality were <0.01 for all organ dysfunction subtypes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Modified SOFA Organ System Score Value** | | | | |
| **Organ System** | **0** | **1** | **2** | **3** | **4** |
|  |  |  |  |  |  |
| **Liver** |  |  |  |  |  |
| Number | 25,191 (83.5) | 2,777 (9.2) | 1,776 (5.9) | 286 (1.0) | 133 (0.4) |
| *Hospital Mortality* | *2,097 (8.3)* | *342 (12.3)* | *291 (16.4)* | *67 (23.4)* | *50 (37.6)* |
|  |  |  |  |  |  |
| **Nervous** |  |  |  |  |  |
| Number | 15,418 (51.1) | 539 (1.8) | 11,618 (38.5) | 1,743 (5.8) | 845 (2.8) |
| *Hospital Mortality* | *348 (2.3)* | *32 (5.9)* | *1,392 (12.0)* | *564 (32.4)* | *511 (60.5)* |
|  |  |  |  |  |  |
| **Respiratory** |  |  |  |  |  |
| Number | 22,977 (76.2) | 1,489 (4.9) | 1,528 (5.1) | 1,379 (4.6) | 2,790 (9.3) |
| *Hospital Mortality* | *1,268 (5.5)* | *80 (5.4)* | *150 (9.8)* | *274 (19.9)* | *1,075 (38.5)* |
|  |  |  |  |  |  |
| **Coagulation** |  |  |  |  |  |
| Number | 18,617 (61.7) | 7,372 (24.4) | 3,068 (10.2) | 758 (2.5) | 348 (1.2) |
| *Hospital Mortality* | *1,503 (8.1)* | *581 (7.9)* | *464 (15.1)* | *187 (24.7)* | *112 (32.2)* |
|  |  |  |  |  |  |
| **Renal** |  |  |  |  |  |
| Number | 17,675 (58.6) | 7,380 (24.5) | 3,093 (10.3) | 882 (2.9) | 1,133 (3.8) |
| *Hospital Mortality* | *1,060 (6.0)* | *769 (10.4)* | *672 (21.7)* | *211 (23.9)* | *135 (11.9)* |
|  |  |  |  |  |  |
| **Cardiac** |  |  |  |  |  |
| Number | 11,336 (37.6) | 15,990 (53.0) | 214 (0.7) | 1,004 (3.3) | 1,619 (5.4) |
| *Hospital Mortality* | *488 (4.3)* | *1,461 (9.1)* | *42 (19.6)* | *137 (13.7)* | *719 (44.4)* |
|  |  |  |  |  |  |

**Appendix Table 3a.** Odds ratios quantifying the association between each acute organ dysfunction and hospital mortality. Models are adjusted for age, gender, predicted hospital mortality, acute severity of illness (based on the *Laboratory and Acute Physiology Score, version 2)*, comorbid disease burden (based on the *Comorbidity Point Score, version 2*), intensive care unit utilization, and full code status.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Odds Ratio** | **95% Confidence Interval** | **p-value** |
|  |  |  |  |
| Liver | 1.25 | 1.18 – 1.32 | <0.001 |
| Nervous | 1.86 | 1.77 – 1.95 | <0.001 |
| Respiratory | 1.43 | 1.39 – 1.47 | <0.001 |
| Coagulation | 1.24 | 1.19 – 1.30 | <0.001 |
| Renal | 1.14 | 1.10 – 1.19 | <0.001 |
| Cardiac | 1.31 | 1.26 – 1.26 | <0.001 |
|  |  |  |  |

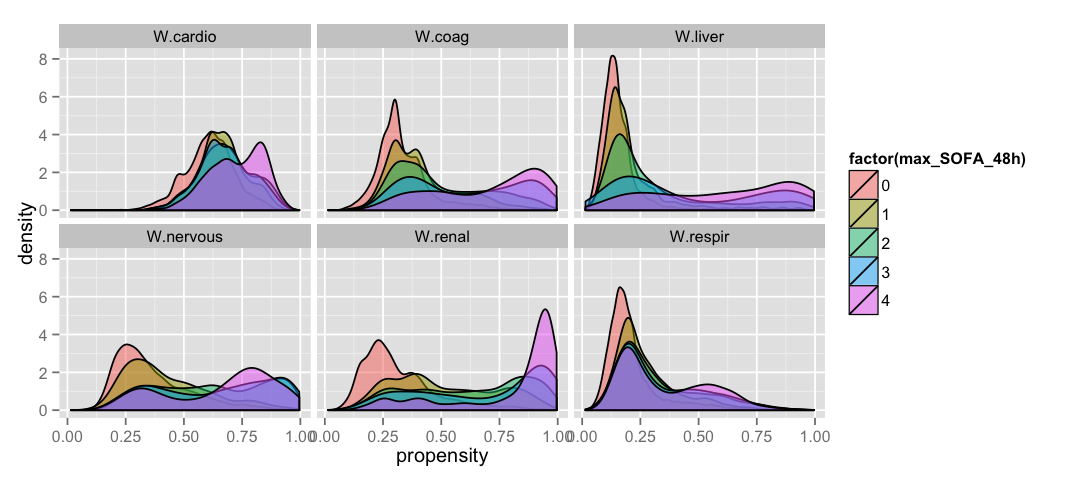
**Appendix Table 3b**. Odds ratios quantifying the association between each acute organ dysfunction and time to mortality beginning from the time of sepsis hospitalization admission (including both death during hospitalization and after hospitalization ).

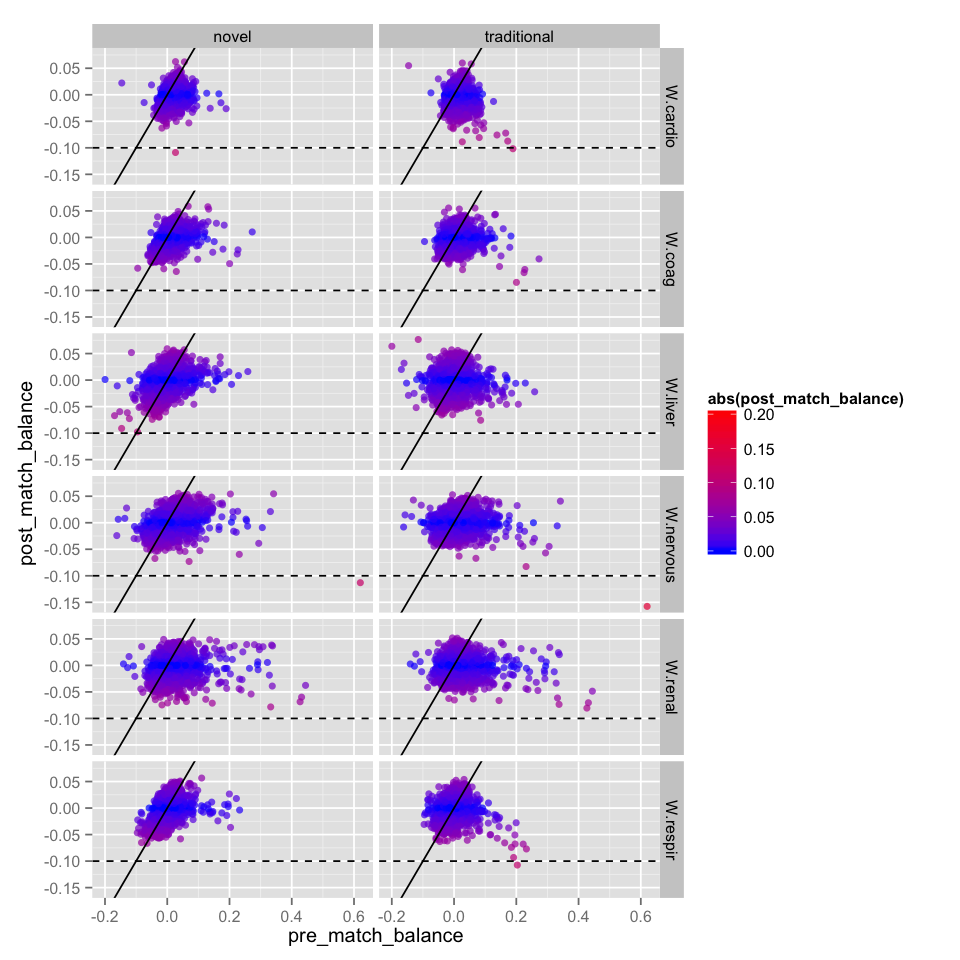
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Odds Ratio** | **95% Confidence Interval** | **p-value** |
|  |  |  |  |
| Liver | 1.14 | 1.09 – 1.19 | <0.001 |
| Nervous | 1.56 | 1.49 – 1.62 | <0.001 |
| Respiratory | 1.13 | 1.08 – 1.78 | <0.001 |
| Coagulation | 1.03 | 0.99 – 1.07 | 0.08 |
| Renal | 1.03 | 0.99 – 1.07 | 0.09 |
| Cardiac | 1.02 | 0.98 – 1.06 | 0.43 |
|  |  |  |  |

**Appendix Figure 2.** The ten most highly predictive features for each organ-specific propensity score model. Features are based on propensity scoring model that included 3,265 possible variables including: age, gender, comorbid disease score (COPS2), hospital admission type (‘ADMIT\_CAT’, i.e. medical versus surgical), and ICD9-based diagnosis and procedural codes. Variable importance scores are based on the number of times a variable is selected for splitting, taking into consideration the resulting improvement in the model from that split, and averaged over all trees.



**Appendix Figure 3.** Estimated propensity scores accurately stratify patients by their numerical SOFA organ subscores, despite being fit to predict a binary indicator of organ failure. Patients who have been estimated at higher risks of suffering a particular organ failure (SOFA subscore > 0) are also more likely to have had higher SOFA subscores for that organ.



**Appendix Figure 4.** Absolute standardized mean differences (ASMD) of each covariate before (x-axis) and after (y-axis) matching for each exposure with and without our novel propensity score modeling methods. Perfect covariate matching would result in a horizontal line of points (all covariates balanced after matching), while random matching would result in a line of points along the *y = x* diagonal (no improvement in balance after matching). Both traditional methods and our novel methods result in reasonable covariate balance (nearly all ASMDs < 0.1) in each matched cohort. The traditional propensity score estimates result in balance plots that skew more along the *y = -x* diagonal, indicating that strong imbalances are overcorrected by those models.

We found that propensity models fit by standard machine learning practice (without use of the out-of-fold estimation method or the one-SE selection rule) resulted in an overcorrection of the largest imbalances present in the full cohort. For instance, in the study of lung failure, the covariate indicating a previous diagnosis of chronic respiratory failure was more prevalent in those who experienced lung failure than those who did not (SMD > 0.2). After matching using a propensity score generated by a model fit with traditional machine learning practice, the effect flipped: chronic respiratory failure was more prevalent in the matched controls than in those who experienced lung failure (SMD < -0.1). In contrast, after matching using a propensity score generated by our methods, chronic respiratory failure was approximately equally prevalent in both groups (SMD ~ -0.04). Avoiding this overcorrection in highly imbalanced covariates came at the expense of attenuating smaller imbalances in other covariates.

**Appendix Table 4.** Adjusted hazard ratios based on multivariable Cox proportional hazards regression models (left) and logistic regression models for 1-year mortality (right) based on SOFA subscore points (top) or binary presence of organ dysfunction (bottom) using propensity score-matched cohorts. Model covariates include maximum hospitalization SOFA subscore values for each organ system, age, predicted mortality, severity of illness (LAPS2 score), chronic comorbid burden (COPS2), intensive care unit utilization, and full code status. Matched cohort size by organ dysfunction was: 7,205 (liver); 13,787 (nervous); 9,701 (respiratory); 14,985 (coagulopathy); 12,522 (renal); and 17,034 (cardiac).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Regression models based on per-point organ dysfunction increments** | | | | |
|  | **Cox Regression for Time-to-Death** | |  | **Logistic Regression for 1-year death** | |
| Organ dysfunction, per point | Hazard Ratio (95% CI) | p |  | Odds Ratio (95% CI) | p |
|  |  |  |  |  |  |
| Liver | 1.07 (1.01 – 1.14) | 0.025 |  | 1.19 (1.08 – 1.31) | 0.001 |
| Nervous | 1.18 (1.15 – 1.22) | <0.001 |  | 1.22 (1.17 – 1.28) | <0.001 |
| Respiratory | 0.96 (0.93 – 0.99) | 0.007 |  | 0.94 (0.90 – 0.99) | 0.011 |
| Coagulopathy | 1.11 (1.08 – 1.15) | <0.001 |  | 1.13 (1.08 – 1.19) | <0.001 |
| Renal | 1.01 (0.98 – 1.04) | 0.623 |  | 1.00 (0.95 – 1.04) | 0.859 |
| Cardiac | 0.94 (0.91 – 0.97) | <0.001 |  | 0.88 (0.83 – 0.92) | <0.001 |
|  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Regression models based on presence of absence of organ dysfunction** | | | | |
|  | **Cox Regression for Time-to-Death** | |  | **Logistic Regression for 1-year death** | |
| Organ dysfunction, binary | Hazard Ratio (95% CI) | p |  | Odds Ratio (95% CI) | p |
|  |  |  |  |  |  |
| Liver | 1.06 (0.95 – 1.18) | 0.286 |  | 1.21 (1.02 – 1.44) | 0.026 |
| Nervous | 1.44 (1.35 – 1.53) | <0.001 |  | 1.52 (1.38 – 1.68) | <0.001 |
| Respiratory | 0.95 (0.87 – 1.03) | 0.180 |  | 0.89 (0.78 – 1.02) | 0.094 |
| Coagulopathy | 1.04 (0.99 – 1.10) | 0.136 |  | 0.98 (0.89 – 1.07) | 0.610 |
| Renal | 0.98 (0.93 – 1.04) | 0.576 |  | 0.96 (0.87 – 1.06) | 0.381 |
| Cardiac | 0.98 (0.93 – 1.03) | 0.387 |  | 0.92 (0.84 – 0.99) | 0.043 |
|  |  |  |  |  |  |

**Appendix Figure 5.** Visual representation of various sensitivity analyses as described in Appendix Table including using organ dysfunction SOFA values as score points (ranging from 0 to 4, top panel) versus as a ‘binary’ indicator of organ failure (bottom panel).





**REFERENCES**

1. Hastie T TR, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer; 2009.

2. Escobar GJ, Gardner MN, Greene JD, Draper D, Kipnis P. Risk-adjusting Hospital Mortality Using a Comprehensive Electronic Record in an Integrated Health Care Delivery System. *Med Care.* 2013;51(5):446-453.