Methods and results

Abhijit Dasgupta

# Methods

We are interested in discovering if the death risk in individuals admitted with sepsis differs between SLE and non-SLE patients within hospitals. In other words, do SLE patients admitted with sepsis experience excess deaths compared to non-SLE patients, within hospitals.

We take a counterfactual approach to this problem. We propose to model the risk of death among the non-SLE patients conditional on various covariates including the admission hospital, age, comorbidities, admission on ventilator, gender, and type of insurance. We would then use this model to predict the risk (or probability) of death during hospitalization for each SLE patient using the observed covariates for that patient. This would give the risk of death for that patient under the assumption that they had the same conditional death risk as non-SLE patients. If we add up these predicted probabilities over patients in a hospital, we would get the *expected* number of deaths (E) among SLE patients under the assumption that they shared the same experience as non-SLE patients. We would compare this with the *observed* number of deaths (O) among SLE patients at that hospital. The O/E ratio is akin to a standardized mortality rate, where the standardized population would be non-SLE patients with sepsis in the sample.

In order to flexibly model the conditional risk of death, we use the XGBoost algorithm [[1]](#footnote-22). This modern form of gradient boosted trees allows for regularization as well as stochastic gradient descent methods to reduce the chance of overfitting, and utilizes decision trees as the learning engine to allow flexible prediction of the outome from heterogeneous responses. For training the XGBoost model,we use the R package xgboost [[2]](#footnote-23), running on R[[3]](#footnote-25) version 3.4. We train this model on the non-SLE patients, using 5-fold cross-validation to train the model parameters (using the R package caret[[4]](#footnote-27)). Our optimal trained model achieved a cross-validated average AUC of 0.86. We then use this trained model to predict the risk of death among SLE patients, by scoring each patient using the model to get a predicted probability of death. Within hospital, we then add these predicted probabilities to obtain the expected number of deaths under the assumption that the risk for SLE and non-SLE patients are the same. Finally, we compare the observed number of deaths per hospital with the expected number of deaths under the model. Note that the expected conditional probabilities of death use information across all hospitals and patients through the XGBoost model. Also note that, despite the hospitals being sampled from the national population of hospitals under the NIS, our analysis is conditional on the hospitals and patients in our dataset and so we do not use the sampling weights provided by NIS in our modeling.

All analyses were performed using R version 3.4 along with the packages xgboost, caret and survey[[5]](#footnote-29).

# Results

Table 1 gives the sampling weight adjusted summary statistics for various measurements utilized in our analysis, stratified by SLE status and vital status. We see that the SLE patients were over 85% female, whereas roughly half of the non-SLE patients were female. We also see that more than two-thirds of those who died were put on a ventilator on admission, compared to less than one-sixth among those who lived. There were no real differences in insurance coverage across the groups. We note that for a hospital to be eligible for this study, it had to have admitted *at least 5 SLE patients with sepsis* over the scope of the survey data (2003-2012 and hospital selected in NIS sample).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | non-sle alive | sle alive | non-sle dead | sle dead |
| Age | 49.23 | 44.60 | 52.71 | 47.92 |
| Female | 49.43 % | 88.82 % | 46.42 % | 85.62 % |
| Elix score | 8.18 | 7.42 | 13.69 | 11.64 |
| On ventilator | 14.07 % | 10.93 % | 66.35 % | 70.15 % |
| On medicare | 29.42 % | 38.93 % | 30.08 % | 42.7 % |
| On medicaid | 21.71 % | 22.53 % | 24.16 % | 19.39 % |
| On private insurance | 36.14 % | 31.73 % | 32.56 % | 27.67 % |
| On other insurance | 12.7 % | 6.78 % | 13.12 % | 10.24 % |

We next look at the distribution of the observed/expected ratio of deaths by hospital derived using our machine learning scheme. We find that 163 hospitals did not see any sepsis-related deaths among SLE patients, and so have an O/E ratio of 0. Among the rest of the 423 hospitals, we see that the O/E ratio could be as high as 10.48, with the median value being 1.47, and the 3rd quartile being 2.04.

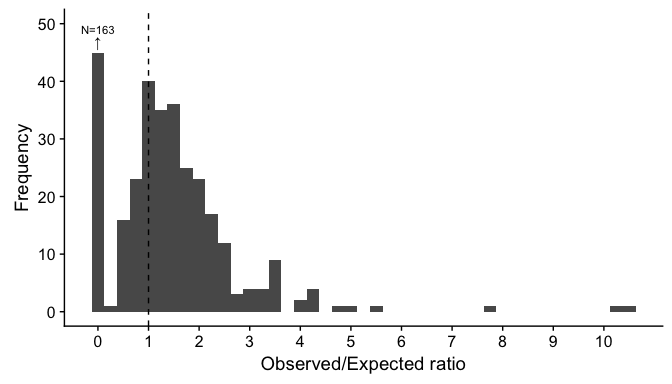


Figure 1: Distribution of observed/expected ratio for deaths from sepsis across 423 hospitals selected from NIS who admitted at least 5 SLE patients with sepsis within the scope of this study

The O/E ratio gives the conditional relative risk of death from sepsis by SLE status. However, it is entirely feasible that one could have a hospital with a high death rate from sepsis, but a low relative risk from SLE, and vice versa. Figure 2 investigate the comparative effects of the relative risk of death from SLE and the absolute risk of death from sepsis by hospital. Here "A" refers to the absolute risk of death by hospital (i.e. the probability of death from sepsis that is observed), and "R" refers to the relative risk of death from sepsis due to SLE status. The cutoff for denoting low and high relative risks was set at 2, i.e., more than a 2-fold increased risk of death due to sepsis was considered high. The cutoff for the absolute death rate from sepsis being called high is the median value of 12.95%. We see that even though the vast majority of hospitals in this study have a reasonable death rate from sepsis, there are some hospitals whith very high death rates where there is no effect of SLE, some hospitals with high absolute death rates where the rates for SLE patients are even higher, and some hospitals where the absolute death rate is low, but the SLE patients have a much higher (4-6, upto 10 times higher) death rate.

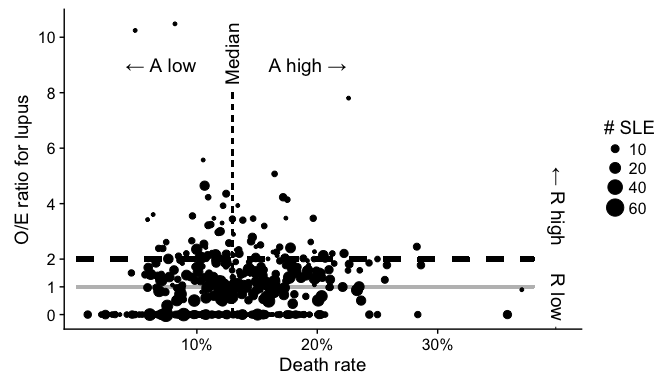


Figure 2: Relation between the hospital death rate for lupus and the relative risk of death by SLE status by hospital. Circle sizes are proportional to the number of SLE patients with sepsis seen at the hospital over the study period.

The classification of hospitals into R low/high and A low/high is then used to see what kinds of hospital-level characteristics might be predictive of this joint status. We look at this by fitting a multiclass decision tree using the CART methodology[[6]](#footnote-33), as implemented in the R package rpart[[7]](#footnote-34). Figure 3 shows the fit decision tree, where the table each node gives the relative frequency of each of the 4 categories of hospitals at that node. We find that SLE experience, i.e. the number of SLE patients seen at the hospital, and region segregate the classes of hospitals best. Hospitals with low SLE experience and low absolute death rates have a 3-fold higher chance of being classified (R high, A low) than hospitals with high SLE volume. Among high SLE experience hospitals, Midwest hospitals are less likely to have a high relative risk due to SLE than the other regions.

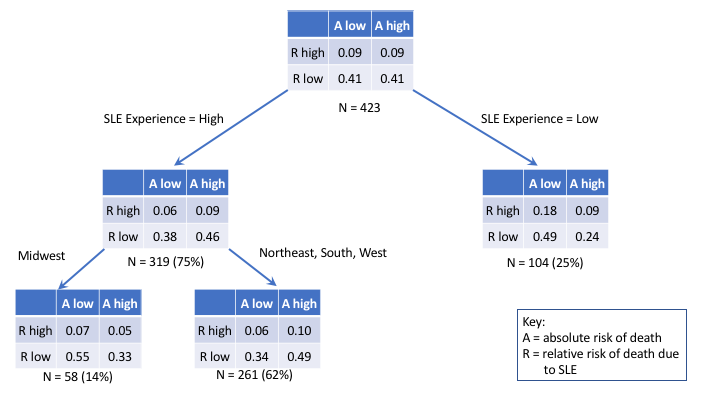


Figure 3: Decision tree of hospital status based on absolute and relative risks of death, with hospital characteristics like SLE experience, location, bed size and whether it was a rural or urban hopsital being predictors. The tables are in the same orientation as Figure 2

1. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM. [↑](#footnote-ref-22)
2. Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich and Yuan Tang (2017). xgboost: Extreme Gradient Boosting. R package version 0.6-4. <https://CRAN.R-project.org/package=xgboost> [↑](#footnote-ref-23)
3. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>. [↑](#footnote-ref-25)
4. Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2017). caret: Classification and Regression Training. R package version 6.0-77. <https://CRAN.R-project.org/package=caret> [↑](#footnote-ref-27)
5. T. Lumley (2017) "survey: analysis of complex survey samples". R package version 3.32. [↑](#footnote-ref-29)
6. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press. [↑](#footnote-ref-33)
7. Terry Therneau, Beth Atkinson and Brian Ripley (2017). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-11. <https://CRAN.R-project.org/package=rpart> [↑](#footnote-ref-34)