Accurate and interpretable intensive care risk adjustment for fused clinical data with generalized additive models

Ben J. Marafino, BS[1], R. Adams Dudley, MD, MBA[2], Nigam H. Shah MBBS, PhD[3], Jonathan H. Chen, MD, PhD[3]

[1]Biomedical Informatics Training Program, Stanford University, Stanford, CA; [2]Center for Healthcare Value, Philip R. Lee Institute for Health Policy Studies, and Department of Pulmonary and Critical Care Medicine, University of California, San Francisco, San Francisco, CA; [3]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA

Abstract

Risk adjustment models for intensive care outcomes have yet to realize the full potential of data unlocked by the increasing adoption of EHRs. In particular, they fail to fully leverage the information present in longitudinal, structured clinical data – including laboratory test results and vital signs – nor can they infer patient state from unstructured clinical narratives without lengthy manual abstraction. A fully electronic ICU risk model fusing these two types of data sources may yield improved accuracy and more personalized risk estimates, and in obviating manual abstraction, could also be used for real-time decision-making. As a first step towards fully "electronic" ICU models based on fused data, we present results of generalized additive modeling applied to a sample of over 36,000 ICU patients. Our approach outperforms those based on the SAPS and OASIS systems (AUC: 0.908 vs. 0.794 and 0.874), and appears to yield more granular and easily visualized risk estimates.

1. Introduction

Intensive care in the United States consumes nearly 1% of the country's GDP annually and accounts for 13% of all hospital costs, as of 2010.[1] The human toll of ICU care in the U.S. also presents a large burden: nearly 6 million patients are admitted to the ICU annually,[2] and mortality rates in ICUs are estimated to be about 12%, on average.[3] In addition, large variations in the quality of ICU care are also observed between hospitals, and much of this variation remains even after adjusting for disparities in case mix.[4–6] Taken together, these findings underscore the critical need for performance measurement and benchmarking of ICU outcomes, chiefly of mortality and length of stay. Such initiatives are driven by the application of accurate ICU risk adjustment models, including the Acute Physiology and Chronic Health Evaluation (APACHE),[7] the Simplified Acute Physiology Score (SAPS),[8] and the Mortality Probability Model (MPM).[9] In addition, these models have also been used to adjust for case mix in observational studies,[3] to compare treatment arms in RCTs,[10] and to inform ICU decision-making, triage, and resource allocation.[11]

However, existing ICU models exhibit a series of limitations. As the increasing adoption of EHRs in the United States and elsewhere promises to make a wider variety of data available for these models to ingest, recent work in this area has focused on merely adapting existing models, including, for example, SAPS[12] and APACHE[13] to the EHR. However, in doing so, such approaches centered around adaptation neglect the wealth of information available in the form of unstructured, free-text clinical narratives, as well as trends in longitudinal laboratory test results and vital signs, possibly limiting predictive accuracy. If not EHR-adapted, these models then require manual chart abstraction by trained staff; one study estimated this abstraction process to take 37 and 20 min per chart, on average, for APACHE and SAPS, respectively.[6] The time and cost burden associated with chart abstraction limits these models' utility for real-time ICU decision-making, and makes it a nontrivial task to compute these risk scores retrospectively; in one study, the cost to abstract the data required to compute APACHE scores for a 60,000-patient ICU cohort was estimated to approach $2 million.[12]

In this work, we present steps towards "fully electronic" ICU risk adjustment based on generalized additive models (GAMs) utilizing features built from *fused* clinical data to predict in-hospital mortality. By fused clinical data, we denote data which combines both structured and unstructured data sources; here, the former are derived from longitudinal laboratory test results and vital sign measurements from ICU flowsheets, while the latter derive from the free text of clinical narratives. Fused data of this form have previously been used to predict code status[14] and colorectal surgical complications,[15] but have yet to be used to predict ICU mortality or to estimate risk in the clinical setting more generally, aside from one

instance where topics derived from latent Dirichlet allocation were used in conjunction with SAPS for this task.[16] Indeed, the use of fused data in clinical predictive modeling more generally appears to be rare, with nearly all published models leveraging either only structured or unstructured data sources, but not both in a single model.[17] Such an approach, especially when combined with a flexible and interpretable method such as GAMs, could yield a richer model giving more personalized risk estimates obtained by the interaction of features derived from both types of sources. For example, trends in a patient's Glasgow Coma Score (GCS) during their first 24 hours in the ICU could be interacted with mentions of postoperative status in provider notes to personalize their risk estimate to an extent not possible if relying on changes in GCS alone.

2. Methods

The data used in this study were drawn from the Medical Information Mart for Intensive Care-III (MIMIC-III) ICU database, developed from EHR, telemetry, and other data routinely collected for patients admitted between 2001 and 2012 at Beth Israel Deaconess Medical Center in Boston, Massachusetts, a tertiary care and teaching hospital.[18] For this study, we identified the first ICU stay lasting > 4 h for each patient and selected all data that were collected up to 24 hours following ICU admission representing the laboratory tests and vital sign measurements listed in Table 1. This is in line with other ICU models, including APACHE IV and SAPS II, which look up to 24 hours following ICU admission; however, these models use only the 'worst' value, i.e., either the highest or lowest value within this window, depending on the variable. We also selected all provider notes, including physician progress notes, nursing notes, postoperative notes, and radiology reports written within this window, and did not differentiate on the basis of the author. The outcome used was in-hospital mortality.

To model the relationships between these features and mortality, we relied on an approach using generalized additive models (GAMs).[19–21] GAMs represent an extension of generalized linear models (GLMs), where the predictors are related to the outcome via smooth, and possibly nonlinear, functions. The parametric forms of these functions can be prespecified before fitting the GAM and then the parameters estimated from the data, or otherwise be taken to be nonparametric and having arbitrary shape, with the latter approach being more common. Classes of functions commonly used in GAMs include locally weighted regression (LOESS) smoothers, smoothing splines, and regression splines. Here, we use smoothing splines, which afford greater flexibility than regression splines or LOESS smoothers, but tend to be more computationally expensive.

The GAM was compared to a series of models run against the fused dataset as described above, specifically L_2-regularized (or ridge) logistic regression, linear support vector machines (SVMs), and gradient boosted trees implemented using xgboost.[22] We also obtained performance estimates for a logistic regression models based on both the SAPS and Oxford Acute Severity of Illness Score (OASIS)[23] systems, which can easily be computed from data elements in MIMIC-3. We used the area under the receiver operating characteristic curve (AUC) was used a metric to compare models. Estimates of and 95% confidence intervals for the AUC for each model were obtained by 10 repetitions of 10-fold cross-validation (CV). All models and related procedures were implemented in R (version 3.3.3) using the caret and gam packages.

The physiologic data, comprising laboratory tests and vital sign measurements, were collected sequentially at varying sampling rates, and so each test or vital sign were initially represented by a time series. For example, heart rate was recorded hourly, while laboratory tests were taken less regularly. For each time series, in order to capture gross temporal variation, we engineered a set of derived features based on summary statistics, namely the mean value, standard deviation, maximum and minimum, last minus first value recorded as well as the absolute value of this difference,, and the slope of the linear trend fit to the data. The full list of data elements and derived features used in our experiments is given in Table 1. We fit a GAM model using only these features derived from structured data as a baseline for the fused-data model.

For all fused-data models, prior to each fold of CV, and owing to computational considerations, we also separately fit a LASSO-based classifier, as in [24], to the free text narratives in the training set in order to prune the set of unstructured term mentions that eventually served as input to GAM, together with the features derived from structured data elements. On the order of 10^5 unique terms were present across all the notes in our corpus, and this step resulted in roughly 500 unique terms being selected over each CV fold.

Without such a step, our GAM would have had to scale to accommodate on the order of $O([10^5+100]^2)$ possible interactions, a model fit which would have proved infeasible to compute. Throughout this process, each document comprising all the notes associated with a patient was represented as a bag of words without normalization, meaning that ontology mapping, negation detection, or other similar techniques were not used. Finally, the per-document frequencies associated with each term were then transformed into term frequency-inverse document frequency (tf-idf)[25] features which served as input to the GAM. We also performed a sensitivity analysis using the sublinear term frequency, i.e., log (1+tf), in place of the raw term frequency in order to adjust for note length, which we hypothesized could be associated with mortality.

Table 1. List of structured data sources and the types of *derived features* engineered from each source. All derived features are with respect to a window of maximum length 24 hours following ICU admission.

Laboratory tests	**Vital signs**	**Derived feature types**
Blood urea nitrogen	Heart rate	Mean
Bilirubin	Respiratory rate	Standard deviation
Creatinine	Temperature	Maximum
Lactate	Mean arterial pressure	Minimum
Glucose	S_aO_2 (oxygen saturation)	Last value minus first value (Δ*featurename*)
Sodium	FiO_2	Absolute value of difference between last and first values
Potassium	Glasgow Coma Score (GCS) – total	Slope of linear trend fit to data using least squares
Bicarbonate	GCS – eye response	
Hematocrit	GCS – motor response	
White blood cell count	GCS – verbal response	
Platelet count		
Arterial P_aCO_2		

As the data used in this study were de-identified, this study was deemed to be exempt from review by the Institutional Review Board of the Stanford University School of Medicine.

3. Results

The characteristics of the dataset are presented in **Table 2**. The data reflect a rich case mix, with at least six different types of ICUs represented, including coronary care and cardiac surgery recovery units, medical ICUs, surgical ICUs, and a combined trauma/surgical ICU.

Table 2. Characteristics of the dataset.

Patients, total number	36,043
Deaths (%)	3,895 (10.8%)
Age, mean (IQR)	61.9 (51-76)
Of which male (%)	20,836 (57.8%)
Type of ICU	
Coronary care	5,255 (14.6%)
Cardiac surgery recovery unit	7,394 (20.5%)
Medical (including Neuro ICU)	12,549 (34.8%)
Surgical	5,963 (16.5%)
Trauma/Surgical	4,882 (13.6%)

The results of various models are presented in **Table 3**. The GAM outperformed a logistic regression model based on SAPS, with 10-fold cross-validated estimates of AUCs of 0.908 for the GAM versus 0.794 for logistic regression on SAPS, and 0.874 for logistic regression on OASIS. In addition, the performance of the GAM compares well to those of other models tested on the fused dataset; while an AUC of 0.910 was estimated for the gradient boosting machine (GBM) – the best-performing model in terms of raw AUC --

this difference was not statistically significant at the 0.05 confidence level, as the 95% CI for the estimate of the AUC obtained for the GBM failed to exclude 0.908. Furthermore, in a sensitivity analysis, using the a normalized tfidf metric incorporating a sublinear term frequency did not improve performance compared to the raw tfidf metric for each unstructured input, and so all performance estimates presented are based on the use of the latter metric. Finally, a GAM fit on fused data appears to yield statistically significantly better performance compared to a GAM fit on features derived from structured data alone.

Table 3. Model performance comparison.

Model	AUC (95% CI)
Logistic regression on SAPS score only	0.794 (0.790-0.798)
Logistic regression on OASIS score only	0.874 (0.864-0.881)
Logistic regression on fused dataset (FUSED)	0.857 (0.841-0.872)
Gradient boosting machine on FUSED	0.910 (0.901-0.920)
Support vector machine on FUSED	0.873 (0.855-0.893)
Generalized additive model on structured features	0.853 (0.840-0.866)
Generalized additive model on FUSED	**0.908 (0.898-0.917)**

Examples of univariate risk curves estimated by the GAM for single features are presented in **Figure 1**. Bivariate risk surfaces estimated for pairs of features, where both derive from unstructured data are presented in **Figure 2**, while those estimated for pairs where one feature derives from structured data and the other from unstructured data are presented in **Figure 3**. The univariate risk surfaces in **Figure 1** appear to more accurately recapitulate the nonlinear nature of the relationships between these features and mortality. While the GAM does not provide summary estimates of the relative influence of each feature as a GLM would by estimating a single coefficient, some of the most influential features, as given by the *p*-value for the significance of the smooth term, were the features derived from GCS scores, particularly the mean GCS over 24 hours, the difference in GCS (ΔGCS) over this window, and the slope of the linear trend fit to these scores. Other influential features included patient age, and their mean sodium, potassium, and lactate levels, the smooths of which are shown in **Figure 1**. Some of the most influential unstructured features, again selected on the basis of their *p*-value, included "expired", "CMO", "midline shift", "posturing", and "BMT" (bone marrow transplant).

The bivariate plots presented in Figures 2 and 3 exhibit some interesting properties. First, these plots can be used to assess linear dependence between the features, as can be seen in the plots in **Figure 2**, as well as in **Figure 3A**. Second, a "phase transition" can be observed in **Figure 3B**, where the relationship between the windowed mean of the mean arterial pressure (MAP) and a term mention related to vasopressor medication, here "pressors", begins to exhibit linear dependence for mean MAPs below 65 mm Hg. Similar patterns were observed for other related term mentions, including "levophed", and "dopamine", among others.

Third, in **Figure 3C**, the contour lines appear to recapitulate the shape of the univariate curve estimated for ΔGCS in **Figure 1C**, but rotated clockwise by 90 degrees, and we also observe that the risk estimates for ΔGCS are further stratified on the basis of postoperative status, as measured by mentions of "POD" (post-op day) in notes. Similarly, patients can be stratified both on the basis of their mean GCS during their first ICU day and also of their having experienced an overdose (**Figure 3D**), and the relationship between the two features also appears to exhibit linear dependence, in that post overdose status appears to confer a protective effect and vice versa. The highest-risk region lies in the bottom left of the plot, corresponding to other ICU patients with low mean GCSs who were *not* post overdose status. We cover some of these points in more depth in the Discussion section.

Figure 1. Bivariate risk curves for four features derived from physiologic measurements common in existing ICU risk models: serum sodium, potassium, and lactate, as well as the Glasgow Coma Score (GCS). However, for sodium and potassium levels, we have taken the 24-hour windowed mean, and for lactate and GCS, we have taken the windowed 24-hour difference. The grey regions denote 95% confidence intervals estimated by the GAM, and a rugplot lies at the bottom of each figure, giving the distribution of the feature among all patients in the dataset.

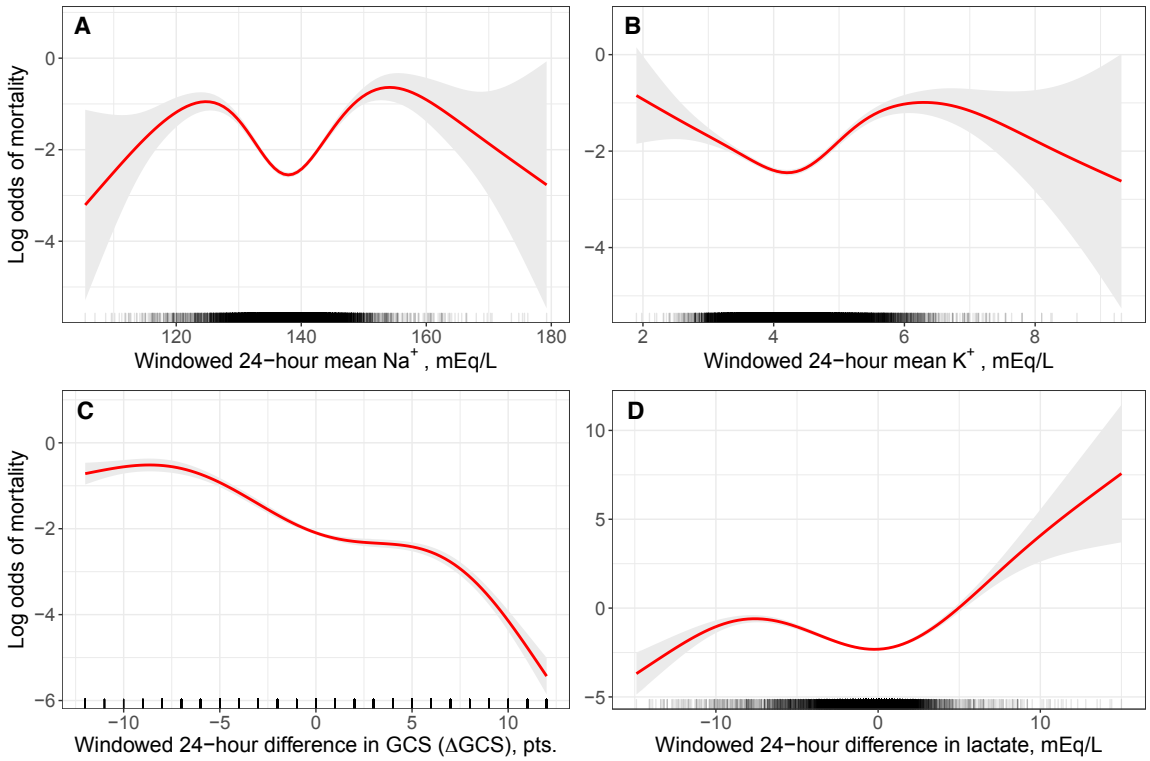


Figure 2. Examples of bivariate risk surfaces estimated for pairs of features derived from unstructured data. Risk estimates are represented by a red (low risk) to white (high) spectrum; the green lines denote contours joining areas of the plot having equal risk.

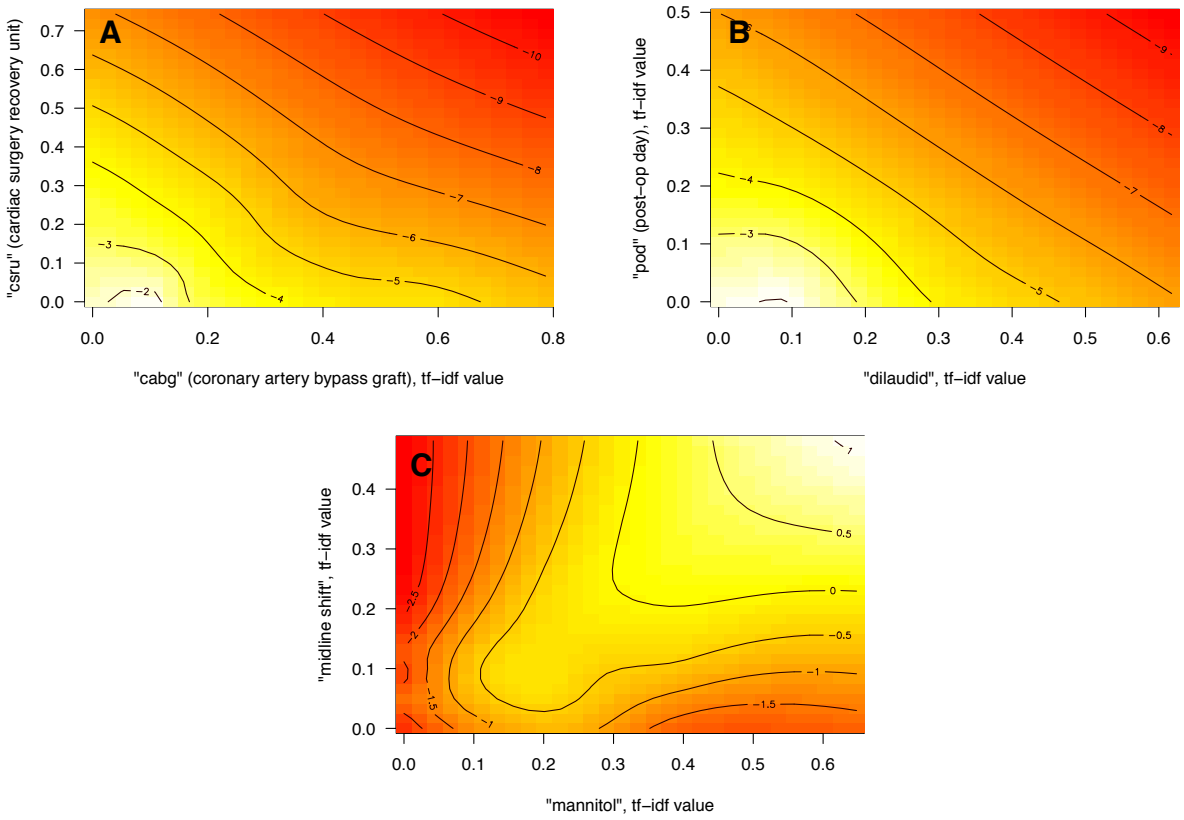
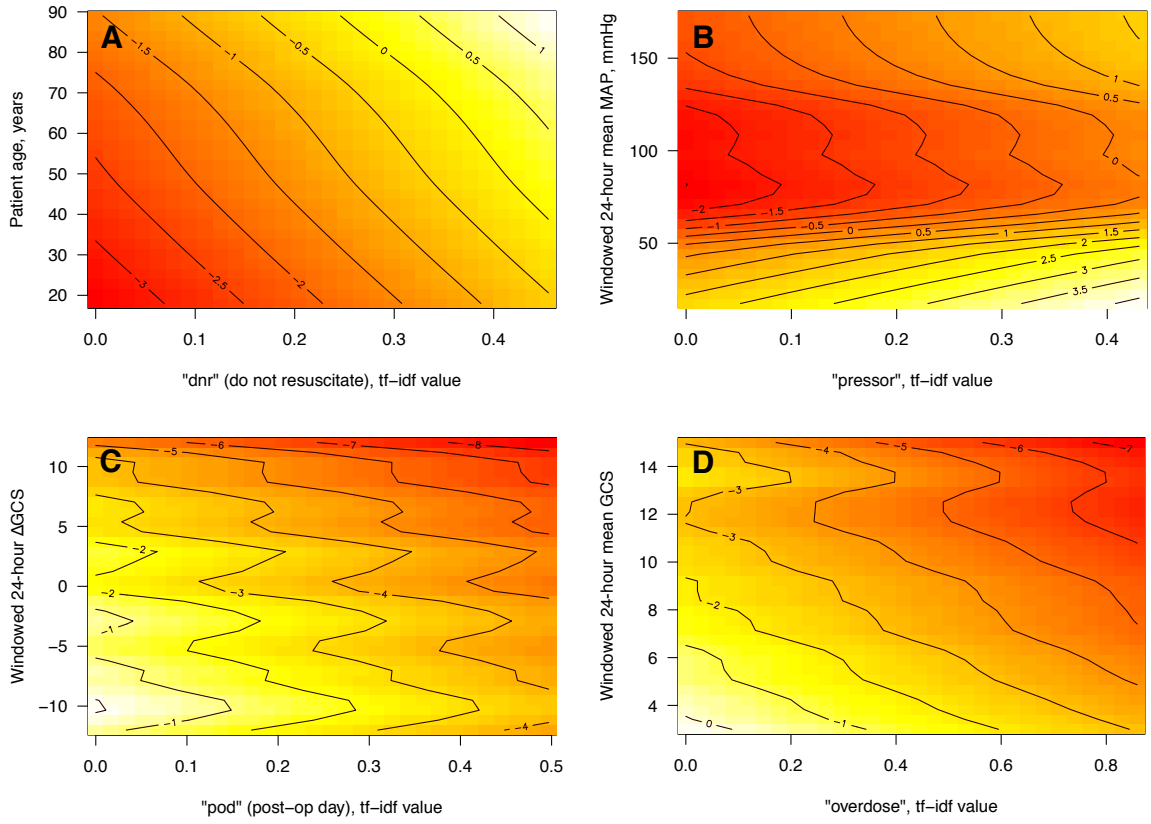


Figure 3. Bivariate risk surfaces for pairs of features, where one derives from unstructured data (x-axis) and the other from structured data sources (y-axis).



4. Discussion

Here, we contribute a risk adjustment model for ICU outcomes based on generalized additive modeling applied to fused clinical data, which comprises features derived from both structured and unstructured data sources. Our work highlights the utility of GAMs for predicting ICU mortality: not only did GAMs appear to significantly outperform existing ICU risk modeling methodologies – SAPS and OASIS – they also appeared to perform just as well as the current "gold standard" of predictive modeling – gradient boosting machines (GBMs) – applied to fused datasets. GAMs also appear capable of robustly estimating complex risk surfaces in order to produce more personalized risk estimates for ICU patients, and to a granularity not possible with other classifiers, including GBMs. In particular, our work demonstrates the utility of features derived from unstructured text, especially in conjunction with more traditional structured features via an interaction. However, further work remains to be done in order to be able to fully integrate features derived from text into ICU models, specifically with regard to normalization, which could possibly include mapping the terms onto an ontology and applying negation detection, among other methods, and also with regard to minimizing possible biases introduced by some of these terms, including, e.g., by "expired," and to a lesser extent, by "DNR" (do not resuscitate) and "CMO" (comfort measures only). Nevertheless, it appears clear that in this setting that these features provide additional value beyond that yielded by features derived from structured data elements, which should motivate their use in such models.

Overall, the GAM approach appears to provide more easily interpretable estimates of risk for both the univariate and bivariate cases. The estimated risk curves agree more readily with clinical intuition, compared to those that would be obtained via other models: a GLM, for example, would estimate a straight line through each plot in **Figure 1**, which would fail to capture the true shape of the nonlinear relationship

between these features and mortality. In addition, these risk curves superficially resemble those used by the Rothman Index (RI),[26] but are distinct in that the RI curves rely on polynomial regression and were estimated separately for each feature before fitting the model, while our approach uses more flexible smoothing splines and estimates all curves and surfaces together in a single model, potentially facilitating scalability and portability. Furthermore, and more generally, the RI also relies on the use of a customized data collection instrument separate from the main patient record, while our approach leverages only those data routinely collected during the process of care.

Examining the estimated risk curves for mean serum sodium and serum potassium over the up to 24 hour window following to ICU admission yielded by the GAM, we observe that they are U-shaped, with the minimum risk lying within common reference ranges for these tests, as would be expected. Similarly, the estimated risk curve for ΔGCS, or the difference between last and first Glasgow Coma Scores (GCSs) within this up to 24 hour window, again agrees with clinical intuition – that steeper declines in a patient's level of consciousness are associated with a poorer prognosis. However, due to the lack of extreme-valued data, the risk estimates in these cases may not be reliable, and this is reflected by the width of the confidence intervals estimated in these ranges. For example, the relationship between extremely high or low sodium values and mortality, as suggested by the univariate plot in **Figure 1A**, would almost certainly not be borne out in reality; more data would be required to be able to obtain more accurate estimates of the true relationship.

Of particular interest are the bivariate risk surfaces for pairs of features estimated by the GAM in **Figure 2**. In order to gain some intuition, we first examine the risk surfaces estimated for pairs of terms taken from unstructured clinical narratives that we would almost certainly expect *a priori* to occur together in notes, and hence to exhibit interaction in a GAM. Here, these pairs are 'CABG' (coronary artery bypass graft) and 'CSRU' (cardiac surgery recovery unit), as well as 'POD' (post-operative day) and 'dilaudid' (an opioid analgesic commonly given after surgery). The plots of each risk surface are given in **Figure 2**. Again, whiter colors represent higher risk, while redder colors correspond to lower risk, and the green contour lines join points having equal risk.

The gradients of the contours in each plot in **Figure 2** are diagonal, indicating an linear interaction between the terms in each pair, as would be expected. A diagonal gradient implies such an interaction as it shows that the risk contributions of an increase of a certain magnitude in the value of either feature are both approximately equal. In addition, the estimated risk decreases smoothly from the lower left to the upper right of each plot, which agrees with existing knowledge that patients admitted to ICUs postoperatively generally are at lower risk of in-hospital mortality.[7] The significance of these interactions lies in that they can characterize the extent to which one feature acts as a proxy for the other in such models, allowing sources of redundancy as well as novelty to be identified – which may potentially prove significant when working with features derived from heterogeneous data sources.

Furthermore, in **Figure 2**, the bivariate risk surface estimated for "mannitol" (a medication used to reduce cerebral edema) and "midline shift" also implies a linear interaction between these features, though not over the whole range of each, as can be observed in the other two plots in **Figure 2**. Here, the risk increases superlinearly along the diagonal region of the plot, and suggests a synergistic relationship between the two features, as can be observed from the steepness of the risk surface in this region, as implied by the increased widths between the contours. This observation suggests that patients experiencing cerebral edema to an extent requiring osmotic therapies, such as mannitol, are indeed at much higher risk than would be suggested by the presence in a provider note of either "midline shift" or "mannitol" alone (mannitol has other uses, e.g., as a laxative).

In particular, bivariate risk surfaces can be estimated for pairs of features where one derives from structured data and the other from unstructured data, as in **Figure 3**. The risk surface estimated for "DNR" (do not resuscitate) and patient age is presented in **Figure 3A**. It is known that increasing patient age is associated with use of DNR orders, and this is borne out here, as the diagonal gradients again suggest the presence of an interaction. In addition, risk also increases as both the number of DNR mentions and age increase. Moreover, **Figure 3B** also illustrates an interesting property captured by these GAMs: below a windowed mean value of mean arterial pressure (MAP) of 65 mm Hg, a "phase transition" is observed. At this point,

the risk surface appears to abruptly shift from a regime where little to no interaction exists between these features, to one which exhibits robust linear interaction: the lower the value of mean MAP, the more frequently mentions of "pressor" appear, and the mortality risk increases concomitantly. A similar pattern with the same "phase transition" is also observed for other terms relating to vasopressors, including "levophed" and "dopamine", among others. The existence of such a "phase transition" within these estimated risk surfaces is significant in that it delineates specific regions of the feature space where one feature can act as a proxy for the other and further informs the use of these features in risk models.

There are several limitations associated with our study. First, even though the data reflect a rich case mix and were collected over a 11-year-long period, they are representative of only one institution, and the relative portability of ICU risk models based on GAMs, compared to other classifiers, to other sites is unknown. Indeed, owing to their flexibility, and hence propensity to overfit, it is certainly possible that GAMs may not prove as portable as other classifiers,. Second, we did not apply normalization in the preprocessing pipeline, and so did not perform negation detection or mapping terms onto an ontology, nor were the computed tfidf metrics normalized to account for note length, but our sensitivity analysis found that the latter modification did not change performance significantly. However, we felt that the development and validation of a such a pipeline was outside the scope of this study, which aimed to principally establish the feasibility of fully electronic ICU risk models based on fused data, and to demonstrate the value of these unstructured features in such models. Finally, the only comparators representing existing ICU risk modeling methodology available to us were the SAPS-I and OASIS systems, which may offer somewhat limited performance compared to the state-of-art in scoring systems, including APACHE IV, that rely on a broader range of features and so capture more information about a patient's physiological state, but currently require costly manual chart abstraction. Ideally, a model based on the APACHE IV score would have been used as a comparator, but the estimated costs of data collection for this cohort – upwards of $1 million, based on other studies[6,12] – precluded us from doing so.

Moreover, the full extent of the utility of using features derived from clinical narratives to predict ICU and other outcomes is currently unknown. Several examples exist in the literature where features derived from nursing notes have been used to predict ICU mortality,[16,24] but to the best of our knowledge, no models currently deployed for this task rely on features derived from unstructured text data. While including these features in models could, as we have shown, potentially improve the predictive performance and facilitate interpretability of such models, their use also presents several unique challenges, given that documentation patterns may vary substantially between institutions. First, the mention of certain terms associated with documentation of patient death, including, e.g., "expired," proves sufficient to predict the outcome with certainty, and as such, may suppress the contribution of other predictors and introduce at least one source of bias, owing to variability in documentation patterns.

To a lesser degree, this phenomenon also occurs with other terms such as "DNR" (do not resuscitate) and "CMO" (comfort measures only); indeed, it is known that DNRs are associated with mortality.[27] Other similarly influential terms include "sepsis", and those associated with neurological exams indicative of comatose status, e.g., "corneals" (as in "corneals absent") and "posturing". While there may not exist substantial variation in documentation patterns with regard to patient death, several studies have demonstrated institutional and regional variation in DNR ordering patterns,[28,29] and such variation could present a potential source of bias -- one which could act in concert with variation in documentation patterns, although the full extent of the latter type of variation is presently unknown. Second, with the knowledge of the high influence of certain terms in these models, providers may then bias their documentation behaviors in such a way to "game" a model, e.g., by repeating terms associated with high risk while writing notes so as to inflate estimated mortality risks for their patients, thus improving the apparent risk-adjusted performance of their ICUs. However, these implementation challenges could be mitigated with new informatics methods built into the preprocessing pipelines that ingest the data for use by these models. Mitigation strategies could involve intelligently filtering potentially biased term mentions in unstructured narratives, as well as methods to characterize variation between providers' lexical styles in notes, or to clamp the contributions of such a subset of features in these models.

5. Conclusion

In this paper, we demonstrated the utility of an approach based on generalized additive models for ICU mortality prediction on fused clinical data which combines features derived from both structured and unstructured data sources. The GAM approach offers an unified framework which outperforms an existing modeling paradigm based on SAPS, and in fact yields performance comparable to gradient boosting machines, and which allows for the estimation of complex – yet easily interpreted risk surfaces for pairs of features. In particular, our approach demonstrates the value added by the inclusion of features derived from unstructured narratives to further stratify mortality risk, particularly in concert with features derived from structured data sources and engineered so as to capture temporal variation. This approach may hold value for models estimated and deployed in other settings beyond the ICU to produce more personalized risk estimates for patients. However, there remain implementation challenges in utilizing unstructured data sources to their fullest extent in such models, but these could be mitigated with the development of adjunctive informatics methods.

Acknowledgements

BJM is supported in part by the National Library of Medicine (NLM) training grant T15 LM007033; JHC is supported in part by NIH Big Data 2 Knowledge Award Number K01ES026837 via the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health or of Stanford Health Care.

References

1. Halpern NA, Pastores SM. Critical Care Medicine Beds, Use, Occupancy, and Costs in the United States: A Methodological Review. *Crit Care Med*. 2015;43(11):2452-2459. doi:10.1097/CCM.0000000000001227.
2. Society of Critical Care Medicine. Critical Care Statistics. http://www.sccm.org/Communications/Pages/CriticalCareStats.aspx. Accessed September 22, 2017.
3. Zimmerman JE, Kramer AA, Knaus WA. Changes in hospital mortality for United States intensive care unit admissions from 1988 to 2012. *Crit Care*. 2013;17(2):R81. doi:10.1186/cc12695.
4. Render ML, Kim HM, Deddens J, et al. Variation in outcomes in Veterans Affairs intensive care units with a computerized severity measure. *Crit Care Med*. 2005;33(5):930-939. http://www.ncbi.nlm.nih.gov/pubmed/15891316.
5. Rosenthal GE, Harper DL, Quinn LM, Cooper GS. Severity-adjusted mortality and length of stay in teaching and nonteaching hospitals. Results of a regional study. *JAMA*. 1997;278(6):485-490. http://www.ncbi.nlm.nih.gov/pubmed/9256223.
6. Kuzniewicz MW, Vasilevskis EE, Lane R, et al. Variation in ICU risk-adjusted mortality: Impact of methods of assessment and potential confounders. *Chest*. 2008;133(6):1319-1327. doi:10.1378/chest.07-3061.
7. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med*. 2006;34(5):1297-1310. doi:10.1097/01.CCM.0000215112.84523.F0.
8. Moreno RP, Metnitz PGH, Almeida E, et al. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med*. 2005;31(10):1345-1355. doi:10.1007/s00134-005-2763-5.
9. Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA. Assessing contemporary intensive care unit outcome: An updated Mortality Probability Admission Model (MPM0-III). *Crit Care Med*. 2007;35(3):827-835. doi:10.1097/01.CCM.0000257337.63529.9F.
10. The National Heart, Lung and BIARDS (ARDS) CTN. Efficacy and Safety of Corticosteroids for Persistent Acute Respiratory Distress Syndrome. *N Engl J Med*. 2006;354(16):1671-1684. doi:10.1056/NEJMoa051693.
11. Hyzy RC, Jacobs S, Lee B, al. et. ICU Scoring and Clinical Decision Making. *Chest*. 1995;107(6):1482-1483. doi:10.1378/chest.107.6.1482.

12. Liu V, Turk BJ, Ragins AI, Kipnis P, Escobar GJ. An Electronic Simplified Acute Physiology Score-based Risk Adjustment Score for Critical Illness in an Integrated Healthcare System. *Crit Care Med*. 2013;41(1):41-48. doi:10.1097/CCM.0B013E318267636E.
13. Chandra S, Kashyap R, Trillo-Alvarez CA, et al. Mapping physicians' admission diagnoses to structured concepts towards fully automatic calculation of acute physiology and chronic health evaluation score. *BMJ Open*. 2011;1(2):e000216. doi:10.1136/bmjopen-2011-000216.
14. Lojun SL, Sauper CJ, Medow M, Long WJ, Mark RG, Barzilay R. Investigating Resuscitation Code Assignment in the Intensive Care Unit using Structured and Unstructured Data. *AMIA Annu Symp Proc*. 2010;2010:467-471. http://www.ncbi.nlm.nih.gov/pubmed/21347022.
15. Soguero-Ruiz C, Hindberg K, Mora-Jiménez I, et al. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *J Biomed Inform*. 2016;61:87-96. doi:10.1016/j.jbi.2016.03.008.
16. Lehman L, Saeed M, Long W, Lee J, Mark R. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annu Symp proceedings*. 2012;2012:505-511. http://www.ncbi.nlm.nih.gov/pubmed/23304322. Accessed September 22, 2017.
17. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Informatics Assoc*. 2017;24(1):198-208. doi:10.1093/jamia/ocw042.
18. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. doi:10.1038/sdata.2016.35.
19. Hastie T, Tibshirani R. Generalized Additive Models. *Stat Sci*. 1986;1(3):297-310. doi:10.1214/ss/1177013604.
20. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman and Hall; 1990.
21. Wood SNW. *Generalized Additive Models: An Introduction with R*. Vol 62.; 2006. doi:10.1111/j.1541-0420.2006.00574.x.
22. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. March 2016. doi:10.1145/2939672.2939785.
23. Johnson AEW, Kramer AA, Clifford GD. A New Severity of Illness Scale Using a Subset of Acute Physiology and Chronic Health Evaluation Data Elements Shows Comparable Predictive Accuracy. *Crit Care Med*. 2013;41(7):1711-1718. doi:10.1097/CCM.0b013e31828a24fe.
24. Marafino BJ, John Boscardin W, Adams Dudley R. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J Biomed Inform*. 2015;54:114-120. doi:10.1016/j.jbi.2015.02.003.
25. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. *J Doc*. 2004;60(5):503-520. doi:10.1108/00220410410560582.
26. Rothman MJ, Rothman SI, Beals J. Development and validation of a continuous measure of patient condition using the Electronic Medical Record. *J Biomed Inform*. 2013;46(5):837-848. doi:10.1016/j.jbi.2013.06.011.
27. Fuchs L, Anstey M, Feng M, et al. Quantifying the Mortality Impact of Do-Not-Resuscitate Orders in the ICU. *Crit Care Med*. 2017;45(6):1019-1027. doi:10.1097/CCM.0000000000002312.
28. Zingmond DS, Wenger NS. Regional and Institutional Variation in the Initiation of Early Do-Not-Resuscitate Orders. *Arch Intern Med*. 2005;165(15):1705. doi:10.1001/archinte.165.15.1705.
29. Jayes RL, Zimmerman JE, Wagner DP, Knaus WA. Variations in the use of do-not-resuscitate orders in ICUs: Findings from a national study. *Chest*. 1996;110(5):1332-1339. doi:10.1378/chest.110.5.1332.