Personalized Mortality Prediction for the Critically Ill Using a Patient Similarity Metric and Bagging

Joon Lee[1]

Abstract— **Conventional mortality prediction in intensive care is based on severity of illness scores created from heterogeneous patient data that usually perform well at the population level but not necessarily at the patient level. With the emergence of large-scale electronic medical records, it is now feasible to utilize only those past patients that are clinically similar to a given index patient for whom mortality prediction is needed. Identification of similar patients can be achieved via a patient similarity metric (PSM) that quantifies the extent of similarity between two patients. Extending from a previous study, the present study aimed to investigate PSM-based bagging instead of hard-thresholding to mitigate the shortcomings identified in the previous study. Based on intensive care data from 17,152 patients, a cosine-similarity PSM and three predictive models were deployed. Besides bagging, the same methods as the previous study were used to enable a valid comparison. With a bootstrap size of 1000, the results showed that bagging led to a similar predictive performance for logistic regression (mean area under the receiver operating characteristic curve [95% confidence interval]: 0.815 [0.809, 0.821]) but worse performances for death counting (0.663 [0.655, 0.672]) and decision tree (0.510 [0.501, 0.520]), in comparison with hard-thresholding. Decreasing the bootstrap size to 500 had minimal effect on predictive performance. Future research should investigate other PSM types.**

I. INTRODUCTION

Mortality prediction in the intensive care unit (ICU) has been studied extensively for its potential importance in guiding treatment planning and allocation of scarce ICU resources. Mortality prediction is closely related to severity of illness (SOI) scores such as APACHE IV [1] and SAPS 3 [2] that are widely used in the ICU. Although these SOI scores can be used to estimate mortality risk for individual patients, they are more commonly used as a descriptive statistic for comparing the characteristics of different patient cohorts [3]. One of the main reasons why ICU SOI scores have seen limited utility in prognostication for individual patients is because they were originally developed based on heterogeneous, multi-center patient data in order to maximize predictive accuracy for average patients rather than unique individual patients. As a result, ICU SOI scores tend to achieve very good predictive performance at the population level but not necessarily at the patient level.

From machine learning's perspective, using large-scale, heterogeneous data to train a mortality prediction algorithm may not be ideal for making a prediction for a particular, index patient, because many of the patients in the training data are likely somewhat different from that index patient. This is especially true for ICU patients who usually exhibit many different permutations of admitting diagnosis, co-morbidities, medications, therapies, surgeries, and previous medical history. My colleagues and I have previously conducted studies to show that identifying and only utilizing similar patients in predictive modeling leads to improved mortality prediction performance [3], [4]. This paradigm of personalized mortality prediction is timely given the widespread adoption of electronic medical record (EMR) systems in North America as well as the resulting emergence of large-scale medical data. With big data, it is feasible to exclude irrelevant data from training without sacrificing sample size too much.

Personalized predictive analytics has been applied in non-health fields including product recommender systems based on collaborative filtering in e-commerce [5] and consumer credit scoring [6]. In health, Sun et al. [7] have developed a patient similarity measure that combines supervised machine learning and clinician input.

In [4], my colleagues and I used a patient similarity metric (PSM) based on cosine similarity to quantify the extent of similarity between two patients. Once all pairwise PSM values were computed, a hard threshold was applied to include only the N most similar patients in training, where N was varied. However, the major shortcomings with this hard-thresholding approach were: 1) N must be sufficiently large to ensure that all categories of categorical predictor variables and all possible outcomes are present in the training data (when selected patients are too similar, variety is lost); 2) optimizing N becomes challenging if predictive performance does not clearly peak at a particular N value or if the peak occurs at a small N value that cannot be evaluated due to the previous shortcoming described above; and 3) patients with PSM values below the threshold had zero probability of being included in the training data, which can be sub-optimal since the PSM itself may not be perfect.

In order to address the shortcomings with hard-thresholding outlined above, the objective of the present study was to extend our previous study [4] by replacing hard-thresholding with a softer, weighted approach. Specifically, bootstrap aggregating (i.e., bagging) with PSM values as resampling weights was employed to compile a personalized training dataset for each patient.

II. METHODS

In order to make a direct comparison with the results reported in [4], the present study analyzed the same patient

[1]Joon Lee is with the School of Public Health and Health Systems, University of Waterloo, Waterloo, ON, Canada N2L 3G1. joon.lee@uwaterloo.ca

978-1-5090-2455-1/16/$31.00 ©2016 IEEE

data using the same PSM and predictive models. The methods of the present study are identical to those of [4] except the important difference in how bootstrapping was employed instead of hard-thresholding, as well as in how an ensemble approach was utilized instead of a single predictive model. Hence, much of the methods described in [4] is repeated in this section.

A. Patient Data

The patient data for this study were extracted from MIMIC-II [8]. MIMIC-II is an ICU database that includes clinical data from over 29,000 adult ICU admissions (version 2.6 [9]) at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA, USA. Because MIMIC-II is a public, de-identified database, the need to obtain informed consent or research ethics approval was waived.

A number of variables were extracted from each ICU admission as predictors. First, the minimum and maximum values of the following vital signs were extracted from each 6-hour period during the first 24 hours in the ICU: heart rate, mean blood pressure, systolic blood pressure, SpO2, spontaneous respiratory rate, and body temperature. Second, again from the first 24 hours in the ICU, the minimum and maximum values of the following blood lab tests were extracted: hematocrit, white blood cell count, glucose, HCO3, potassium, sodium, blood urea nitrogen, and creatinine. The rationale for using minimum and maximum values was that either the minimum or maximum could be the worst value, and worst values are known to be predictive of mortality outcomes, as can be seen in SAPS calculation [10]. Third, the following additional variables were pulled out: admission type (elective, urgent, emergency), gender, ICU service type (medical, surgical, coronary care, cardiac surgery), receipt of vasoactive agents during the first 24 hours in the ICU (binary), use of mechanical ventilation or Continuous Positive Airway Pressure during the first 24 hours in the ICU (binary), admission age, minimum (i.e., worst) Glasgow Coma Scale, and the total urinary output from each 6-hour period during the first 24 hours in the ICU. Mortality at 30 days post-hospital-discharge, represented as a binary variable, was the target outcome to be predicted.

ICU admissions with missing data were excluded from the study. Moreover, ICU admissions were treated as distinct patients and no special consideration was given to ICU admissions from the same patient. Not only that the vast majority of the patients included in MIMIC-II have only one ICU admission each (there are 1.24 ICU admissions per patient in MIMIC-II), but also this was a conscious decision to objectively identify similar clinical cases regardless of patient identity. Hence, note that ICU admissions are referred to as "patients" in this article.

All patient data were extracted from MIMIC-II using Structured Query Language (SQL) in Oracle SQL Developer (version 3.2.09).

B. Patient Similarity Metric

The cosine similarity PSM was defined as follows:

$$PSM(\mathbf{P_1}, \mathbf{P_2}) = \frac{\mathbf{P_1} \bullet \mathbf{P_2}}{\|\mathbf{P_1}\|\|\mathbf{P_2}\|} \quad (1)$$

$\mathbf{P_1}$ and $\mathbf{P_2}$ are the predictor vectors from two different patients, while $\bullet$ and $\|\cdot\|$ represent the dot product and Euclidean vector magnitude, respectively. Because this PSM is the cosine of the angle between $\mathbf{P_1}$ and $\mathbf{P_2}$, it is naturally normalized between -1 (minimum similarity) and 1 (maximum similarity).

Each continuous predictor variable was linearly transformed to fit the range between -1 and 1. For the categorical predictors (e.g., gender, admission type), the product between two vectors in that particular dimension in the feature space was assigned a value of 1 if the two patients belonged to the same category, or a value of -1 otherwise.

C. Personalized Mortality Prediction Models

Three types of predictive models were evaluated in 10-fold cross-validation: death counting (DC, the mortality rate among similar patients as the predicted mortality risk), logistic regression (LR), and decision tree (DT). For each patient in the test data as the index patient (i.e., in a leave-one-out fashion), the steps below were followed:

1) All pairwise PSM values between the index patient and every patient in the training data were calculated.
2) A bootstrap sample of 500 or 1000 similar patients was constructed by resampling the training data with replacement and with the PSM values as resampling weights (i.e., the greater the PSM, the more likely the corresponding patient is included in the bootstrap sample).
3) Each of the three predictive models was trained using the bootstrap sample as training data, and subsequently yielded a number between 0 and 1 as the predicted mortality risk for the index patient.
4) Steps 2 and 3 above were repeated 100 times for each of the two bootstrap sizes.
5) For each predictive model and for each bootstrap size, the final prediction for the index patient was the average of the 100 predictions arising from the different bootstrap samples.

For performance evaluation, the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) were employed. AUPRC is an informative performance metric for skewed datasets such as the one investigated in this study (the overall mortality rate was much less than 50%) [11]. AUPRC values range from zero to one, where one indicates a perfect prediction performance. Unlike AUROC, however, random guessing does not yield an AUPRC of 0.5 for skewed datasets.

In order to mitigate the effects of the skewed dataset in this study, the 10-fold cross-validation incorporated stratified sampling to ensure that the ratio between the positive (expired) and negative (survived) cases in each fold was similar to that in the entire dataset. Thus, the patients were divided into two groups based on mortality outcome, and

then random assignment to the 10 cross-validation folds was carried out in each group independently.

All computations and analyses were conducted in R (version 3.1.1).

III. RESULTS

A total of 17,152 adult ICU admissions in MIMIC-II had complete data and were included in the study. The overall 30-day mortality rate was 15.1%.

Fig. 1 and Fig. 2 show the AUROCs and AUPRCs, respectively, from the three predictive models. The two bootstrap sizes are also compared. With the bootstrap size of 1000, LR achieved a mean AUROC [95% confidence interval] of 0.815 [0.809, 0.821] and a mean AUPRC of 0.449 [0.437, 0.461]), while DT resulted in a mean AUROC of 0.510 [0.501, 0.520] and a mean AUPRC of 0.161 [0.154, 0.167]). Both LR and DT showed no significant difference between bootstrap sizes of 500 and 1000, in terms of both AUROC and AUPRC.

On the other hand, the mean AUROC of DC was significantly greater with the bootstrap size of 1000 than 500 (0.663 [0.655, 0.672] vs. 0.637 [0.626, 0.647], p=0.001). DC did not show a significant difference between the two bootstrap sizes in terms of AUPRC.

Overall, LR resulted in the best predictive performance, whereas DT exhibited the worst performance. In comparison with the best performance reported in [4] for each model, only LR achieved a reasonably similar performance while DC and DT substantially underperformed, in terms of both AUROC and AUPRC. Compared to the performances of the two widely used ICU SOI scores reported in [4] (SAPS and SOFA), DC achieved a similar performance while LR and DT outperformed and underperformed them, respectively.

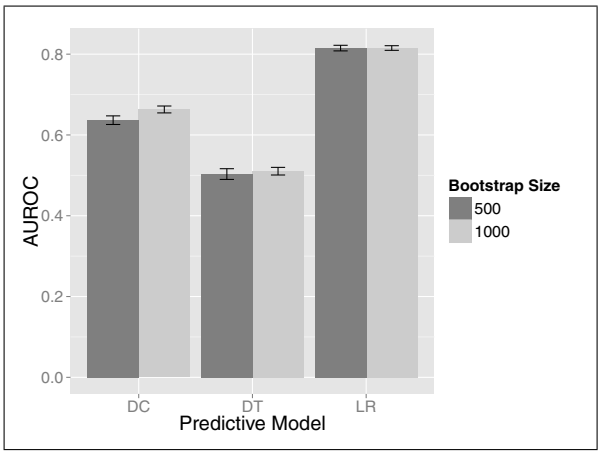

Fig. 1. AUROCs from three personalized models with bagging, with a comparison between two bootstrap sizes. Bars and error bars indicate mean values and their 95% confidence intervals, respectively, based on 10-fold cross-validation. DC: death counting; DT: decision tree; LR: logistic regression.

IV. DISCUSSION

The present study investigated a new method to utilize the cosine-similarity PSM introduced in [4], which deployed bagging instead of hard-thresholding to compile a personalized training dataset for each index patient that only includes similar patients. Compared to hard-thresholding, the results showed that bagging resulted in a similar performance for LR but worse performances for DC and DT. While it is true that bootstrapping can help avoid the two major problems with hard-thresholding (namely, a lack of variety in training data and the difficulty with optimizing the hard threshold), it is important to note that predictive performance degraded substantially for two of the three models.

The results from the present study should be considered preliminary since only two hand-picked bootstrap sizes (i.e., 500 and 1000) were evaluated. Since other bootstrap sizes may lead to different results, it is premature to make a definitive conclusion about the general predictive performance of PSM-based bagging. However, the results of this study revealed that bootstrap size had no significant impact on predictive performance, with the exception that DC's AUROC significantly improved by increasing the bootstrap size from 500 to 1000. It is anticipated that very small bootstrap sizes would likely lead to the same problem that hard-thresholding faces: a lack of variety in categorical predictors and in the outcome variable (e.g., if the index patient is male, every case in the personalized training data may end up being male as well with a small bootstrap size). Conversely, very large bootstrap sizes would naturally lead to increased computational burden, which can become a major issue particularly for ensemble methods like bagging that need to train many models. Although bootstrap size often does not need to be exhaustively investigated, the need for an evaluation of different bootstrap sizes undermines the advantage of bagging over hard-thresholding.

Among the three models, the predictive performance of DT suffered the most with bagging in comparison with hard-thresholding. The fact that its performance was essentially no better than random guessing implies that bagging completely erased the predictive utility of the variables used in this study. This result is surprising given that the combination of DT and bagging is similar to random forests which often yield very good performance in many real-life applications. It may also be the case that DT is more vulnerable to contamination from dissimilar patients in training data than LR or DC (with bootstrapping, it is possible to have a small number of very dissimilar patients in training data). Furthermore, this speculative reason may apply to all three models in general since bagging failed to improve upon the previously reported performances. The results from this study seem to suggest that excluding dissimilar patients from training data is at least as important as including similar patients, and this observation was also made in [4].

As more and more electronic medical data are becoming available, there has been a rapidly increasing interest in harnessing the power of big health care data in recent years [12], [13]. In particular, intensive care is one of the most data-intensive specialties in medicine, and Celi et al. [14] have described the potential role of data analytics in making decision making more evidence-based and in reducing health care costs in the ICU. Accurate prognostic information from

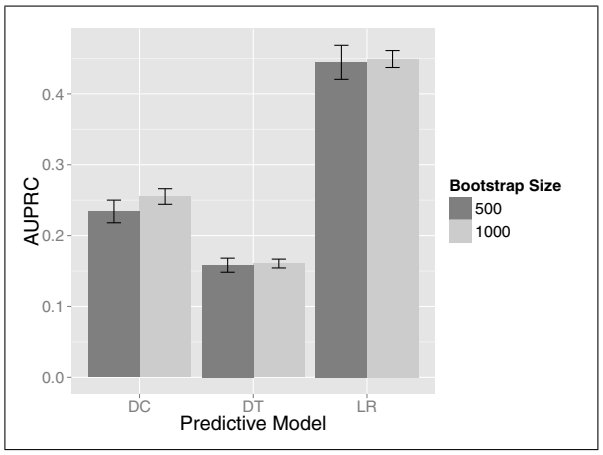

Fig. 2. AUPRCs from three personalized models with bagging, with a comparison between two bootstrap sizes. Bars and error bars indicate mean values and their 95% confidence intervals, respectively, based on 10-fold cross-validation. DC: death counting; DT: decision tree; LR: logistic regression.

personalized algorithms can greatly facilitate treatment planning, allocation of scarce ICU resources, and conversations between clinicians and patients/their families. Rich datasets like MIMIC-II enable detailed patient similarity matching in a way that is analogous to genomics-based personalized medicine. In the era of big data, it is important to understand the extent to which personalized care is feasible with EMR data. More future studies in this area are both anticipated and needed.

For future work, an investigation of other PSM types would be worthwhile, since the difference between bagging and hard-thresholding reported in this study could be specific to the cosine-similarity PSM that was used. Furthermore, inspired by the combination of DT and bagging, it would be interesting to apply the case-specific random forest (CSRF) [15] to MIMIC-II data. Similarly to our PSM-based methods, CSRFs compile and utilize a personalized training dataset for each index case. However, two major differences between CSRFs and the DT+bagging combination used in this study are: 1) CSRFs employ randomly selected subsets of features; and 2) CSRFs use the random forest proximity measure to weight training cases for bootstrapping.

V. CONCLUSIONS

In comparison with hard-thresholding, bagging failed to improve the mortality prediction performances of three personalized models. The results support hard-thresholding as a better approach for mortality prediction in the ICU, at least for the cosine-similarity PSM.

ACKNOWLEDGMENT

The author was supported by a Discovery Grant (RGPIN-2014-04743) from the Natural Sciences and Engineering Research Council of Canada (NSERC). The author would also like to thank the University of Waterloo for providing general support and resources for research.

REFERENCES

[1] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, "Acute physiology and chronic health evaluation (apache) iv: Hospital mortality assessment for todays critically ill patients," *Critical care medicine*, vol. 34, no. 5, pp. 1297–1310, 2006.

[2] R. P. Moreno, P. G. Metnitz, E. Almeida, B. Jordan, P. Bauer, R. A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, J.-R. Le Gall, *et al.*, "Saps 3from 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 medicine*, vol. 31, no. 10, pp. 1345–1355, 2005.

[3] J. Lee and D. M. Maslove, "Customization of a severity of illness score using local electronic medical record data," *Journal of intensive care medicine*, p. 0885066615585951, 2015.

[4] J. Lee, D. M. Maslove, and J. A. Dubin, "Personalized mortality prediction driven by electronic medical data and a patient similarity metric," *PLoS ONE*, vol. 10, no. 5, 2015.

[5] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[6] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 160, no. 3, pp. 523–541, 1997.

[7] J. Sun, F. Wang, J. Hu, and S. Edabollahi, "Supervised patient similarity measure of heterogeneous patient records," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 1, pp. 16–24, 2012.

[8] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database," *Critical care medicine*, vol. 39, no. 5, p. 952, 2011.

[9] J. Lee, D. J. Scott, M. Villarroel, G. D. Clifford, M. Saeed, and R. G. Mark, "Open-access mimic-ii database for intensive care research," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 8315–8318.

[10] J.-R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers, "A simplified acute physiology score for icu patients." *Critical care medicine*, vol. 12, no. 11, pp. 975–977, 1984.

[11] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[12] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

[13] S. Schneeweiss, "Learning from big health care data," *New England Journal of Medicine*, vol. 370, no. 23, pp. 2161–2163, 2014.

[14] L. Anthony Celi, R. G. Mark, D. J. Stone, and R. A. Montgomery, "big data in the intensive care unit. closing the data loop," *American journal of respiratory and critical care medicine*, vol. 187, no. 11, pp. 1157–1160, 2013.

[15] R. Xu, D. Nettleton, and D. J. Nordman, "Case-specific random forests," *Journal of Computational and Graphical Statistics*, no. just-accepted, pp. 00–00, 2014.