MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190219

Combining Structured and Unstructured Data for Predicting Risk of Readmission for Heart Failure Patients

Satish M. Mahajan[a], Rayid Ghani[b]

[a] Veterans Affairs Palo Alto Health Care System, Palo Alto, California, USA,
[b] Center for Data Science & Public Policy, University of Chicago, Chicago, Illinois, USA

Abstract

Researchers have studied many models for predicting the risk of readmission for heart failure over the last decade. Most models have used a parametric statistical approach while a few have ventured into using machine learning methods such as statistical natural language processing. We created three predictive models by combining these two techniques for the cohort of 1,629 patients from six hosptials using structured data along with their 136,963 clinical notes till their index admission, stored in the EMR system over five years. The AUCs for structured and combined models were very close (0.6494 and 0.6447) and that for the unstructured model was 0.5219. The clinical impact of the models using decision curve analysis showed that, at a threshold predicted probability of 0.20, the combined model offered 15%, 30%, and 70% net benefit over its individual counterparts, treat-all, and treat-none strategy respectively.

Keywords:

Heart Failure; Machine Learning; Electronic Health Records

Introduction

The Affordable Care Act of 2010 in the United States mandated penalties for health organizations with high readmission rates in an effort to improve quality of care and patient outcomes. The Centers for Medicare and Medicaid Services (CMS) implemented the law via the Readmissions Reduction Program (HRRP) [10]. This program has survived three repeal efforts in the past 18 months, and, in fact, has continued to improve its risk assessment strategies for readmissions. The risk formula now looks at readmission rates for six major diseases (heart failure, elective hip/knee arthroplasty, pneumonia, chronic obstructive pulmonary disease, myocardial infraction, and coronary artery bypass graft) for stratified hospital comparisons and uses it for prospective payments to hospitals for Medicare and Medicaid patients.

This change has launched the interest and development of risk prediction models for patient readmissions to hospitals. Availability of data in Electronic Health Records (EHR) systems has pushed this development further. Many variations of models for predicting readmissions have been developed, particularly for heart failure, in the last few years [8]. Heart failure (HF) epidemiology highlights this interest: HF affects 6.6 million adults in the United States alone and 550,000 new cases are reported annually [3]. It is the most frequent and expensive disease category for 30 day readmissions [2]. The chronicity of HF disease is further marked by repeated hospitalizations and consumes 70% of total cost related to the disease [4]. Moreover, as a part of the American Heart Association's Policy Statement on forecasting the future of cardiovascular diseases in the United States, Heidenreich et al. [5] provide projections for the 2010 to 2030 timeframe for HF: an increase of 215% in direct costs, 80% in indirect costs, and 25% in prevalence respectively.

A closer examination of HF predictive models for readmission reveals their specific characteristics: mortality and readmission appear to be influenced by different predictors and hence many models have preferred to treat them as independent outcomes. Many recent studies have used a 30-day timeframe for modeling readmission risk in line with the CMS lead, whereas older studies have used a somewhat variable timeframe ranging from seven days to one year [8]. The 30-day timeframe appears to be a strong representative of readmissions for many diseases, as a similar readmission pattern arises for different diseases based on time-to-event analyses [6]. Furthermore, most models have considered all-cause readmission to encourage hospitals to implement intervention programs to improve overall care—such as better discharge coordination, medication reconciliation, and post-discharge follow-up—instead of using a narrow set of condition-specific solutions. In terms of data sources used for predictor selection, recent models prefer to use EHR systems over registry or claims-based datasets [8].

Some predictive models [7] have used parametric algorithms derived mostly from generalized linear methods bolstered with modern strategies, such as variable class weighting, cross-validation, and regularization. These models emphasize the importance of interpretability; however, they require significant construction time especially with larger predictor sets and their interactions. Furthermore, they asymptotically become numerically unstable and even fail to build due to combinatorial explosion during variable selection process. Meanwhile, other predictive models [9] have deployed non-parametric methods mostly using machine learning techniques with large datasets comprising many predictors. This class of models focuses more on predictive accuracy at the cost of sensitivity to the configuration parameters of the underlying algorithms as well as to the features of datasets. Howerver, combining and comparing both the approaches with the same underlying patient population remains unexplored.

In this study, we built three predictive models: one based on the parametric approach using structured predictors readily available in the EHR system; the second based on a non-parametric approach using Natural Language Processing (NLP) on unstructurred data representing patient notes in the EHR system; and a third that combined predictors from both the approaches. We further compared discriminative power,

calibration, and usefulness of each model in clinical decision making.

Methods

Patient Population and Data Extraction

We used EHR data at six hospitals from the Veterans Health Administration system to derive and validate the predictive models for this observational retrospective cross-sectional study. We selected all the hospitals from one network of hospitals from the western region of the United States for this study. The datasets from all hospitals under the health system are extracted and loaded into a central repository on a regular basis and we leveraged this infrastructure for the data collection for our study.

Based on our review of past literature [8] describing the predictive models built for estimating risk of readmission for heart failure, we used variables from clinical, administrative, and psychosocial categories that are routinely collected during episodes of care. Application of the same data policy framework imposed by the EHR system helped us to develop common logic and programs for data extraction. Datasets for five consecutive calendar years from 2011 to 2015 were extracted using International Classification of Diseases version 9 – Clinical Modification (ICD-9-CM) codes for heart failure as a principal discharge diagnosis. We used 2011 – 2014 data for derivation cohort and 2015 data for validation cohort formations. This setup provided an external temporal validation for the models. Since the validation cohort consisted of patients from different regions, the setup also provided an external geographic validation for the models.

Primary Outcome

We adhered to the commonly adopted definition of any-cause 30-day readmission in the literature [6], which excludes hospitalizations with a length of stay less than one day and any elective hospitalizations. If the patient encountered multiple readmissions within the 30-day timeframe from the discharge date of the index admission, only the last readmission episode was considered so that the latest health status for the patient is used for predicting the risk of future readmission. This setup provided us with the sample of patients that was blinded for outcome by programming for the above rules. The sample was also statistically independent and mutually exclusive across the two classes of admission groups of non-readmitted and readmitted patients.

Predictors

Structured data predictors were broadly classified into clinical, administrative, and psychosocial categories based on prior literature survey and expert clinical consultation [8] and are shown in Table 1. All the variables were collected either on or before the discharge date but closest to it depending on the data availability.

Unstructured data for the patients in the cohort were obtained from clinical notes—such as history and physician note at admission, progress notes, social worker notes, and discharge summary—till the date of discharge for the index admission. Since the bulk extraction of the structured and unstructured data from the EHR system was carried for the study cohort, predictor assessment was blinded to the other predictor and outcome variables.

Table 1 – Structured Dataset

Clinical Set	Predictor Name
Labs & Vitals	Sodium
	Potassium
	Blood Urea Nitrogen
	Creatinine
	Hemoglobin
	Hematocrit
	Glucose
	Albumin
	B-Natriuretic Peptide
	Systolic/Diastolic Blood Pressure
	Pulse
	Respiratory Rate
Administrative & Psychosocial Sets	**Predictor Name**
Demographics	Age
	Sex (Male/Female)
	Race (Native American/Asian/ Black/Hispanic/White/Not specified)
	Marital Status (Married/Divorced/ Separated/Unknown)
	Insurance Type (Multiple Insurance/ Blue Cross-Shield/Special Medicare /Disability/Income Medicaid/None)
	Residential Area (Urban/Rural/Highly Rural/Unknown)
Pre-Index Admission Factors	Appointments in Past Year
	No Show to Appointment in Past Year
	ED Visits in Past Year
	Prior Diagnoses
	Admissions in Previous Year
	Telemetry Monitor During Ind. Adm. (Yes/No)
	Index Adm. via ED (Yes/No)
	Length of Stay
	Concurrent Procedures
Comorbidities and Concurrent Procedure	Alcohol Abuse (Yes/No)
	Cardiac Arrhythmia (Yes/No)
	Coronary Artery Disease (Yes/No)
	Cancer (Yes/No)
	Cardiomyopathy (Yes/No)
	Cerebrovascular Accident (Yes/No)
	Depression (Yes/No)
	Diabetes Mellitus (Yes/No)
	Drug Abuse (Yes/No)
	Functional Disability (Yes/No)
	Liver Disease (Yes/No)
	Lung Disease (Yes/No)
	Lung Disease (Yes/No)
	Protein Caloric Malnutrition (Yes/No)
	Psychiatric Disorder (Yes/No)
	Rheumatic Disease Group (Yes/No)
	Renal Disease Group (Yes/No)
	Vascular Disease Group (Yes/No)
	Aortic Valve Disorder (Yes/No)
	Cancer related (Yes/No)
	Cardiac Devices (Yes/No)
	Cardiac Surgery (Yes/No)
	Coronary Angioplasty (Yes/No)
	History of Mechanical Ventilation Devices (Yes/No)

Statistical Analysis

The extraction of data from the EHR system yielded a sample size of 1629 admissions with 114 variables with dummy variable formats for categorical variables. We did not have any missing values for the dependent variable; a few independent variables such as BNP and blood glucose had up to 5% missing values. We used multiple imputation by chained equations resampled over five imputed datasets for the missing values assuming missingness at random. We created separate dummy variables for missing values with higher rate of missingness for the categorical variables. This strategy allowed us to use all the available records in the analysis. Continuous variables were also examined for nonlinear effects and transformations were carried wherever necessary.

We extracted 54 different types of clinical notes totaling 136,963 (102,055 for derivation cohort and 34,908 for validaton cohort) for our study and could not find any notes for 10 patients for the timeframe of interest. We eliminated these patients from the structured dataset for fair comparison of models using structured and unstructured datasets. Our final structured dataset thus contained 1,619 total patients (1,279 for derivation cohort and 340 for validation cohort). The corpus datasets for derivation and validation cohorts were created separately and processed for tokenization and stop words. They were then treated with term frequency-inverse document frequency statistic for vectorization with bigram range and term frequency thresholds of 0.01 and 0.80 for minimum and maximum cutoff respectively. We linked the structured and unstructured datasets on patients and their admission episodes for creating a combined dataset.

We fitted all three models using logistic regression with their respective feature sets to predict the outcome of 30-day readmission and reported their results for the validation datasets. We did not apply any special variable selection methods in this study. All confidence intervals were calculated at 95% level using 2,000 stratified bootstrap replicates. We further calculated and plotted the discriminative power using c-statistic, precision, recall, and model calibration for each model along with the top 10 words most predictive of readmission in the unstructured data. We also carried out decision curve analysis using the three models to understand their relative utilities with respect to treat-none and treat-all strategies. We used Microsoft SQL Server Version 12 Release 2, R version 3.0.2 (The R Foundation for Statistical Computing Platform), and Python version 3.6.4 for data extraction and statistical analyses. This Pilot Project under Quality Improvement initiatives was exempted from the review of full committee Institutional Review Board.

Results

Our validation cohort's empirical readmission rate was 32.7% excluding repeated patient readmissions. We created and fitted three separate models with structured, unstructured, and the combined datasets. Figure 1(A) shows precision and recall plotted on the left and right Y-axis respectively against percentage of population on the X-axis for the models. The Area-Under-Curve (AUC) or c-statistic was calculated for the models and plotted as Receiver Operating Characteristic (ROC) curves as shown in Figure 2(A). The AUC values with 95% confidence interval were: structured model: 0.6494 [0.5885-0.7103]; unstructured model: 0.5219 [0.5157-0.5281]; combined model: 0.6447 [0.6386-0.6508].

Calibrations with ideal, apparent, and bias-corrected fits with 300 bootstrap repetitions for the models were carried out and plotted as actual probabilities on Y-axis against model-predicted probabilities on X-axis as shown in Figure 1(B). The word sets with the top 10 words that were most predictive of readmission versus non-readmission using notes are shown in Figure 2(B). The Feature Importance Plot for the combined dataset is shown in Figure 2(C).

Finally, we carried out decision curve analysis (DCA) [11] for the models to understand net benefit for each model at various thresholds of predicted probabilities. It is shown in Figure 2(D).

Discussion

As shown in Figure 1(A), precision and recall seem to have benefited from combining the structured and unstructured datasets especially when examining the 20% of the population with the highest precision: the combined dataset precision climbed to 70% from 45-50% precision of the individual datasets. Similarly, the combined dataset recall climbed to 40% from 20-30% recall of the individual datasets. Precision measures the accuracy of the model when it predicts an index admission as likely to have a readmission. Recall measures the ability of the model to find index admissions that have readmissions. As there is a tradeoff between these two measures, they can be varied and plotted for various thresholds of population as well.

As shown in Figure 2(A), the ROC curves represent the discriminative ability of the models by plotting the ratio of true positives to predicted positives on Y-axis against the ratio of true negatives to predicted negatives on X-axis. The curve that represents the maximum pull in the northwest direction of the graph signifies the best discrimination, while the diagonal axis represents random prediction. The unstructured data model had poor discrimination in our study, whereas the structured data model and combined model had similar discrimination with the combined model showing slightly better performance for non-readmitted patients. Our empirical datasets represented high class-imbalance (~30% readmitted versus 70% non-readmitted patients) and hence precision-recall curves might be better measures of model performance in the absence of tuning performed on the class-weight parameter.

Calibration plot scheme using bootstrapping method works as an efficient bias estimator since the difference between fits of the whole sample and a new sample is estimated by the difference between fits of the whole sample and a sample with replacement. The plots in Figure 1(B) show presence of over-fitting for predicted probabilities over 0.35 for all the models. The ideal fits with no over-fitting and no over-estimation are shown as the dotted diagonal axis at 45^{o} in the plots. All the models also show under-fitting below 0.35 of predicted probabilities. Moreover, all the models are under-estimated with positive Y-axis intercepts indicating need for better functional form for the models.

When we identified the top 10 words for both readmitted and non-readmitted patients as shown in Figure 2(B), there was only one dyad (*pulmonary discomfort*) common to both the classes. However, closer examination of the word sets reveals that some other words, even though written differently, are semantically equivalent: for example, word pairs such as *gcs commands* (Glasgow Command Scale) and *respond meaningfully*; *regarding mrsa* (Methicillin-Resistant Staphylococcus Aureus) and *mrsa status*. Such semantically

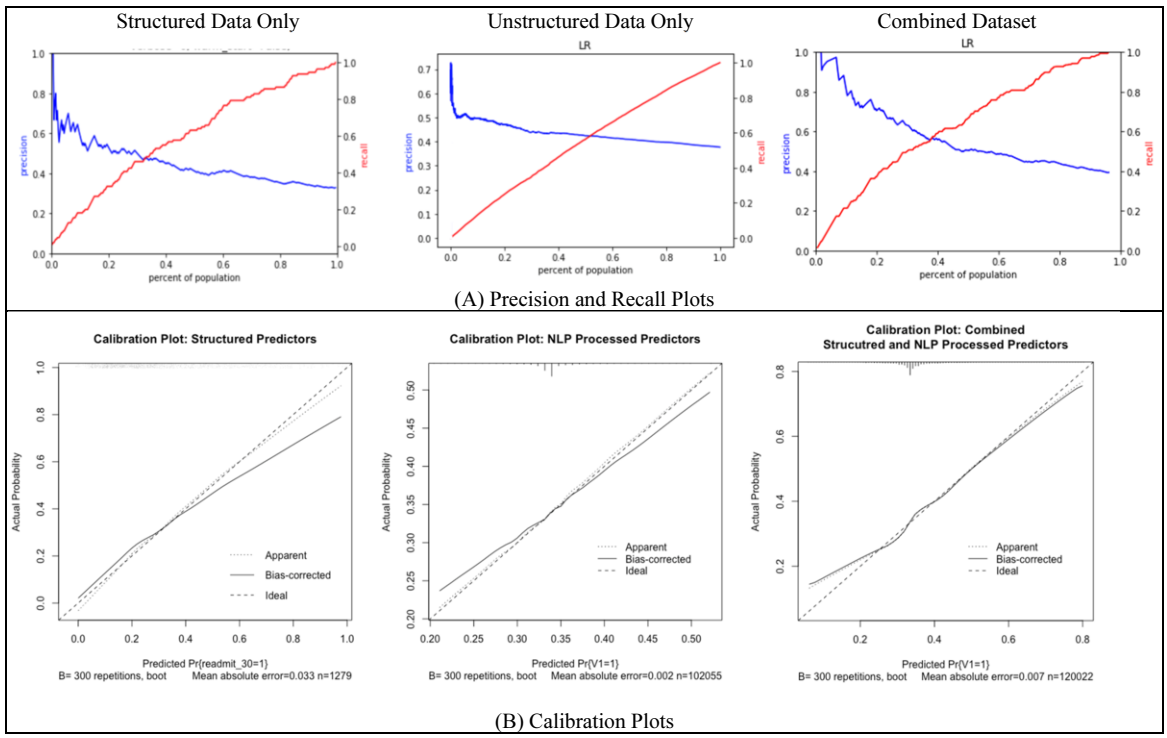

Figure 1 – Plots with Validation Cohort for Models

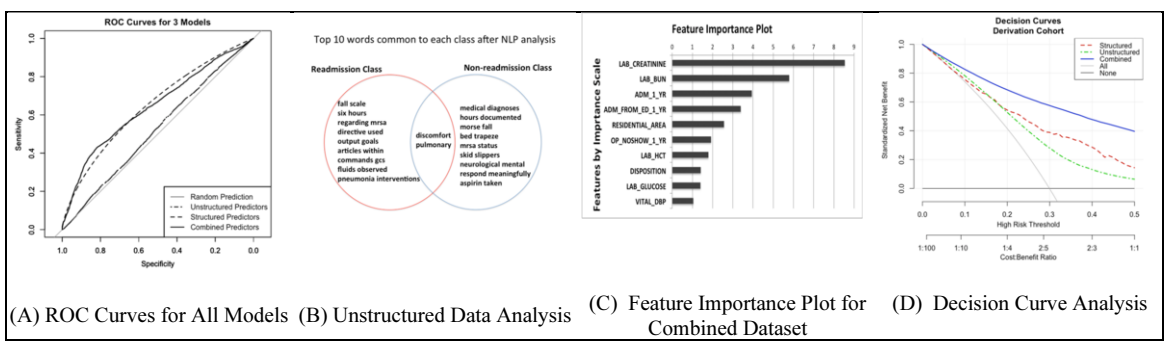

Figure 2 – Decision Analysis for Models

similar words need to be clustered together to further guide supervised learning. Given a large set of clinical documents for each patient episode, it would also be useful to further analyze specific documents with targeted word selections that have greater impact on readmission prediction.

The combined dataset has over 4900 predictors and hence we showed a plot of relative importance of the top 10 predictors in the model in Figure 2(C). Laboratory tests such as creatinine and hematocrit from the clinical set of structured predictors along with some administrative predictors such as patient hospital boarding characteristics and office visits status have strong influence on the readmission risk in our model. These predictors seem to be in line with the ones reported by the previous studies. Particularly, no predictor from the unstructured dataset made up to the top 10 list of the influencers of the patient readmission in the combined set.

Note that logistic regression is used as a baseline for the model development in all three. This is merely for convenience during methods development in this pilot study. The demonstrated methods could be used with any other parametric or machine learning models or the combination of both for structured and NLP datasets.

DCA is a method of evaluating usefulness of predictive models in clinical settings. It combines strengths of traditional biostatistical methods and decision-analytic methods while eliminating their limitations. For example, AUC for the model does not convey how high of an AUC is needed in order to deploy it in clinical practice. On the other hand, traditional decision analysis involves complex mathematics not suitable for continuous outcomes that is typical of predictive models. DCA is based on the relative harms caused by false positives and false negatives and is expressed as a threshold for

probabilities predicted by the models. As seen in Figure 2(D), the use of any predictive model provides positive standardized net benefit over both *treat none* (X-axis) and *treat all* (lowest curve cutting X-axis at risk threshold of 0.30) strategies. Either structured or unstructured data model appears to provide about the same net benefit of 55% over *treat none* and 15% over *treat all* strategies at risk threshold of 0.20. Moreover, the combined model (topmost curve in Figure 2(D)) shows consistently higher standardized net benefit over the other two models (15%) as well as *treat all* (30%) and *treat none* (70%) strategies.

Clinical Impact and Implications for Practice

We have combined and compared two seemingly different techniques for predicting 30-day readmissions for heart failure using most commonly available data in the EHR systems. Methods presented here pave the foundation for combining automated harvesting of predictors from unstructured data with carefully selected predictors from structured dataset. We have further demonstrated the ability to select the most useful model that has the greatest clinical impact. Decision analysis in this study has shown that the use of *some* predictive model is better than both using *no* model at all and treating *all* the admitted patients with the same readmission reduction program interventions. This appears to be valid despite having average descriminative capabilities (AUC = 0.6447) of the predictive model. Discharging physicians and case managers could make decisions for post-discharge care with their patients by identifying risks and benefits using such models. Even though the model development and associated activities of data processing pipeline might seem difficult to implement in practice, application of the selected risk threshold and outcome prediction could be captured in a simple mobile application for the clinicians in future implementations.

Limitations

This study represents a pilot project that brings together techniques for combining structured and unstructured data into a model for predicting readmissions for heart failure. It has used data from all the clinical notes for the patients till the index admission. Targeted use of notes with specific topic and word importance might help in better performing NLP system and remains unexplored in this study. Readmissions outside the index hosptials, even though deemed low, are not considered in this study. Finally, the study has applied external temporal validation but external geographical validation will further evaluate generalizability of this approach.

Conclusions

We built separate predictive models with the same EHR system's structured and unstructured data with two different modeling techniques. We compared the performance metrics for the two models and their combination. Our analysis showed that there was some benefit in combining these data although AUC for structured data model (0.6494) was not very different from the combined data model (0.6447). The combined model showed better results in terms of calibration plot and precision-recall curves. The decision curve analysis for assessing the clinical impact of practical usefulness of the models showed that the combined model offered 15% net benefit over its individual counterparts at a threshold predicted probability of 0.20. We have described the model development and validation efforts for this study using the Transparent Reporting of a multivariate prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [1].

Acknowledgements

Part of this work was carried under the collaboration between co-authors via 2018 Data Science Rotations for Advancing Discovery Trip program. This work was performed under operational and quality improvement initiatives at Veterans Affairs Palo Alto Health Care System and views expressed are those of the authors and not necessarily those of the system.

References

[1] G.S. Collins, J.B. Reitsma, D.G. Altman, and K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, *BMJ* **350** (2015), g7594.

[2] N. Fida and I.L. Pina, Trends in heart failure hospitalizations, *Curr Heart Fail Rep* **9** (2012), 346-353.

[3] A.A. Fitzgerald, L.A. Allen, and F.A. Masoudi, The evolving landscape of quality measurement for heart failure, *Ann N Y Acad Sci* **1254** (2012), 131-139.

[4] M.J. Hall and C.J. DeFrances, National Hospital Discharge Survey, in: *Advance data from vital and health statistics*, National Center for Health Statistics, Hyattsville, MD, 2001.

[5] P.A. Heidenreich, J.G. Trogdon, O.A. Khavjou, J. Butler, K. Dracup, M.D. Ezekowitz, et al., C. American Heart Association Advocacy Coordinating, C. Stroke, R. Council on Cardiovascular, Intervention, C. Council on Clinical, E. Council on, Prevention, A. Council on, Thrombosis, B. Vascular, C. Council on, C. Critical, Perioperative, Resuscitation, N. Council on Cardiovascular, D. Council on the Kidney in Cardiovascular, S. Council on Cardiovascular, Anesthesia, C. Interdisciplinary Council on Quality of, and R. Outcomes, Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association, *Circulation* **123** (2011), 933-944.

[6] H.M. Krumholz, L. Horwitz, C. Partovian, Z. Lin, J. Herrin, J. Grady, M. Conover, J. Montague, C. Dillaway, K. Bartczak, J. Ross, S. Bernheim, and E. Drye, Hospital-Wide (All-Condition) 30-Day Risk-Standardized Readmission Measure, in, Yale New Haven Health Services Corporation/Center, New Haven, CT, 2008.

[7] S.M. Mahajan, P. Burman, A. Newton, and P.A. Heidenreich, A Validated Risk Model for 30-Day Readmission for Heart Failure, *Studies in health technology and informatics* **245** (2017), 506-510.

[8] S.M. Mahajan, P.A. Heidenreich, B. Abbott, A. Newton, and D. Ward, Predictive models for identifying risk of readmission after index hospitalization for heart failure: A systematic review, *European Journal of Cardiovascular Nursing* (2018).

[9] S.M. Mahajan, A.S. Mahajan, R. King, and S. Negahban, Predicting Risk of 30-Day Readmissions Using Two Emerging Machine Learning Methods, *Studies in health technology and informatics* **250** (2018), 250-255.

[10] United States Center for Medicare and Medicaid Services, Readmissions Reduction Program (HRRP), 2018.

[11] A.J. Vickers and E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, *Med Decis Making* **26** (2006), 565-574.

Address for correspondence

Satish M. Mahajan, VA Palo Alto Health Care System, Mailstop 118R, 3801 Miranda Ave, Palo Alto, CA 94304, USA.

Email: satish.mahajan@va.gov.