Five Year Life Expectancy Calculator for Older Adults

Ankit Agrawal[1], Jason Mathias[2], David Baker[2], Alok Choudhary[1]
[1]Department of Electrical Engineering and Computer Science
[2]Northwestern Memorial Hospital
Northwestern University
Corresponding author email: ankitag@eecs.northwestern.edu

***Abstract*—Incorporating accurate prognostic information into clinical decision making could advance evidence-based, person-centered healthcare by more effectively targeting healthcare services to those patients most likely to benefit. Here we describe the deployment of predictive models for five year life expectancy of patients, built on electronic health records (EHR) of nearly 7,500 patients aged 50 and above, with one or more visits to a large, academic, multispeciality hospital in a year, exploring more than 75 modeling configurations. The online web-tool takes a non-redundant subset of 24 patient attributes as input and generates a patient-specific prediction of 5-year survival. The online five year life expectancy calculator is available at http://info.eecs.northwestern.edu/FiveYearLifeExpectancyCalculator**

***Keywords*-Healthcare informatics, supervised learning, ensemble learning, electronic healthcare records, five-year life expectancy**

I. Motivation

When making healthcare decisions, failure to consider an individual patient's prognosis can lead to poor quality care and waste healthcare resources. First, patients with a good prognosis may fail to receive beneficial services that could improve their quality of life and/or longevity. For example, healthy older patients often have low rates of cancer screening despite the potential benefits. Second, patients with a poor prognosis may receive services that are not beneficial, may be harmful, and/or waste financial resources. For example, chronically ill patients with limited life expectancy frequently receive preventive services for which the potential risks (e.g., decreased quality of life, invasive follow-up tests or treatments) outweigh any potential benefits Finally, patients with a poor prognosis may fail to receive beneficial services that could improve their quality of life and preserve their independence. For example, few patients participate in advanced care planning despite evidence that doing so increases patient satisfaction, preserves independence, enables person-centered care at the end of life, and relieves caregiver burden.

Development and deployment of accurate life expectancy prediction models can also have a tremendous economic impact. The Centers for Disease Control and Prevention estimates that there are more than 150,000 surgical-site infections annually [1], and it can cost $11,000 to $35,000 per patient, i.e., about $5 billion every year. Accurate predictions and risk estimation for healthcare outcomes can potentially avoid thousands of complications, resulting in improved resource management and significantly reduced costs.

Big data analytics [2] and the advent of the 4th paradigm of science [3] in healthcare provides unprecedented opportunities to utilize the big healthcare data itself as a resource to get actionable insights. A variety of healthcare data is increasingly becoming available, such as electronic healthcare records, genomic sequence data, x-ray images, social media, etc. There has been a growing interest in data-driven analytics on such heterogenous biological and healthcare data [4], [5], [6], [7], [8], [9], [10], [11].

Motivated by the above challenges and opportunities, this demonstration paper presents an online health informatics tool to predict 5-year life expectancy of older adults, which is a key piece of information required to make informed clinical decisions for older adults. The models deployed in the tool are a result of the application of many supervised learning techniques on a high-dimensional electronic healthcare records (EHR) database. The development of the specific model deployed in the tool was described earlier in [12], and was shown to outperform other better known prognostic indices, like the Charlson Comorbidity Index [13] and Walter Life Expectancy Index [14]. In this paper we re-analyze that data and deploy the most accurate predictive models for predicting 5-year survival in a user-friendly web-tool.

II. Design

The overall data-driven process is depicted as a block diagram in Figure 1. EHR data is extracted for patients with at least one visit to NMFF in 2003. Nearly 1,000 attributes were derived, in a bid to provide domain knowledge to the predictive model. Multiple rounds of automatic attribute selection interleaved with manual inspection and selection were performed to identify a small non-redundant attribute set with good predictive power, to construct the 5-year life expectancy prediction database. Supervised learning techniques are then used to learn predictive models, which are evaluated using standard validation techniques, and the most accurate models are deployed in an online user-friendly web-tool that can estimate the 5-year life expectancy of a patient as a relative probability value.

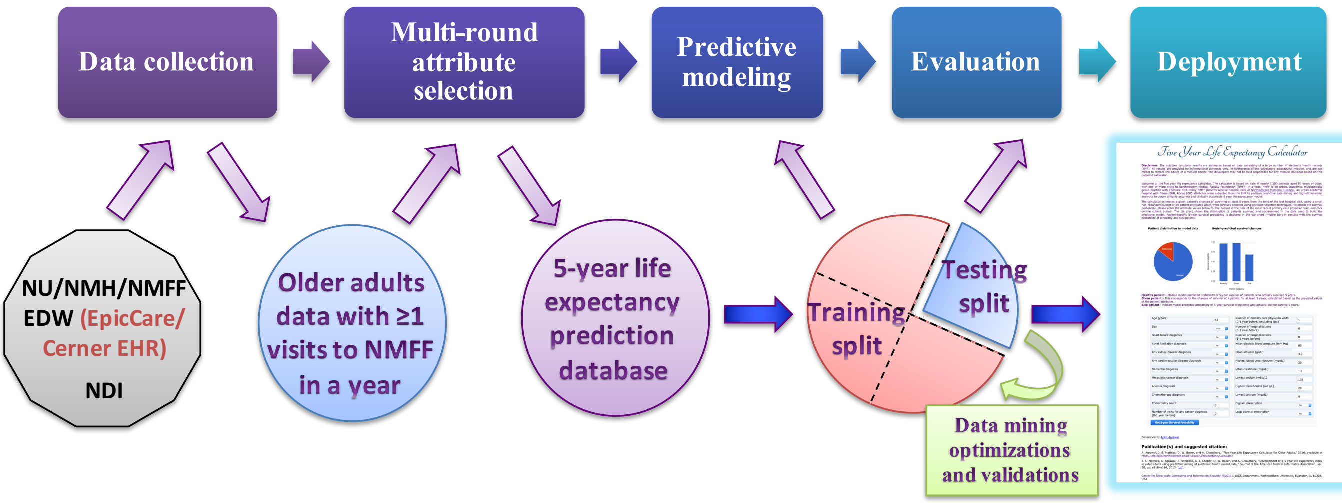

Fig. 1. The design of the data mining workflow used in this work.

III. Development and Functionality

A. Data

We use the same dataset as used in [12]. Patient-level EHR data was extracted from the Enterprise Data Warehouse (EDW) implemented by Northwestern University (NU), Northwestern Memorial Hospital (NMH), and Northwestern Medical Faculty Foundation (NMFF) using Cerner and Epic EHR. Patients with at least one visit to NMFF in 2003 were selected. This was linked with the National Death Index for the years 2003-2008, as we are interested in modeling 5-year survival, which is recommended to be considered while making decisions about preventive service use (e.g. cancer screening).

A total of 980 predictive attributes for 7,463 patients were derived. These attributes included all a priori plausible predictors of mortality available within the EHR, including 11 sociodemographic attributes, 117 comorbidities, 20 vital signs, 120 laboratory results, 664 possible medications, and 48 healthcare utilization attributes. Please refer to [12] for details. Feature selection techniques were used to find a subset of 52 features that were highly correlated with the outcome but weakly correlated amongst themselves. This set was manually reviewed to remove certain attributes of low face validity, with potentially problematic reliability, and some redundant features, resulting in a smaller subset of 32 attributes. This set was again analyzed with automated feature selection techniques reducing it to 23, to which sex was added for a final set of 24 attributes, plus the dichotomous outcome attribute, which denoted whether or not the patient survived at least 5 years. The 5-year life expectancy prediction database, therefore, had 7,463 instances and 25 attributes.

B. Methods

We used more than 75 configurations of classification schemes in this study, including both direct application of classification techniques and constructing their ensembles using various ensembling techniques (compatible combinations). Evaluation metrics included the area under the ROC curve (c-statistic), classification accuracy, precision, recall, and F-measure.

C. Results

Table I presents the top five modeling techniques. All results are based on 10-fold cross-validation. In addition, the training and testing times for each model, and the model size is also listed. WEKA software [15] verison 3.7.13 was used for all analytics with default parameters, unless otherwise stated. Table I is sorted by the AUC metric. The predictive models used in [12] were based on Rotation Forest ensembling technique with alternating decision trees as the underlying base modeling technique, which ranks 2^{nd} in our current experiments. The best accuracy was obtained by the Rotation Forest ensembling technique with LogitBoostBasedADTree (LADTree) as the underlying base modeling technique (AUC=0.8612). However, statistical significance testing revealed that the AUC obtained by the $RotationForest\_LADTree$ is not statistically distinguishable from $RotationForest\_ADTree$ at p=0.05. This is not surprising as both the models are essentially based on the principle of alternating decision trees.

D. Five year life expectancy calculator

We have created an online 5-year life expectancy calculator that can take as input the values of the 24 non-redundant patient attributes (see Table II), and estimate the patient's chances of surviving at least 5 years from the time of the last hospital visit.

To improve the calibration of the models, Stacking technique with logistic modeling as the meta learning technique was used to calibrate the predictions from the $RotationForest\_ADTree$ model. In this way, the final

TABLE I

TOP FIVE MODELING TECHNIQUES (10-FOLD CROSS-VALIDATION RESULTS)

Modeling Scheme	*AUC*	*Accuracy (%)*	*Precision*	*Recall*	*Fmeasure*	*TrainTime (s)*	*TestTime (s)*	*ModelSize (bytes)*
RotationForest_LADTree	0.8612	90.29	0.8869	0.9029	0.8855	50.33	0.13	23249416
RotationForest_ADTree	0.8609	90.19	0.8856	0.9019	0.8808	27.65	0.35	538272
Bagging_MLP	0.8534	89.82	0.8800	0.8982	0.8817	63.29	0.20	113977
RotationForest_MLP	0.8534	89.90	0.8813	0.8990	0.8834	56.23	0.67	615019
RandomSubSpace_MLP	0.8522	89.88	0.8812	0.8988	0.8739	30.76	0.28	117328

TABLE II

PATIENT ATTRIBUTES USED AS INPUT IN THE FIVE YEAR LIFE EXPECTANCY CALCULATOR

Attribute	*Brief description*
Age	Numeric age of the patient
Sex	Gender of the patient (male/female)
Heart failure diagnosis	Whether or not patient has been diagnosed with heart failure (yes/no)
Atrial fibrillation diagnosis	Whether or not patient has been diagnosed with atrial fibrillation (yes/no)
Any kidney disease diagnosis	Whether or not patient has been diagnosed with any kidney disease (yes/no)
Any cardiovascular disease diagnosis	Whether or not patient has been diagnosed with any cardiovascular disease (yes/no)
Dementia diagnosis	Whether or not patient has been diagnosed with dementia (yes/no)
Metastatic cancer diagnosis	Whether or not patient has been diagnosed with metastatic cancer (yes/no)
Anemia diagnosis	Whether or not patient has been diagnosed with anemia (yes/no)
Chemotherapy diagnosis	Whether or not patient has been given chemotherapy (yes/no)
Comorbidity count	Number of comorbidities
Number of visits for any cancer diagnosis (0-1 year before)	Number of doctor visits for any cancer diagnosis in the past year
Number of primary care physician visits (0-1 year before)	Number of primary care physician visits in the past year
Number of hospitalizations (0-1 year before)	Number of hospitalizations in the past year
Number of hospitalizations (1-2 years before)	Number of hospitalizations between one and two years prior to the current date
Mean diastolic blood pressure	Numeric value of mean diastolic blood pressure (in mm Hg)
Mean albumin	Numeric value of mean albumin (in g/dL)
Highest blood urea nitrogen	Numeric value of highest blood urea nitrogen (in mg/dL)
Mean creatinine	Numeric value of mean creatinine (in mg/dL)
Lowest sodium	Numeric value of lowest sodium (in mEq/L)
Highest bicarbonate	Numeric value of highest bicarbonate (in mEq/L)
Lowest calcium	Numeric value of lowest calcium (in mg/dL)
Digoxin prescription	Whether or not the patient has been prescribed dogoxin (yes/no)
Loop diuretic prescription	Whether or not the patient is on loop diuretic prescription (yes/no)

$Stacking\_RotationForest\_ADTree$ model is deployed in the 5-year life expectancy calculator presented here. The screenshot of the calculator is depicted in Figure 2, and it is available online at http://info.eecs.northwestern.edu/FiveYearLifeExpectancyCalculator.

To obtain the survival probability, the attribute values for a patient at the time of the most recent primary care physician visit needs to be submitted in the form available on the website. A pie chart shows the distribution of patients survived and not-survived in the data used to build the predictive model. Patient-specific 5-year survival probability is depicted in the bar chart in context with the survival probability of a healthy and sick patient, which are essentially the median risk of death of patients who actually survived and did not survive 5-years respectively, as calculated by the model. The bar chart has 3 bars. The middle bar denotes the patient-specific risk, and the left (right) bars denote the healthy (sick) patient risk. The patient-specific risk is thus put in context of the healthy and sick patient risk for an informative comparison.

IV. SIGNIFICANCE

From a research perspective, this work investigates the applicability of predictive modeling techniques to predict patient-specific healthcare outcomes. In particular, here we focus on predicting the 5-year survival chances of an older adult, by comparing over 75 supervised modeling configurations on a EHR dataset of about 7,500 patients from Northwestern Memorial Hospital.

From a practical point of view, we have deployed the most accurate predictive models in a user-friendly web-tool for easy access. Although a slightly more accurate model compared to [12] was found, its performance was found to be not statistically distinguishable at p=0.05, and thus only the original model was deployed in the calculator. To make the developed predictive models for 5-year survival readily accessible, we have created an online five year life expectancy calculator that can take patient characteristics as input and make patient-centered life expectancy predictions. The primary advantage of this tool is that it is a general calculator and not disease-specific, and offers the capability of quickly estimating the chances of 5-year survival of an older individual at the point of care, thereby facilitating personalization, assisting in clinical decision support, and enhancing informed patient consent. The deployed tool is expected to be a useful resource for a variety of stakeholders in healthcare, such as patients, doctors and healthcare providers, researchers in this field, insurance companies, and so on.

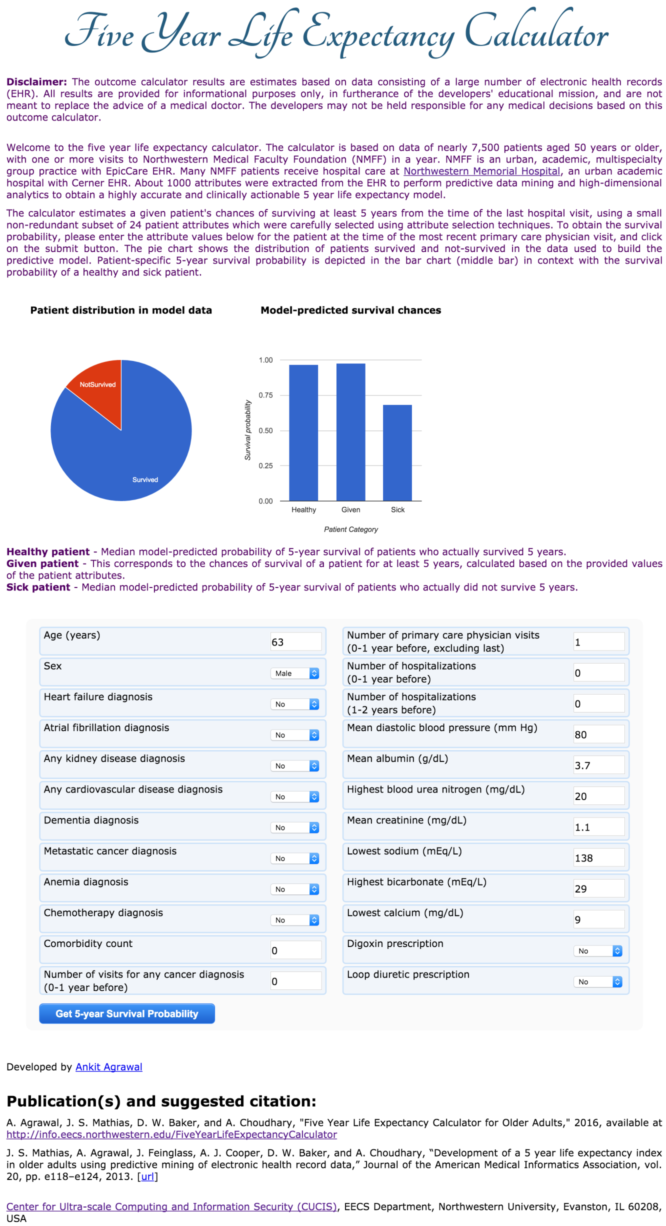

Fig. 2. A screenshot of the deployed 5-year life expectancy calculator.

Acknowledgment

This work is supported in part by the following grants: NSF awards IIS-1343639, CCF-1409601; DOE awards DE-SC0007456, DE-SC0014330, NIST award 70NANB14H012; AFOSR award FA9550-12-1-0458.

References

[1] S. S. Magill, J. R. Edwards *et al.*, "Multistate point-prevalence survey of health care-associated infections," *New England Journal of Medicine*, vol. 370, no. 13, pp. 1198–1208, 2014, pMID: 24670166.

[2] A. Agrawal and A. Choudhary, *Big data analytics for deriving predictive healthcare insights*, ser. Handbook of Health Services Research. Springer, 2016.

[3] T. Hey, S. Tansley, and K. Tolle, Eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research, 2009.

[4] A. Agrawal and X. Huang, "Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 1, pp. 194–205, 2011.

[5] S. Misra, A. Agrawal, W.-k. Liao, and A. Choudhary, "Anatomy of a hash-based long read sequence mapping algorithm for next generation dna sequencing," *Bioinformatics*, vol. 27, no. 2, pp. 189–195, 2011.

[6] A. Agrawal and A. Choudhary, "Identifying hotspots in lung cancer data using association rule mining," in *2nd IEEE ICDM Workshop on Biological Data Mining and its Applications in Healthcare*, 2011, pp. 995–1002.

[7] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "A lung cancer outcome calculator using ensemble data mining on seer data," in *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics (BIOKDD)*, 2011, pp. 1–9.

[8] A. Agrawal, R. Al-Bahrani, J. Raman, M. J. Russo, and A. Choudhary, "Lung transplant outcome prediction using unos data," in *Proceedings of the IEEE Big Data Workshop on Bioinformatics and Health Informatics (BHI)*, 2013, pp. 1–8.

[9] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G.-Z. Yang, "Big data for health," *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, pp. 1193–1208, 2015.

[10] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: Demonstration on flu and cancer," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 2013, pp. 1474–1477.

[11] Y. Xie, Z. Chen, Y. Cheng, K. Zhang, A. Agrawal, W.-k. Liao, and A. Choudhary, "Detecting and tracking disease outbreaks by mining social media data," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013, pp. 2958–2960.

[12] J. S. Mathias, A. Agrawal, J. Feinglass, A. J. Cooper, D. W. Baker, and A. Choudhary, "Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data," *Journal of the American Medical Informatics Association*, vol. 20, pp. e118–e124, 2013.

[13] A. J. Perkins, K. Kroenke *et al.*, "Common comorbidity scales were similar in their ability to predict health care costs and mortality," *Journal of clinical epidemiology*, vol. 57, no. 10, pp. 1040–1048, 2004.

[14] L. C. Walter and K. E. Covinsky, "Cancer screening in elderly patients: a framework for individualized decision making," *Jama*, vol. 285, no. 21, pp. 2750–2756, 2001.

[15] M. Hall, E. Frank *et al.*, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.