Five Year Life Expectancy Calculator for Older Adults

Ankit Agrawal1, Jason Mathias2, David Baker2, Alok Choudhary1
1Department of Electrical Engineering and Computer Science
2Northwestern 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 heterogeneous 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.
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] version 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 2nd 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 RotationForestLADTree is not statistically distinguishable from RotationForestADTree 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 RotationForestADTree model. In this way, the final...
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 an 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.
Five 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**


**Development by Ankit Agrawal**

**Publication(s) and suggested citation:**


Center for Ultra-scale Computing and Information Security (CUCIS), EECS Department, Northwestern University, Evanston, IL 60208, USA

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