Prediction in claims databases for epidemiological research

David A. Pritchard, PhD, Target RWE

Prediction opportunities in epidemiological research

Healthcare data is becoming larger, more complex, and more integrated which opens up exciting new innovative opportunities. Some research questions that machine learning can help to answer include:

- Patient / physician characteristics associated with lower diagnosis rates
- Identifying high risk populations for disease areas
- Signal detection and syndromic definition of diseases
- Effective patient and physician targeting strategy for interventions to reduce the burden of disease
- Identifying the most effective populations in which to administer a particular treatment

Selecting an index date

The index date is the time to start follow up

- Some studies are naturally anchored on a specific event such as diagnosis of a particular disease
- For a general population, the first non-urgent outpatient office visit is often a reasonable choice

Prediction considerations and challenges

Inclusion / exclusion criteria

- The inclusion / exclusion criteria are the list of conditions that a patient has to satisfy to be included in the study data.
- Determines the population that you hope your model can generalize to. Examples: patients over 18, diabetic patients, patients without a complicating comorbidity (study dependent)

Feature engineering overview for longitudinal healthcare data

A central challenge for prediction using longitudinal healthcare data is how to construct the input data (feature engineering).

- With a slight oversimplification, the data for a given patient can be considered as being a stream of events such that each event encodes a *what happened*, and a *when it happened*
- Traditional machine learning models assume a fixed-length covariate vector for each patient

Feature engineering techniques

The goal is to transform each patient's longitudinal history into a fixed-length covariate vector that summarizes as much of the predictive information as possible. Some available techniques include:

- Subject-matter experts postulate what information is likely to be predictive of outcome
- Automated variable creation
  - Aggregate information into time intervals. E.g. how many times did a particular event occur during a given interval
  - Hard to know what the vocabulary (i.e. the universe of possible events) should be. One approach is to consider all codes observed above a certain threshold proportion

Deep learning potential

- One of the potential advantages of deep learning is that the model can itself learn a suitable data representation
- Recurrent neural networks are designed to process sequential data of variable length such as is present in healthcare data

Training models with right censored data

There are several approaches to handling right censored data for prediction problems.

- Complete case analysis
- Using methods explicitly designed for right censored data such as penalized Cox proportional hazards, support vector regression for right censored data, survival trees / forests
- Using inverse probability censoring weighting (IPCW) with existing methods not designed for right censored data
  - Estimate the survival distribution of the censoring times
  - All patients who did not have an event and were censored before the end of follow up are dropped from the study data, and all other patients are upweighted according to their inverse probability of being uncensored

Prediction metrics for right censored data

Measuring the accuracy for machine learning methods is challenging since we don't know everybody's outcome. The most commonly used prediction metrics for right censored data are *Harrell's C index* and the *net reclassification improvement*.

Study overview

The study goal was to predict the likelihood that subjects would suffer a hip fracture (among other fracture types) within one year of follow-up

- Based on an administrative claims database
- Used the first non-urgent outpatient office visit as the index date
- The main inclusion criteria were:
  - Observed an outpatient office visit during the study period
  - Subject age 50 years or older at time of index
  - Continuous enrollment of at least 730 days prior
- There were 2.1 million patients in the study population, and 75% had a full year of follow up
- Among patients for whom full follow-up was observed, 0.62% observed a hip fracture

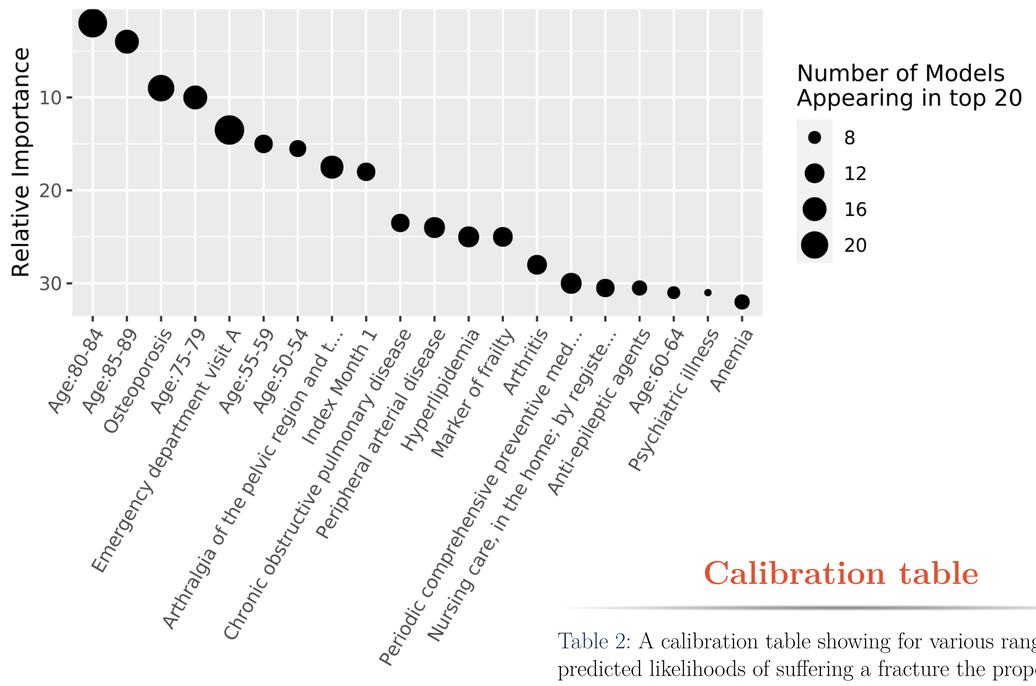
Modeling choices

- Feature engineering:
  - Constructed 53 expert-input variables, including demographics information, various comorbidities, and conditions or treatments thought to be associated with risk of fracture such as ASVCD, glucocorticoids, SSRI therapy, opioids, etc.
  - Derived indicator variables for an additional 800 variables that appeared in at least 1% of the training data, and filtered down to the 50 most correlated variables using sure independence screening
- Used IPC weighting to construct patient weights to account for censoring
- Results are shown for lasso. Other models tested for performance include random forests, gradient boosting machines, deep neural networks, and stacked ensembles models

Example fracture prediction study

Variable importance measures

Figure 1: Multiple machine learning models were trained across various fracture outcomes. The variables with the best mean relative importance are shown from left to right. The circle size shows how consistently a variable was important



Confusion matrix

Table 1: The confusion matrix for a prediction threshold chosen to maximize the F1 score

		Predicted	
		No event	Hip fracture
Actual	No event	310,690	1,649
	Hip fracture	6,003	249

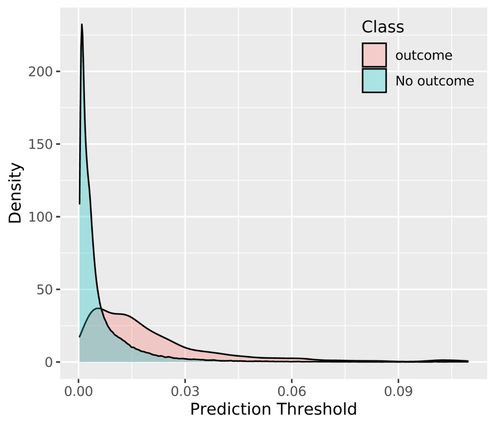
Calibration table

Table 2: A calibration table showing for various ranges of predicted likelihoods of suffering a fracture the proportion of patients who actually observed a fracture

Bucket	Proportion	Count
0.00 to 0.02	0.0042	296,318
0.02 to 0.04	0.0262	16,422
0.04 to 0.06	0.0360	3,613
0.06 to 0.08	0.0448	1,271
0.08 to 0.10	0.0356	506
0.10+	0.0586	461

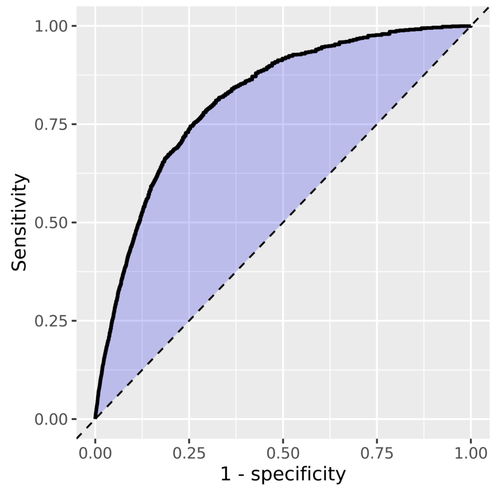
Class-conditional densities curve

Figure 2: The test data are separated according to their true outcome, and their density functions are constructed using the predicted likelihoods of suffering a fracture



ROC curve

Figure 3: The receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) is 0.82



TARGET RWE
BETTER EVIDENCE • BETTER HEALTH