Name:	Noémie Elhadad
Affiliation:	Columbia University
Email:	noemie@gmail.com
Presentation type (select one):	Poster

Deep Survival Analysis

Rajesh Ranganath[1], Adler Perotte[2], Noémie Elhadad[2], David Blei[2]
[1]Princeton University, Princeton, NJ, USA; [2]Columbia University, New York, NY, USA

Abstract

The OHDSI datasets provide an unprecedented opportunity to investigate predictive modeling at the population and individual levels. Here we investigate the task of survival analysis in the context of observational health record data. We present Deep Survival Analysis, a novel hierarchical generative approach to survival analysis. It departs from previous approaches in two primary ways: (1) all observations, including covariates, are modeled jointly conditioned on a rich latent structure; and (2) the observations are aligned by their failure time, rather than by an arbitrary time zero like in traditional survival analysis. Further, it (3) handles heterogeneous (continuous and discrete) data types that occur in an electronic health record in a scalable manner. We validate the deep survival analysis model on stratifying patients according to risk of developing Coronary Heart Disease. We train and test the model on a dataset of 313,000 patients corresponding to 5.5 million months of observations. When compared to the clinically validated Framingham CHD risk score, our model is significantly superior in stratifying patients according to their risk.

Introduction

Our goal is to leverage observational data like the ones in OHDSI to estimate the time of a future event of interest for given individual, namely, to carry out survival analysis. When used at the point of care, accurately estimating the time to an event can improve clinical decision support by allowing physicians to take risk-calibrated actions. To learn useful estimates however, large amounts of longitudinal patient trajectories are needed. We illustrate our work in the context of survival analysis for coronary heart disease (CHD). CHD, also known as coronary artery disease or ischemic heart disease, is the most common type of heart disease and the leading cause of death worldwide causing 1 in every 4 deaths. There are many effective lifestyle interventions and preventive therapies to reduce risk of morbidity and mortality of CHD, but because they themselves can be risky for patients, there is great value in identifying accurately the patients that are at high risk of a CHD event. Beyond CHD, there are many conditions where a similar need for estimating risk arises, and clinicians have routinely been relying on clinically validated risk scores for individual patients to make treatment decisions (e.g., prostate cancer, breast cancer, stroke).

The standard approach to developing risk scores hinges on regressing covariates to the time of failure on a curated set of patient data. The significant covariates in the analysis are then summarized in an easy-to-use table (e.g., the Wilson et al. table for CHD [1]. However, this approach has serious limitations when it comes to using it with data derived from observational data like in the EHR. First, regression requires complete measurement of the covariates for all patients; in practice, many are missing. Second, all patients are aligned based on some initial event (e.g., entry into trial, onset of a disease related to event of interest, start of medication, etc.). Third, the relationship between the covariates and the time of the medical event is assumed to be linear, possibly with some interaction terms.

Methods

We present a novel model for survival analysis from EHR data, which we call deep survival analysis. Full details of the model are available [2]. For the sake of space, we describe the main contributions of the model and our experiments here. Deep survival analysis handles the biases and other inherent characteristics of observational EHR data, and enables accurate risk scores for an event of interest. The key contributions of our method are:

• Deep survival analysis models covariates and survival time in a Bayesian framework, thus handling the missing covariates prevalent in EHR and OHDSI data;

• Deep exponential families [3], a deep latent variable model, forms the backbone of the generative process. This results in a non-linear latent structure that captures complex dependencies between the covariates and the failure time;

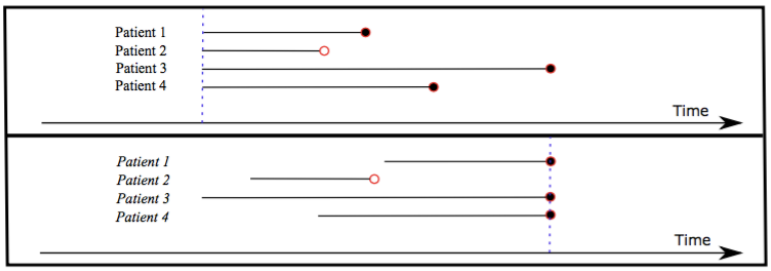

Figure 1. A comparison of traditional survival analysis (top frame) and failure aligned survival analysis (bottom frame). A filled circle represents an observed event, while an empty circle represents a censored one. In the case of standard survival analysis patients in a cohort are aligned by a starting event. In failure aligned survival analysis, patients are aligned by a failure event.

• Rather than enforcing an artificial time 0 alignment for all patients, deep survival analysis aligns all patients by their failure time (i.e., the event occurs or data is right censored) (see Figure 1);

• Good preprocessing of EHR data allows deep survival analysis to include heterogeneous data types. We include vitals, laboratory measurements, medications, and diagnosis codes;

Results

We experiment with the Columbia University OHDSI site dataset and focus on 313,000 adult patient records (309K for training, 2K for validation, and 2K for testing). We used deep survival analysis to assess the risk of coronary heart disease. In our experiments, we vary the dimensionality of the deep exponential family latent structure to assume the values of $K \in \{5, 10, 25, 75, 100\}$. The baseline clinically validated CHD risk score [1] yielded 68.06% in concordance over the held-out test set. In comparison, our model yielded 75.34% concordance (K=50) (Table 1).

While the concordance metric enables the comparison of the deep survival model to the baseline method, it captures only roughly the accuracy of the temporal prediction of the models. In the case of the deep survival model, we are able to compute the predictive likelihood of the held-out set according to the model. This enables us to capture how well the model predicts failure in time. Table 2 shows predictive likelihood, evaluated as the expected log probability of the observed time until failure under the deep survival analysis model conditioned on the observed covariates for a given patient in a given month. The diagnosis-only model yielded the best predictive likelihood.

Table 1. Concordance the deep survival analysis on held-out set for different values of K and for the baseline risk score.

Model	Deep Survival K=5	Baseline CHD risk score	Deep Survival K=10	Deep Survival K=100	Deep Survival K=25	Deep Survival K=75	Deep Survival K=50
Concordance (%)	65.81	68.06	68.22	71.02	73.77	74.80	75.34

Table 2. Predictive likelihood of the deep survival model (K=50) for individual data types.

Data type included in model	Medications only	Vitals only	Laboratory tests only	Diagnoses only
Predictive likelihood	-1.12538	-1.06543	-0.974539	-0.721999

Conclusion

While traditional survival analysis techniques require carefully curated research datasets, our approach handles the unavoidable data sparsity and heterogeneity of EHR observations. Our approach holds particular promise for developing risk scores from observational data for conditions where there is no known risk score.

References

1. Wilson PWF, DAgostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation, 97(18):1837–1847, 1998.
2. Ranganath R, Perotte AJ, Elhadad N, Blei DM. Deep Survival Analysis. In Machine Learning for Healthcare (MUCMD). 2016.
3. Ranganath R, Tang L, Charlin L, Blei DM. Deep Exponential Families. In Artificial Intelligence and Statistics. 2015.