Deep Survival Analysis

Rajesh Ranganath RAJESHR@CS.PRINCETON.EDU
Princeton University
Princeton, NJ 08540

Adler Perotte ADLER.PEROTTE@COLUMBIA.EDU
Columbia University
New York City, NY, 10032

Noémie Elhadad NOEMIE.ELHADAD@COLUMBIA.EDU
Columbia University
New York City, NY, 10032

David Blei DAVID.BLEI@COLUMBIA.EDU
Columbia University
New York City, NY, 10027

Abstract

The electronic health record (EHR) provides an unprecedented opportunity to build actionable tools to support physicians at the point of care. In this paper, we introduce *deep survival analysis*, a hierarchical generative approach to survival analysis in the context of the EHR. It departs from previous approaches in two main 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 as in traditional survival analysis. Further, it handles heterogeneous data types that occur in the EHR. We validate deep survival analysis by stratifying patients according to risk of developing coronary heart disease (CHD) on 313,000 patients corresponding to 5.5 million months of observations. When compared to the clinically validated Framingham CHD risk score, deep survival analysis is superior in stratifying patients according to their risk.

1. Introduction

Our goal is to use electronic health record (EHR) data to estimate the time of a future event of interest, namely, to carry out survival analysis in a healthcare context. Accurately estimating the time to an event improves clinical decision support by allowing physicians to take risk-calibrated actions. As a motivating example, consider coronary heart disease (CHD). It is the leading cause of death worldwide (Hansson, 2005; Pagidipati and Gaziano, 2013). This condition, also known as coronary artery disease or ischemic heart disease, is the most common type of heart disease and causes 1 in every 4 deaths. There are effective preventative therapies for CHD that can significantly reduce the risk of morbidity and mortality: antiplatelet therapy (ISIS-2 Collaborative Group, 1988), statin therapy (Scandinavian Simvastatin Survival Study Group, 1994; Sacks et al., 1996; Shepherd et al., 1995),

hypertensive therapy (Neal et al., 2000), and lifestyle interventions (Hjermann et al., 1981; Schuler et al., 1992). Given the numerous effective strategies for primary prevention (no prior CHD event) and secondary prevention (prior history of CHD event), there is great value to identifying those individuals at high risk of experiencing a CHD event. This is particularly important because these interventions, albeit effective, are not risk free.

The challenge of administering treatment based on risk pervades the clinical decision process, and risk scores are in use for many conditions, such as prostate cancer (Thompson et al., 2013), breast cancer (Gail et al., 1989), and stroke (Gage et al., 2001).

The standard approach to developing risk scores hinges on using a curated set of patient data to regress covariates to the time of failure. The significant covariates in the analysis are then summarized in a easy-to-use table (for CHD see Wilson et al. (1998)). However, this approach has serious limitations with respect to EHR data. First, regression requires complete measurement of the covariates for all patients; in practice, many are missing. Second, the traditional approach requires 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.); EHR data does not enjoy a natural alignment. Third, the relationship between the covariates and the time of the medical event is assumed to be linear, possibly with some interaction terms; this limits the kind of relationships that may be found.

In this paper we propose a novel model for survival analysis from EHR data, which we call *deep survival analysis*. Deep survival analysis handles the biases and other inherent characteristics of EHR data, and enables accurate risk scores for an event of interest. The key contributions of this work are:

- Deep survival analysis models covariates and survival time in a Bayesian framework. This simplifies working with the missing covariates prevalent in the EHR.
- Deep exponential families (Ranganath et al., 2015b), 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.
- Rather than enforcing an artificial time zero alignment for all patients, deep survival analysis aligns all patients by their failure time (i.e., the event occurs or data is right censored).
- Good preprocessing of EHR data allows deep survival analysis to include heterogeneous data types. In our study, we include vitals, laboratory measurements, medications, and diagnosis codes.

We studied a large dataset of 313,000 patient records and used deep survival analysis to assess the risk of coronary heart disease. Deep survival analysis better stratifies patients than the gold-standard, clinically validated CHD risk score (Wilson et al., 1998).

This paper is organized as follows. Section 2 reviews the fundamentals of traditional survival analysis and motivates the need for better modeling tools for EHR data. Section 3.1 reviews deep exponential families (Ranganath et al., 2015b) and Section 3.2 discusses our alignment

strategy for deep survival analysis. Section 3.3 describes the modeling assumptions behind deep survival analysis; Section 4.2 gives details of our scalable variational inference algorithm. Section 4 describes the clinical scenario of CHD, data, experimental setup, baseline, and evaluation metrics. Finally, Sections 5 and 6 discuss our results and conclusions.

2. Survival Analysis

In this section, we provide background on the task of survival analysis. We review the traditional approaches, along with several variants that are relevant to our work. We then delve into two of the primary limitations of current survival analysis techniques, which hinder their use in EHR data.

2.1 Traditional Survival Analysis

Survival analysis models the time to an event from a common start (Kaplan and Meier, 1958). Examples of survival data include time to delivery from conception and time to retirement from birth. Survival observations consist of two varieties. The first are observations for which the exact failure time is known. The second, called censored observations, are observations for which the failure time is known to be greater than a particular time. Both types are represented as (t, c), pairs of positive times and binary censoring status.

Survival modeling assumes the observations, both censored and uncensored, come from an unknown distribution. The two traditional methods for estimating the survival distribution are the Kaplan-Meier estimator (Kaplan and Meier, 1958) and Cox proportional hazards (Cox, 1972). The Kaplan-Meier estimator forms a nonparametric estimator for the survival function, one minus the cumulative distribution function. Intuitively, it breaks a stick of length one at points proportional to the fraction of patients that survived until that point. Cox proportional hazards generalizes this estimator to include covariates. Finally, Bayesian variations of these methods place priors on the parameters (Hjort, 1990).

2.2 Limitations of Traditional Survival Analysis for EHR Data

There are three significant limitations to using traditional survival methods on EHR data.

First, EHR data is usually high-dimensional and very sparse (Hripcsak and Albers, 2013; Pivovarov et al., 2014). This makes it difficult to use traditional conditional models, which cannot easily handle missing covariates

Second, traditional methods require aligning all patients based on a synchronization event. This event can be entry to a clinical trial, the date of an intervention, or the onset of a condition. However, the EHR for a patient can begin at any point in their lifetime and at any point in their disease progression. Thus careful definition of entry point into study are required when experimenting with traditional survival techniques on EHR data (e.g., Hagar et al. (2014); Perotte et al. (2015)). In this work, we seek methods that are able to evaluate risk at any point in time, not only points in time that correspond to such a synchronization event.

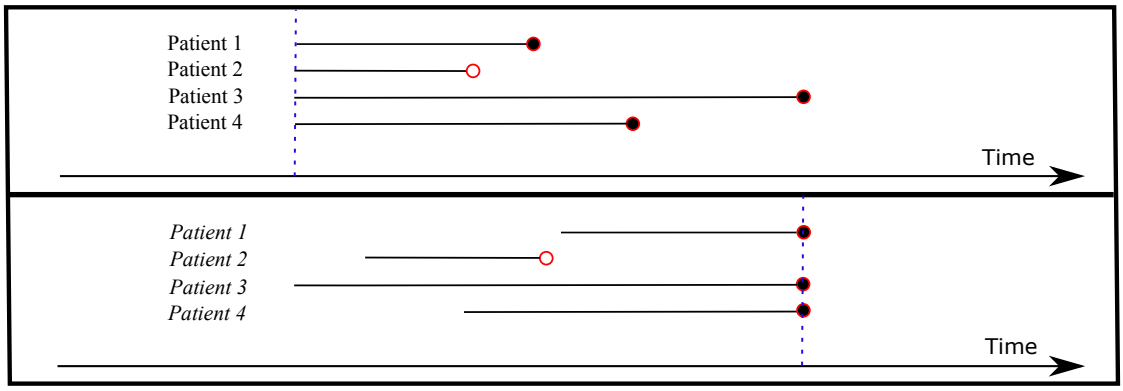

Figure 1: A comparison of traditional survival analysis (top frame) and failure aligned survival analysis (bottom frame). A filled circle represents an observed event, while an empty circle represents a censored one. In the case of standard survival analysis, patients in a cohort are aligned by a starting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

Third, regression-based approaches to survival analysis often assume a linear function of the covariates. Nonlinear interaction terms are sometimes introduced but must be limited (often based on expert opinion) because of the combinatorial explosion of possibilities. Models with greater flexibility can incorporate nonlinear relationships between combinations of covariates and time-to-event.

3. Deep Survival Analysis

We introduce *deep survival analysis*, a hierarchical generative approach for survival analysis. It departs from previous approaches in two primary ways. First, all observations, including covariates, are modeled jointly and conditioned on a rich latent structure. Second, patient records align by their failure time rather than by entry time, thus resolving the ambiguity of entry to the EHR.

3.1 Deep Exponential Families

When not modeling the covariates, missing covariates are usually imputed using population-level statistics. In contrast, we build a joint model for both the covariates and the survival times, where the covariates and survival times are specified conditioned on a latent process. This strategy requires a rich latent process; we use deep exponential families (Ranganath et al., 2015b).

Deep exponential families (DEF) are a class of multi-layer probability models built from exponential families (Brown, 1986). In deep exponential families, each observation has L layers of latent variables. Each layer conditions on the previous layer of latent variables.

Formally, let n be the index of the data, $\text{EXPFAM}(\cdot)$ be an exponential family distribution with natural parameter, $g(\cdot; W)$ denote a link function with parameters W with prior $p(W)$, and η be a hyperparameter. The generative process for the latent variables is

$$z_{n,L,j} \sim \text{EXPFAM}_L(\eta)$$
$$z_{n,\ell,j} \sim \text{EXPFAM}_\ell(g_\ell(z_{n,l+1}, W_{l,j})).$$

The observations for the ith data point are drawn conditional on the vector $z_{i,1}$. We shorthand the draw of the last layer, $z_{n,1}$ as $z_n \sim \text{DEF}(\mathbf{W})$. DEFs have been successful at modeling text, recommender systems, and images. They handle missing data better than competitive latent variable models and state-of-the-art density estimators (Rezende et al., 2014; Ranganath et al., 2015b). Thus, they are a promising prior for the latent structure to model survival in the EHR.

3.2 Alignment by Failure

Censored survival observations are pairs (t_i, c_i), where t_i denotes the time of the ith observation, and c_i marks whether failure or censoring occurs at that time. Traditionally, the time t_i is measured from a common start point for each observation, such as birth or pregnancy. As we mentioned earlier, this type of alignment is inappropriate in the context of the EHR because people enter the EHR in different ways relative to to their underlying health. This limitation of survival analysis was acknowledged by McCullagh (2013), but their solution has only been tested on a small dataset, and does not directly apply to censored observations.

We consider an event-centric ordering, which measures time backwards from the event of interest, rather than measuring time forward from an artificial start time. At the event all patients share the defining characteristics of the event, thus patients are similar at time zero under this alignment. We handle censored observations as interval observations.

For each time point, this alignment models the time to failure from that time point. This is a positive number which decreases when approaching failure. Censored events differs. For censored patients, their failure must happen (if it happens) after their last interaction with the EHR; their time of interest is an interval greater than the time of their last visit to the EHR. As such, event-centric ordering of data consists of pairs (t_i, c_i). See Figure 1 for a graphical comparison of this versus the standard survival setup.

In this approach, every interaction with the EHR has a (possibly censored) time from event/failure associated with it. This means the interactions can be modeled exchangeably, which trades statistical efficiency of persisting patient information over time for computational efficiency. We take this approach. Each event of interest in the EHR represents a different survival alignment frame. Thus, we can investigate several survival tasks by choosing an event of interest and aligning by its timing. This contrasts traditional survival analysis, which requires a careful decision about the start time.

To model the time from event, we use a Weibull distribution, a popular distribution in survival analysis. Let λ be the scale and k be the shape, the Weibull distribution is

$$p(t) = k/\lambda\,(t/\lambda)^{k-1}\,e^{-\left(\frac{t}{\lambda}\right)^k}.$$

It has support over the positive reals and its parameters are constrained to be positive. Its expectation is $\lambda\Gamma(1+\frac{1}{k})$. The parameter k control how the density looks. When $k<1$ most of the mass is concentrated near zero; when $k=1$ this distributions matches the exponential; when $k>1$, the distribution places mass around its expectation. For a censored observation, The likelihood is the amount of probability the model places after censoring, i.e., one minus the cumulative distribution function. For the Weibull, this is $\exp(-\left(\frac{t}{\lambda}\right)^k)$, which gets large as the scale grows.

3.3 Generative Process for Deep Survival Analysis

Let $\mathbf{x}$ denote the set of covariates, $\boldsymbol{\beta}$ be the parameters for the data with some prior $p(\boldsymbol{\beta})$, k be a fixed scalar, and let n index an observation. The generative model for deep survival analysis is

$$\begin{aligned}
b &\sim \text{Normal}(0, \sigma_b)\\
a &\sim \text{Normal}(0, \sigma_W)\\
z_n &\sim \text{DEF}(\mathbf{W})\\
\mathbf{x}_n &\sim p(\cdot \mid \boldsymbol{\beta}, z_n)\\
t_n &\sim \text{Weibull}(\log(1+\exp(z_n^\top a + b), k).
\end{aligned} \tag{1}$$

The latent variable z_i comes from a DEF which then generates the observed covariates and the time to failure. The function $\log(1+\exp(\cdot))$, called the softplus, maps from the reals to the positives to output a valid scale for the Weibull. Given covariates $\mathbf{x}$, the model makes predictions via the posterior predictive distribution:

$$p(t \mid \mathbf{x}) = \int_z p(t \mid z) p(z \mid \mathbf{x}) dz.$$

The complexity of the predictions depends on the complexity of the distribution z. Note this predictive distribution exists and is consistent even if data are missing.

For electronic health records the $\mathbf{x}$ contain several data types. We consider laboratory test values (labs), medications (meds), diagnosis codes, and vitals. We assume each of these data types are generated independently, conditional on the latent structure

$$p(\mathbf{x}_n \mid \boldsymbol{\beta}, z_n) = p(x_n^{\text{labs}} \mid z_n, \beta^{\text{labs}}) p(x_n^{\text{meds}} \mid z_n, \beta^{\text{meds}}) p(x_n^{\text{diagnoses}} \mid z_n, \beta^{\text{diagnoses}}) p(x_n^{\text{vitals}} \mid z_n, \beta^{\text{vitals}}).$$

We emphasizes that they are marginally dependent.

The data types in the EHR can be grouped by whether they are real valued (labs and vitals) or counts (diagnoses and medications). Next we define the likelihood for each group.

Real-Valued Observations. Real-valued observations in EHR are heavy tailed and are prone to data entry errors (Hauskrecht et al., 2013). This leads to extreme outliers that may badly corrupt estimates of non-robust models such as those based on the Gaussian. We model the real-valued data with the Student-t distribution, a continuous mixture of Gaussians across scales, which is more robust to outliers. Given parameters, $\beta_{W_i}^{\text{labs}}, \beta_{b_i}^{\text{labs}}$ and degrees of freedom ν, the conditional density of the ith lab, is

$$p(x_{n,i}^{\text{labs}} \mid \beta_{W_i}^{\text{labs}}, \beta_{b_i}^{\text{labs}}, z_n) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}} \left(1 + \frac{(x_{n,i}^{\text{labs}} - (z_n^\top \beta_{W_i}^{\text{labs}} + \beta_{b_i}^{\text{labs}})^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

This is a Student-t distribution whose mode is at $z_n^\top \beta_{W_i}^{\text{labs}} + \beta_{b_i}^{\text{labs}}$ a function of both the data point specific latent variables and the parameters shared across data points. The degrees of freedom controls to which extent the distribution resembles a Naussian, where large values look more Gaussian. We place Gaussian priors on both W_i and b_i. The likelihood follows similarly for the vitals.

Count Valued Observations. Unlike the laboratory tests and the vitals, the count-valued observations are highly dimensional and sparse. To handle counts robustly, we model them as binary values, one if the count is non-zero and zero otherwise. We model the ith medication with parameters $\beta_{W_i}^{\text{meds}}$ as

$$p(x_{n,i}^{\text{meds}} \mid \beta_{W_i}^{\text{meds}}, z_n) \sim \text{Bernoulli}(1 - \exp(z_n^\top \beta_{W_i}^{\text{meds}}))),$$

where $\beta_{W_i}^{\text{meds}}$ has a log-Gaussian prior. This likelihood has the added benefit that the total likelihood and its gradient can be computed in time proportional to the number of nonzero elements (Ranganath et al., 2015a). We overdisperse the Bernoulli likelihood by the number of medications to balance this component with the time from failure. The diagnoses are modeled in the same manner.

4. Experimental Setup

We apply deep survival analysis to data from a large metropolitan hospital. We use the fitted model to predict coronary heart disease risk (CHD). Access to data and experiments was approved after review by the Columbia University Institutional Review Board.

4.1 Dataset

Our dataset comprises the longitudinal records of 313,000 patients from the Columbia University Medical Center clinical data warehouse. The patient population included all adults (>18 years old) that have at least 5 months (not necessarily consecutive) where at least one observation was recorded.

The patient records contain documentation resulting from all settings, including inpatient, outpatient, and emergency department visits. Observations included 9 vital signs, 79 laboratory test measurements, 5,262 medication orders, and 13,153 diagnosis codes.

Data Preprocessing. All real-valued measurements and discrete variables were aggregated at the month level, leading to binned observations for each patient and for each month the patient had any recorded observation. The expected value over the course of each month was computed for continuous measurements such as vitals and laboratory measurements and the presence of discrete elements such as medication orders and diagnosis codes was encoded as a binary variable.

4.2 Baseline and Model Setup

Baseline. The Framingham CHD risk score was developed in 1998 and is one of the earliest validated clinical risk scores. It is a gender-stratified algorithm for estimating the 10-year coronary heart disease risk of an individual. Aside from gender, this score takes into consideration age, sex, LDL cholesterol, HDL cholesterol, blood pressure, diabetes, and smoking. For example, a 43-year-old (1 point) male patient with an LDL level of 170 mg/dl (1 point), an HDL level of 43 mg/dl (1 point), a blood pressure of 140/90 (2 points), and no history of diabetes (0 points) or smoking (0 points) would have a risk score of 5 which would correspond to a 10 year CHD risk of 9% (Wilson et al., 1998).

The score was validated using curated data from the Framingham Heart Study. It was shown to have good predictive power of 10-year risk with a concordance of 0.73 for men and 0.77 for women. However, this score has lower performance when applied to EHR data (Pike et al., 2016).

Model Setup and Hyperparameters. We set the shape of the Weibull to be 2. The exponential family used inside the DEF is a Gaussian. The mean and inverse softplus variance functions for each layer are a 2 layer perceptron with rectified linear activations. We set Normal priors to have mean zero and variance one.

We let all methods run for 6,000 iterations and assess convergence on a validation set. On a 40-core Xeon Server with 384 GB of RAM, 6,000 iterations for all patients in the training set completed in 7.5 hours. Due to the high variance in patient record lengths, we subsample observations during inference inversely to the length of the patient records.

Inference. We approximate the posterior distribution with variational inference (Jordan et al., 1999) for the observation specific latent variables in the DEF and do maximum-a-posteriori inference on the parameters. We choose the approximating family to be the mean-field family where each latent variables gets its own independent parameterization. We use black box variational methods (Ranganath et al., 2014) with reparameterization gradients (Kingma and Welling, 2014; Rezende et al., 2014) to approximate the posterior without needing model specific computation. To scale to the large data, we subsample data batches (Hoffman et al., 2013) of size 240 and parallelize computation across data in a batch. We use RMSProp with scale 0.0001 and Nesterov momentum of 0.9 as learning rates during optimization.

4.3 Evaluation

Of the 313,000 patients in the study, 263,000 were randomly selected for training, 25,000 for validation, and 25,000 for testing. We assessed convergence with the validation cohort and evaluated concordance on the test cohort. A CHD event was defined as the documentation of any ICD-9 diagnosis code with the following prefixes: 413 (angina pectoris), 410 (myocardial infarction), or 411 (coronary insufficiency). In our experiments, we vary the dimensionality of z_n to assume the values of $\mathrm{K} \in \{5, 10, 25, 75, 100\}$. The layer size for the perceptrons were set to equal the dimensionality of z_n in each experiment. We evaluate both the baseline risk score and deep survival analysis with the concordance (Harrell et al., 1982).

While concordance enables the comparison of deep survival analysis to the baseline, it only roughly captures the accuracy of the temporal prediction of the models. In deep survival analysis, we can compute the predictive likelihood of the held-out set according to the model, which enables us to capture how well the model predicts failure in time. For internal model validation, we thus rely on predictive likelihood. Predictive likelihood is evaluated as the expected log probability of the observed time until failure conditioned on the observed covariates for a given patient in a given month.

5. Results

Missing data is a core challenge of electronic health record data analysis and temporal analysis. We first report the extent of incomplete observations in our dataset. We then report results for the baseline CHD risk model and deep survival analysis. We also report internal validation of deep survival analysis which include only a single data type.

5.1 Missing observations in EHR data

For estimating CHD risk in 10 years, the widely used guideline-based CHD risk score calculators used routinely by clinicians require input of seven variables: age, sex, current smoking status, total cholesterol level, HDL cholesterol level, systolic blood pressure, and whether patient takes blood pressure medication. We examined how many patients had at least one month in their record, where the most basic, critical set of variables were observed (LDL level, HDL level, and blood pressure). In the full dataset, only 11.8% of patients have a complete month, and 1.4% of months are complete.

5.2 Model Performance and Predictive Likelihood

The baseline CHD risk score yielded 65.57% in concordance over the held out test set. Table 1 shows the concordance of the deep survival analysis for different values of K. When considering the full deep survival with all data types considered, the best performance was obtained for K=50.

Model	Concordance (%)
Baseline Framingham Risk Score	**65.57**
Deep Survival Analysis; K=10	69.35
Deep Survival Analysis; K=5	70.45
Deep Survival Analysis; K=25	71.20
Deep Survival Analysis; K=75	71.65
Deep Survival Analysis; K=100	72.71
Deep Survival Analysis; K=50	**73.11**

Table 1: Concordance on a held-out set of 25,000 patients for different values of K and for the baseline risk score. All deep survival analysis dimensionalities outperform the baseline.

When examining the deep survival analysis with the best concordance on the held out set (K=50), we then asked how well each individual data type predicts failure. The following four models were thus trained: deep survival analysis including vitals only, diagnosis codes only, laboratory tests only, and medications only. All models included age and gender. Their individual predictive likelihood was computed on the same month bins, even in the absence of observations of a specific data type. The diagnosis-only model yielded the best predictive likelihood.

Data Type	Likelihood
Medications Only	-1.24899
Laboratory Tests Only	-0.998774
Vitals Only	-0.961827
Diagnoses Only	-0.855385

Table 2: Predictive likelihood of deep survival analysis (K=50) for individual data types. The diagnoses perform best.

6. Discussion

In this paper we introduce a new method for survival analysis built to handle the inherent characteristics of EHR data. While traditional survival analysis requires carefully curated research datasets, our approach easily handles the sparsity and heterogeneity of EHR observations. We estimate deep survival analysis on the entire data from a large metropolitan hospital in a matter of hours. When compared to one of the state-of-the-art, clinically validated risk score in the context of coronary heart disease, deep survival analysis yields a more accurate stratification of patients. Our approach holds particular promise for developing risk scores from observational data for conditions where there is no known risk score.

Acknowledgments

This work is supported by NSF #1344668, NSF IIS-1247664, ONR N00014-11-1-0651, DARPA FA8750-14-2-0009, DARPA N66001-15-C-4032, Adobe, The Sloan Foundation, The Seibel Foundation, and The Porter Ogden Jacobus Fellowship

References

L Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, CA, 1986.

DR Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220, 1972.

BF Gage, AD Waterman, W Shannon, M Boechler, MW Rich, and MJ Radford. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA*, 285(22):2864–2870, 2001.

MH Gail, LA Brinton, DP Byar, DK Corle, SB Green, C Schairer, and JJ Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

Y Hagar, DJ Albers, R Pivovarov, H Chase, V Dukic, and N Elhadad. Survival analysis adapted for electronic health record data: experiments with chronic kidney disease. *Statistical Analysis and Data Mining*, 7(5):385–403, 2014.

GK Hansson. Inflammation, atherosclerosis, and coronary artery disease. *New England Journal of Medicine*, 352(16):1685–1695, 2005.

FE Harrell, RM Califf, DB Pryor, KL Lee, and RA Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

M Hauskrecht, I Batal, M Valko, S Visweswaran, GF Cooper, and G Clermont. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics*, 46(1): 47–55, 2013.

I Hjermann, I Holme, KV Byre, and P Leren. Effect of diet and smoking intervention on the incidence of coronary heart disease. *The Lancet*, 318(8259):1303–1310, 1981.

NL Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

M Hoffman, DM Blei, C Wang, and J Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1303–1347), 2013.

G Hripcsak and DJ Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *The Lancet*, 332(8607):349–360, 1988.

M Jordan, Z Ghahramani, T Jaakkola, and L Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

EL Kaplan and P Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

D Kingma and M Welling. Auto-encoding variational bayes. In *Proceedings ICLR (International Conference on Learning Representations)*, 2014.

P McCullagh. Survival models and health sequences. *arXiv preprint arXiv:1301.2699*, 2013.

B Neal, S MacMahon, N Chapman, J Cutler, R Fagard, P Whelton, S Yusuf, L Agodoa, C Baigent, H Black, JP Boissel, B Brenner, M Brown, C Bulpitt, et al. Effects of ace inhibitors, calcium antagonists, and other blood-pressure-lowering drugs: Results of prospectively designed overviews of randomised trials. *The Lancet*, 356(9246):1955–1964, 2000.

NJ Pagidipati and TA Gaziano. Estimating deaths from cardiovascular disease: A review of global methodologies of mortality measurement. *Circulation*, 127(6):749–756, 2013.

A Perotte, R Ranganath, JS Hirsch, DM Blei, and N Elhadad. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association*, 22(4):872–880, 2015.

MM Pike, PA Decker, NB Larson, JL St. Sauver, PY Takahashi, VL Roger, WA Rocca, VM Miller, JE Olson, J Pathak, and SJ Bielinski. Improvement in cardiovascular risk prediction with electronic health records. *Journal of Cardiovascular Translational Research*, 9(3):214–222, 2016.

R Pivovarov, DJ Albers, JL Sepulveda, and N Elhadad. Identifying and mitigating biases in ehr laboratory tests. *Journal of biomedical informatics*, 51:24–34, 2014.

R Ranganath, S Gerrish, and DM Blei. Black box variational inference. In *Proceedings AISTATS (International Conference on Artificial Intelligence and Statistics)*, pages 814–822, 2014.

R Ranganath, A Perotte, N Elhadad, and DM Blei. The survival filter: Joint survival analysis with a latent time series. In *Proceedings UAI (Uncertainty in Artificial Intelligence)*, 2015a.

R Ranganath, L Tang, L Charlin, and DM Blei. Deep exponential families. In *Proceedgins AISTATS (Internation Conference on Artificial Intelligence and Statistics)*, 2015b.

DJ Rezende, S Mohamed, and D Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings ICML (International Conference on Machine Learning)*, 2014.

FM Sacks, MA Pfeffer, LA Moye, JL Rouleau, JD Rutherford, TG Cole, LW Brown, JW Warnica, JM Arnold, CC Wun, BR Davis, and E Braunwald. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *New England Journal of Medicine*, 335(14):1001–1009, 1996.

Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the scandinavian simvastatin survival study (4s). *The Lancet*, 344(8934):1383–1389, 1994. ISSN 0140-6736.

G Schuler, R Hambrecht, G Schlierf, J Niebauer, K Hauer, J Neumann, E Hoberg, A Drinkmann, F Bacher, and M Grunze. Regular physical exercise and low-fat diet. effects on progression of coronary artery disease. *Circulation*, 86(1):1–11, 1992.

J Shepherd, SM Cobbe, I Ford, CG Isles, AR Lorimer, PW Macfarlane, JH McKillop, and CJ Packard. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *New England Journal of Medicine*, 333(20):1301–1308, 1995.

IM Thompson, PJ Goodman, CM Tangen, HL Parnes, LM Minasian, PA Godley, MS Lucia, and LG Ford. Long-term survival of participants in the prostate cancer prevention trial. *New England Journal of Medicine*, 369(7):603–610, 2013.

PWF Wilson, RB DAgostino, D Levy, AM Belanger, H Silbershatz, and WB Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18): 1837–1847, 1998.