Forensic-Style Analysis with Survival Trajectories

Pranjul Yadav*, Michael Steinbach*, Lisiane Pruinelli† , Bonnie Westra †,
Connie Delaney †, Vipin Kumar *, Gyorgy Simon ‡

*Department of Computer Science and Engineering, University of Minnesota. {yadav023, stei0062, kumar001}@umn.edu
† School of Nursing, University of Minnesota. {pruin001,westr006, delaney} @umn.edu
‡Mayo Clinic, Minnesota. {Simon.Gyorgy} @mayo.edu

***Abstract*—Electronic Health Records (EHRs) consists of patient information such as demographics, medications, laboratory test results, diagnosis codes and procedures. Mining EHRs could lead to improvement in patient healthcare management as EHRs contain detailed information related to disease prognosis for large patient populations. We hypothesize that a patient's condition does not deteriorate at random; the trajectories, sequences in which diseases appear in a patient, are determined by a finite number of underlying disease mechanisms. In this work, we exploit this idea by predicting a patient's risk of mortality in the context of the metabolic syndrome by assessing which of many available trajectories a patient is following and progression along this trajectory. Implementing this idea required innovative enhancements both for the study design and also for the fitting algorithm. We propose a forensic-style study design, which aligns patients on last follow-up and measures time backwards. We modify the time-dependent covariate Cox proportional hazards model to better capture coefficients of covariate that follow a particular temporal sequence, such as trajectories. Knowledge extracted from such analysis can lead to personalized treatments, thereby forming the basis for future trajectory-centered guidelines.**

I. Introduction

The use of large repositories of Electronic Health Records (EHR) data for assessing the risk of adverse outcomes, such as mortality or the development of new complications, is experiencing a rapid growth in popularity. The most common style of analysis for this purpose is based on longitudinal retrospective design, where patients are aligned on a particular point in time (e.g. enrollment into the study), called a *baseline*, their state of health at baseline is characterized by elements present in EHR data (*baseline characteristics*) and they are followed until *last follow-up*, at which point they suffer the adverse outcome in questions or are simply lost to follow-up (are *censored*).

Such studies have enjoyed great success. Strong epidemiological evidence has been discovered, which ultimately influenced health care policy. However, the acceptance and incorporation of these methods into clinical decision support systems is slow. The design underlying this methodology, where a patient's risk is solely based on baseline characteristics, is incompatible with clinical practice. Providers constantly reevaluate a patient's risks and adjust treatment accordingly. When the patient information shows no clear sign of improvement or deterioration, a common approach is to wait and see. As time progresses and the patient's condition further deteriorates, the outcome becomes more apparent and an appropriate intervention can be administered. When the patient's health has deteriorated to the final stages, the outcome can become obvious and also inevitable: there may be no time for a successful intervention. Knowing not the only the risk but also the expected timing of adverse events is important, allowing the care provider to have time to intervene. In this study, we look at a large diabetic population and aim to mimic the clinical process. We assess the a patient's risk at every encounter, taking not only the prior conditions but also their sequence into account.

Our working hypothesis is that a patient's health deteriorates following a (small or large) number of non-random mechanisms. These different mechanisms may affect organs or health indicators (blood sugar, lipid levels, blood pressure) differently, leading to different sequences of diagnoses. Therefore, the order in which the diagnoses appear in a patient's record can be suggestive of the underlying disease mechanism, allowing us to provide the patient a better prognosis.

We make the following contributions:

1) We propose modeling a patient's risk of adverse outcome based on the trajectories they follow and the extent to which they have progressed along these trajectories; thus allowing us to take the sequence of events into account.
2) We introduce the *forensic-style* analysis, which aligns patients on last follow-up and measures time backwards. Measuring time backwards allows us to estimate the time-to-event more directly.
3) We modified the Cox proportional hazard model, using forensic-style analysis, to better model time-dependent covariates. Specifically, we modified how the outcome is designated, allowing the fitting algorithm to better estimate the risk of diseases in earlier stages of the trajectories.

The paper is organized as follows. Section II describes current state of the art techniques used to handle time-to-event data. Section III-A introduces terminology associated with trajectories. Section III-B discusses techniques for extracting frequent trajectories. In Section III-C we present our model and optimization framework. In Section IV, we discuss our results. Finally, Section V presents our conclusions.

II. Related Work

Survival modeling techniques on time-to-event data have been explored widely in the past. Cox regression [1] is one of the most commonly used survival regression models. Standard regularization techniques, developed for other regression methods, have been applied to Cox models, as well. Lasso [2] and elastic-net regularized Cox models [3] have been developed, and have been further extended by regularizing them with convex combinations of L1 and L2 penalties [4]. We are not aware of regularization for time-dependent covariate Cox models [5], which would be a straightforward extension. Chandan et. al [6] proposed an active learning based survival model which uses a novel discriminative gradient based sampling scheme and observed better sampling rates as compared to other sampling strategies. They also proposed correlation based regularizers with Cox regression to handle correlated and grouped features which are commonly seen in many practical problems [7].

III. Methods

We consider seven **disease**s in the context of diabetes. These are hyperlipidemia (HL; high cholesterol), hypertension (HTN; high blood pressure), type-II diabetes mellitus (DM), chronic kidney disease (CKD), ischemic heart disease (IHD), cerebrovascular disease (CVD), and congestive heart failure (CHF). These are chronic diseases; once the presence of the disease has been confirmed, they remain active. The patient may have the condition under control (i.e. a patient can have normal laboratory results), but the disease remains.

A. Trajectory Terminology

Not all mentions of these diseases in the patient's record indicate a new diagnosis. Often, these diagnoses are present for billing purposes, as they complicate treatment. To determine the precedence of the diseases in the trajectories, we need to focus on new (**incident**) diagnoses. The term 'incident' refers to the diagnosis creating a new incidence, as opposed to being a chronic condition in the background that complicates the treatment of a different disease. To identify incident diagnoses, we need to determine the **status** of diseases at any time point.

The disease is **confirmed** if we have evidence that the patient presents with the disease; it can be **ruled out** if we have evidence that the patient does not have the disease; or the status can be unknown otherwise (when we do not have evidence either way).

Definition 1 (Disease confirmed): A disease is confirmed at time t and thereafter, if the patient's record has a diagnosis code, a prescribed medication or an abnormal lab result (if applicable) related to the disease.

Definition 2 (Disease ruled out): A disease is ruled out at time t and before, if no pertinent medication prescription or diagnosis code is present at or before t and a normal laboratory result is present at t.

Definition 3 (Incident diagnosis): A new disease diagnosis is incident at t if the disease is ruled out before t and confirmed after t.

In plain language, a disease diagnosis is new (or an incident diagnosis) if we have evidence that it is new: it was absent before t and is present at t. For IHD, CVD and CHF, we do not have laboratory results to rule them out, so we assume that the first diagnosis of these diseases in our data is an incident diagnosis.

Definition 4 (Background disease): Non-incident diagnosis of a confirmed disease.

A background disease is a confirmed as a preexistent condition or a potentially preexisting condition that we cannot rule out. If a patient enters the study with (say) HTN, then HTN is a background disease (confirmed preexisting condition). If the first appearance of HL (high cholesterol) is a year after enrollment, but the patient does not have cholesterol measurements before the diagnosis, then HL is background (the patient may have had HL all along). If, however, we have a normal cholesterol measurement before the HL diagnosis, then the HL is incident, because we rule it out for (some part of) the first year.

Definition 5 (Precedence): A disease A precedes disease B, $A \to B$, (or B follows A) in a patient, if the patient has an *incident* disease B at time t and A is a background or incident disease before t.

Since B is an *incident* disease, we ruled it out before t, while A could not be ruled out before t, thus A occurred earlier than B.

Definition 6 (Trajectory): A trajectory is a set of diseases, some incident, some background, with precedence information among them. In other words, a trajectory is a partially temporally ordered set of diseases.

Example. $T = (A, B) \to C \to D$ is a trajectory with A and B being background diseases, whose ordering cannot be determined from our data and C and D are incident diseases, hence their ordering is known. The precedence information is transitive, so beside the depicted $A \to C$, $B \to C$ and $C \to D$ precedence relationships, $A \to D$ and $B \to D$ also hold. These diseases are chronic, hence at the time when the patient develops C, he also has A and B; and at the time he develops D, he also has A, B and C.

The central idea in our work is to place patients on trajectories, which requires that we define when a trajectory **applies** to a patient or **matches** a patient's trajectory.

Definition 7 (Sub-trajectory): A trajectory S is a subtrajectory of T if the diseases in S are a subset of the diseases in T and all precedence information in T that relates to the diseases in S holds true in S.

Example. The trajectory $S = B \to C$ is a subtrajectory of T, as it contains a subset of the diseases B and C and all precedence relationships involving B or C in T, namely $B \to C$, also holds true in S.

Definition 8 (Prefix trajectory): A trajectory P is a prefix trajectory of T if P is a subjectory of T and no disease in T precedes any of the diseases in P.

Example. $S = (A, B) \to C$ is a prefix trajectory of T as none of the diseases in T precede the diseases in S. In contrast, $B \to C$ is not a prefix of T as there is a disease A in T that

precedes C in T.

Definition 9 (Matching): A trajectory T applies to a patient with trajectory X (or matches X) iff (i) there exists a prefix P of T that is a subtrajectory of X and (ii) there exists no disease d in X, such that d is not a part of P but is present in T.

Example. Consider a patient trajectory $X = A \to B \to C \to D$. The trajectory $T = A \to B \to E$ matches X with prefix $P = A \to B$, because the only disease in T that is not in P (namely E) is not in X. The clinical motivation behind this definition is that the patient with trajectory X may be following T, just has not progressed to E yet. (He may also follow other trajectories that explain C and D).

B. Algorithm for Extracting the Frequent Trajectories

Our goal is enumerate all trajectories that patients frequently follow that end in mortality. Therefore, for trajectory extraction, we only consider patients who died and do not consider patients who remained alive after last follow-up. We apply the venerable Apriori algorithm [8] to enumerate all sub-trajectories that occur in at least 4 patients. With 2814 total deaths, the support of 4 (support fraction of 4/2814) is the smallest support fraction that does not contain 0 in its 95% confidence interval.

Output. The output of the algorithm is a library (set) of trajectories, which we call **library trajectories** that end in mortality and occur frequently in patients who suffered mortality. Some of these trajectories can be sub-trajectories of each other.

C. Estimating the Relative Risk of the Trajectories

Given a potentially large and redundant set of library trajectories, discovered above, our goal in this section is to a develop a methodology for (i) selecting trajectories and (ii) to estimate the risk of mortality in patients, who may follow zero, one or more of the library trajectories.

C.1. Data Format.

The trajectories are transformed into a binary *design matrix* X. The columns of X correspond to diseases along the trajectories: each disease along each trajectory is mapped to its own column. The rows of X correspond to patients during different time periods, active periods. Therefore, for the ith record, we have the associated trajectory information x_i (ith row in X), the beginning b_i and end e_i time of the active period, the patient id p_i and the outcome y_i. The indicator $A_i(t)$ signals whether record i is active at time t; it returns 1 for $b_i \geq t > e_i$. Note that in our forensic-style analysis, time is measured backwards, so $b_i > e_i$.

For each patient, the active time periods are defined by changes in the trajectory: whenever the patient develops a new disease which corresponds to progression along a trajectory, we add a new record with the appropriate timing information. Therefore each record represents a new state, where the patient has progressed further (has accumulated more diseases).

The outcome y_i is 1 if the patient p_i had an adverse outcome exactly b_i time after the beginning of the record. In contrast to Cox models with varying covariates, the outcome is 1 for all records of the patient. While it may appear that the patient had died multiple times, our definition of b_i (being measured from death) ensures that these "multiple" deaths coincide at the right time point. The baseline hazard can compensate for the multiplicity of deaths.

C.2. Model.

The model is a variant of the Cox Proportional Hazards Regression model. Central to the model is the concept of **hazard**, which we define analogously to the Cox terminology, namely, as the instantaneous probability of death in exactly t time from an event.

$$\lambda_0(t) \exp(x_i \beta) \tag{1}$$

where $\lambda_0(t)$ is a time-dependent baseline hazard that is common across all patients and the trajectories x_i increase the hazard proportionally.

Given our design matrix X described earlier, the expression $x_i\beta$ expands into

$$x_i \beta = \sum_{L \in \mathcal{L}(b_i)} \sum_{d \in L(b_i)} \beta_{L,d} \tag{2}$$

where $\mathcal{L}(b_i)$ is the set of library trajectories that apply to the patient p_i at time b_i, the diseases d are the diseases confirmed for the patient at or before time b_i along the trajectory L, and $\beta_{L,d}$ are the coefficients. The sum $\sum_{d \in L} \beta_{L,d}$ is the (log) relative risk that having reached d along trajectory L confers on the patient. Notice that the (log) relative risk along a trajectory cumulates in a (log-)additive fashion, indicating that each events along the trajectory also confers a proportional hazard.

We can estimate the "probability" of death (technically, expected count of deaths) for patient p at time t as

$$\begin{gathered} \Lambda_p(t) = \sum_{\tau=0}^{t} \lambda_0(\tau) \exp(x_i \beta), \\ \text{for } i: \ A_i(\tau) = 1, p_i = p \end{gathered} \tag{3}$$

for all records i where the record is active at time τ and describes patient p.

C.3. Likelihood.

The likelihood is the probability that for each patient p after developing each disease d, the outcome happens exactly time t after developing the disease.

$$\begin{gathered} \prod_i \left[\frac{\lambda_0(t) \exp(x_i \beta)}{\sum_j A_j(t) \lambda_0(t) \exp(x_j \beta)} \right]^{y_i} \\ \text{for } i: \ t = b_i, \ j: \ b_j \geq t > e_j \end{gathered} \tag{4}$$

Defining the vector of linear risk score u as $u = x\beta$, the log likelihood becomes

$$\ell(u) = \sum_i y_i \left[u_i - \log \sum_j A_j(b_i) \exp u_j \right] \tag{5}$$

C.4. Optimization

Our goal with the optimization is (i) select a subset of the library trajectories for modeling and (ii) estimate their coefficients. We optimize β iteratively through a gradient boosting framework [9], adding a new trajectory in each iteration. Adding a library trajectory, say L, is equivalent to changing the corresponding set of coefficients in β, which we denote by β_L.

Performing boosting (gradient ascent in u-space), leads to the update

$$u^{(k+1)} = u^{(k)} + \gamma \frac{d\ell}{du^{(k)}}, \tag{6}$$

where γ is the learning rate, $u^{(k)}$ is the linear risk score u in the kth iteration and ℓ is the log likelihood function.

In iteration k, we need to find the trajectory that fits $d\ell/du^{(k)}$ the best. Let $\nabla\ell$ denote the gradient $d\ell/du^{(k)}$. We wish to find the trajectory L, with coefficient vector β_L, such that the quantity

$$\min \beta_L \quad (\nabla\ell - x_L\beta_L)'(\nabla\ell - x_L\beta_L)$$

is minimal across all trajectories. The prime sign (') denotes matrix (vector) transposition. Once we find the optimal trajectory along with the optimal β_L, we can update the β vector. The learning rate γ can be determined through line-search or can also be chosen as an arbitrary small number.

Stopping criterion. We stop adding trajectories, when the improvement of ℓ on either the training or a validation set is less then a pre-defined small positive number ε.

Initialization. We can either start with an empty set of trajectories, or we can provide a pre-selected set of trajectories resulting from a greedy coverage of the events in the patient trajectories. For our experiments, we started with an empty set.

Gradient. To derive $\nabla\ell$, we first separate out a particular component u_k from ℓ and then derive the partial derivative with respect to u_k.

$$\begin{aligned}
\ell &= \sum_i \{y_i u_i - y_i \log [A_j(b_i) \exp u_j]\} \\
&= \left\{ y_k u_k - y_k \log \left[A_k(b_k) \exp u_k + \sum_{j \neq k} A_j(b_i) \exp u_j \right] \right\} \\
&+ \sum_{i \neq k} \left\{ y_i u_i - y_i \log \left[A_k(b_k) \exp u_k + \sum_j A_j(b_i) \exp u_j \right] \right\}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial u_k} &= y_k - y_k \frac{A_k(b_k) \exp u_k}{\sum_j A_j(b_k) \exp u_j} \qquad (7) \\
&\quad - \sum_{i \neq k} y_i \frac{A_k(b_i) \exp u_k}{\sum_j A_j(b_i) \exp u_j} \\
&= y_k - \sum_i y_i \frac{A_k(b_i) \exp u_k}{\sum_j A_j(b_i) \exp u_j} \qquad (8)
\end{aligned}$$

The sum iterates over all records i that began during the active time of record k divided by the summed risk of records j that started when i was active. Thus the partial derivative can be restated in a more familiar form

$$\begin{aligned}
\frac{\partial \ell}{\partial u_k} &= y_k - \sum_{\tau=b_k}^{e_k} \left(\sum_i \frac{y_i}{\sum_j A_j(\tau) \exp u_j} \right) \exp u_k, \quad \text{for } i: \ b_i = \tau. \qquad (9) \\
&= y_k - \sum_{\tau=b_k}^{e_k} \lambda_0(\tau) \exp u_k \qquad (10)
\end{aligned}$$

Unlike in the regular Cox models, the gradient is not the residual, only a part of the residual that the corresponding record is responsible for. For gradient boosting, it is not required that the gradient coincides with the residual. The form of the partial derivative, however, suggests a form for the cumulative hazard that parallels the Breslow estimate [10] in Cox models, which we presented in Eq. 3.

IV. Evaluation and Results

We use the clinical data repository of a large health care system situated in the Midwestern United States. Based on data availability, we selected 2005 to 2014 as the study period. We included all adult patients who developed type-II diabetes during this period. Mortality data from the state death registry was available for 8,000 of these patients. Our health care system has a large tertiary care arm, thus many of the patients may receive their primary care (and possibly diabetes care) outside this system leading to large gaps in the data. To exclude such patients, we required the study population to have at least 2 Hemoglobin A1c measurements at least 1 year apart. The final cohort consists of 2,814 *cases* (patients who died) and almost 2,000 *controls* (who were censored). For these patients, we collected diagnoses, lab values, vitals and medication data.

Experiment Designs. Our problem is not a traditional computer science problem hence methods to compare it against are very few, and mostly in biostatistics and epidemiology. We decided to evaluate our algorithm by showing that all innovations we claim improve the performance. Accordingly, we build four models, starting from the simplest model (the one that is typically used to solve this problem) and successively adding our proposed features to it to isolate the effect of each of our contributions.

(1) Enrollment-Aligned Design. The typical approach to time-to-event problems is to conduct a retrospective study, where patients are aligned on their enrollment into the study and are followed until mortality or until they get lost to follow-up (until censoring). The modeling method is Cox proportional hazards model with time-dependent covariates [5]. This is the simplest and most common model to solve our problem.

(2) Outcome-Aligned Design. The next simplest model aligns patients on outcome, which admittedly, is an unusual but reasonable design. Suppose patient i has follow-up T_i. We

select a time T, which is larger than all T_i's and designate T as the last follow-up for all patients. Consequently, in this design, we align patients on their last follow-up (which happens at time T for all patients by design). Their enrollment time into the study will vary, it will be $T - T_i$ for patient i. This design assumes that the baseline hazard depends on time from death. This stands in sharp contrast with the assumption of the Enrollment-Aligned Design, which assumes that the baseline hazard depends on time from enrollment. Therefore, the Outcome-Aligned Design incorporates exactly one aspect of the proposed forensic-style analysis: alignment on outcome (i.e. alignment on the last follow-up).

(3) Forensic-Style Design. This is the design proposed in this manuscript. Forensic-Style Design is similar to Outcome-Aligned Design in that patients are aligned on last follow-up, but it goes beyond by measuring time backwards. Measuring time backwards allows us to designate *all* records of cases (patients who died at last follow-up) as positive and still retain the correct time of death across all records. Since the likelihood has a different meaning in this design, we use our own fitting algorithm from Section III-C: we use seven "trajectories" each consisting of a single disease, thus our predictors are the diseases. We will refer to this model as 'fast w/o traj' (FAST without trajectories).

(4) Forensic-Style Analysis Via Survival Trajectories (FAST). The experiment is designed using the Forensic-Style Design, but instead of the seven diseases, we use trajectories as predictors. This is precisely the proposed methodology.

Evaluation Method

Given the time-to-event outcome, our evaluation metric is **survival concordance**. This is a widely used metric for time-to-event data. For any two patients, i and j, i having a higher risk of death than j, survival concordance measures the probability that i dies earlier than j. Patient pairs, for which it is not possible to determine whether the higher risk patient dies earlier (e.g. he is still alive at last follow-up), are ignored. Ties (patient pairs with the same risk and same time-to-death) are also ignored. Note that i and j are different patients: two records of the same patient are not compared.

A. Results

In Figure 1, we present the survival concordance of the four models across the 100 bootstrap replications.

Effect of Aligning Patients on Outcome. The 'enrollment' and 'outcome' models use the same fitting algorithm (time-dependent Cox model), the same predictors (the seven diseases) and only differ in the study design: in the 'enrollment' model, patients are aligned on their enrollment into the study, while in the 'outcome' model, they are aligned on their last follow-up. The benefit of aligning patients on last follow-up is clear. In the Enrollment-Aligned Design, time represents time-since-enrollment, which is not associated with death. On the other hand, in case of the Outcome-Aligned Design, time is related to the time of death: time of death happens exactly at the same time point for all patients (by definition). Aligning patients on their time of death (or censoring) allows for more accurate description of the so-called *risk set*: the patients who were under observation at the time, having the potential for an event. (This is the denominator in the likelihood function.)

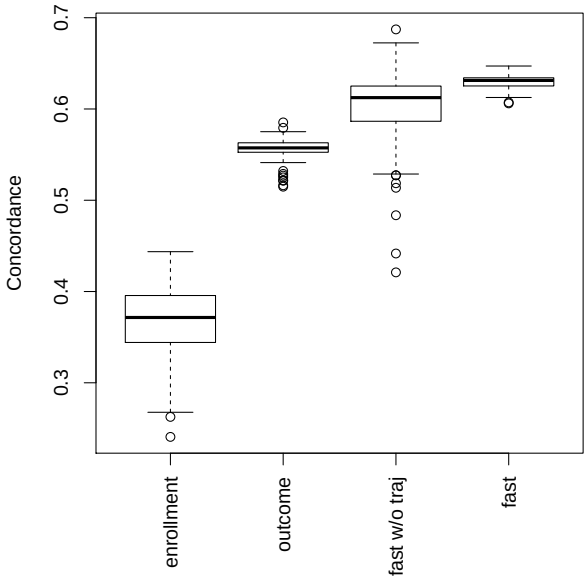

Fig. 1. Concordance of the various designs estimated through bootstrapping.

Effect of Outcome Designation. To assess the effect of the outcome designation, we can compare the 'outcome' model with the 'fast w/o traj'. Both of these models utilize a study design that aligns patients on the outcome; the difference between them lies in measuring time backwards and the outcome designation this change enables. The beneficial effect of this difference is very significant as observed from the paired t-test performed between survival concordance values of the two methods (p-value 1e-16). Our choice of outcome designation was motivated by the following observation. Given a trajectory $a \rightarrow b \rightarrow c$ that ends in death, in the typical study design (Enrollment-Aligned or even Outcome-Aligned), when the patient only has a, or has a and b, his outcome is still designated as 'alive'. Since death rarely follows a or b without c, this designation leads the fitting algorithm to believe that a and b are protective. The result is negative coefficients for these diseases and a survival concordance less than 0.5 (see the 'enrollment' model).

Effect of using trajectories. Finally, 'fast w/o trajectories' and 'fast' differ only in the use of trajectories. The use of trajectories is advantageous (p-value 7.5e-6). When trajectories are not utilized, the model only has seven predictors. Estimating seven coefficients from 12k records contributed by 5k patients is trivial. On the other hand, it is suspected that these seven conditions affect the risk of mortality differently depending on the presence of other conditions.

B. In-Depth Look at the FAST Results

Above, we have shown that the performance of FAST is substantially and (statistically) significantly better than any other model we have considered. In this section, we are going to show some of the resultant models.

covariate	enrollment	outcome	fast w/o traj
HL	–	–	0.00
HTN	–	–	0.00
DM	14.56	16.24	8.21
CKD	-0.45	-0.21	-0.07
IHD	0.42	0.05	0.00
CVD	-0.28	-0.37	-0.01
CHF	-0.24	0.10	0.00

TABLE I
COEFFICIENTS OF THE NON-TRAJECTORY BASED MODELS

Table I presents the coefficients of the three models that do not rely on trajectories. These are coefficients obtained from regular Cox models and thus their interpretation is as follows. For example, the relative risk of mortality that CHF (congestive heart failure) confers on a patient is $\exp(-.24) = .79$; a patient with CHF is 21% less likely to die than the average patient in our cohort. The coefficients of HL and HTN are 0 or NA because these diseases occur in nearly all patients. As a result, their risk is not reasonably estimable. ('fast w/o traj' did not select these variables, either.)

FAST Models.

In this section, we turn our attention to the FAST model. To assess the statistical significance of the coefficients, we ran 500 bootstrap replications, resulting in 500 models, each potentially using a different set of trajectories.

Most of the 500 models used only one (392 models) or two trajectories (34 models) and on the other extreme, there were models using 20, 22, and 28 trajectories (one model each). In Table II, we present some of the frequently selected trajectories.

	HL	HTN	DM	CKD	IHD	CVD	CHF
$DM, HL, HTN \rightarrow CHF$ (263)							
coefs	-0.23	-0.31	0.13	–	–	–	1.21
p-val	0.04	0.00	0.06	–	–	–	0.00
$DM, HL, HTN \rightarrow CKD$ (192)							
coefs	-0.25	-0.30	0.12	1.09	–	–	–
p-val	0.03	0.01	0.04	0.00	–	–	–
$DM, HL \rightarrow HTN \rightarrow CKD$ (41)							
coefs	-0.10	-0.11	-0.04	0.76	–	–	–
p-val	0.02	0.02	0.27	0.00	–	–	–
$HTN \rightarrow HL \rightarrow DM$ (10)							
coefs	0.09	0.13	0.20	–	–	–	–
p-val	0.00	0.00	0.00	–	–	–	–

TABLE II
TRAJECTORIES AND THEIR COEFFICIENTS THAT WERE UTILIZED IN AT LEAST 10 MODELS

The table presents the trajectory, followed by the number of models that utilized this trajectory in parenthesis. We then present the average coefficients (across the models that utilized this trajectory) of the diseases along the trajectory and also the empirical p-value of the coefficient, which is the fraction of bootstrap iterations in which the sign of the coefficient in question was the opposite of the sign of the mean.

V. SUMMARY AND CONCLUSION

In this manuscript we presented Forensic-style Analysis based on Survival Trajectories (FAST). FAST makes two key contributions: it places patients onto disease trajectories to assess their risk of progression to an adverse outcome (mortality in our study) and it performs a forensic-style analysis, where patients are aligned on their last follow-up and time is measured backwards. Measuring time backwards allows a third ancillary contribution: we can designate the outcome as positive for all records of cases (patients who ultimately died), potentially leading to better estimates of the effects of diseases that occur early in the progression. To isolate the effect of our innovations, we successively enhanced the baseline method (which we referred to as Enrollment-Aligned Design) by adding our contributions one at a time. We have thus demonstrated the benefit of aligning patients on outcome when we believe that time-to-death is important; we demonstrated the benefit of our outcome designation and we have also isolated the beneficial effect of using trajectories.

VI. ACKNOWLEDGEMENTS

This study is supported by National Science Foundation (NSF) grant: IIS-1344135. Contents of this document are the sole responsibility of the authors and do not necessarily represent official views of the NSF. This was partially supported by Grant Number 1UL1RR033183 from the National Center for Research Resources (NCRR) of the National Institutes of Health (NIH) to the University of Minnesota Clinical and Translational Science Institute (CTSI).

REFERENCES

[1] David R Cox. Regression models and life-tables. In *Breakthroughs in Statistics*, pages 527–541. Springer, 1992.

[2] Robert Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[3] Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshirani, et al. Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1–13, 2011.

[4] Hao Helen Zhang and Wenbin Lu. Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

[5] Terry Therneau and Cindy Crowson. Using time dependent covariates and time dependent coefficients in the cox model. *Red*, 2:1, 2014.

[6] Bhanukiran Vinzamuri, Yan Li, and Chandan K Reddy. Active learning based survival regression for censored data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 241–250. ACM, 2014.

[7] Bhanukiran Vinzamuri and Chandan K Reddy. Cox regression with correlation based regularization for electronic health records. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 757–766. IEEE, 2013.

[8] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[9] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[10] JA Anderson and A Senthilselvan. Smooth estimates for the hazard function. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 322–327, 1980.