A Robust Framework for Accelerated Outcome-driven Risk Factor Identification from EHR

Prithwish Chakraborty
IBM
Yorktown Heights, New York, USA
prithwish.chakraborty@ibm.com

Faisal Farooq
IBM
Yorktown Heights, New York, USA
ffarooq@us.ibm.com

ABSTRACT

Electronic Health Records (EHR) containing longitudinal information about millions of patient lives are increasingly being utilized by organizations across the healthcare spectrum. Studies on EHR data have enabled real world applications like understanding of disease progression, outcomes analysis, and comparative effectiveness research. However, often every study is independently commissioned, data is gathered by surveys or specifically purchased per study by a long and often painful process. This is followed by an arduous repetitive cycle of analysis, model building, and generation of insights. This process can take anywhere between 1 – 3 years. In this paper, we present a robust end-to-end machine learning based SaaS system to perform analysis on a very large EHR dataset. The framework consists of a proprietary EHR datamart spanning ∼55 million patient lives in USA and over ∼20 billion data points. To the best of our knowledge, this framework is the largest in the industry to analyze medical records at this scale, with such efficacy and ease. We developed an end-to-end ML framework with carefully chosen components to support EHR analysis at scale and suitable for further downstream clinical analysis. Specifically, it consists of a ridge regularized Survival Support Vector Machine (SSVM) with a clinical kernel, coupled with Chi-square distance-based feature selection, to uncover relevant risk factors by exploiting the weak correlations in EHR. Our results on multiple real use cases indicate that the framework identifies relevant factors effectively without expert supervision. The framework is stable, generalizable over outcomes, and also found to contribute to better out-of-bound prediction over known expert features. Importantly, the ML methodologies used are interpretable which is critical for acceptance of our system in the targeted user base. With the system being operational, all of these studies were completed within a time frame of 3 – 4 weeks compared to the industry standard 12 – 36 months. As such our system can accelerate analysis and discovery, result in better ROI due to reduced investments as well as quicker turn around of studies.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Applied computing → Health informatics**; *Life and medical sciences*;

KEYWORDS

health informatics; electronic health records; risk analysis

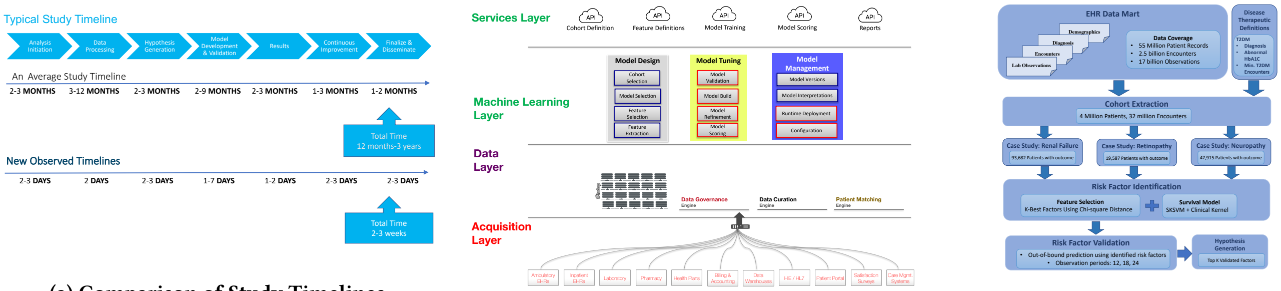
ACM Reference Format:
Prithwish Chakraborty and Faisal Farooq. 2019. A Robust Framework for Accelerated Outcome-driven Risk Factor Identification from EHR . In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330718

1 INTRODUCTION

With the growing prevalence of Electronic Health Record (EHR) systems, large volumes of clinical data are now digitized as part of routine patient care [1, 2, 8, 19]. Real world data (RWD) produced as an outcome of daily clinical situations (contrasted with clinical trials data created in a controlled study), is being increasingly used across healthcare domain such as health economics outcomes research (HEOR), comparative effectiveness of treatments, treatment recommendations, disease progression modeling, and risk analysis and prediction. However, most of these studies are commissioned independently and performed in a decentralized fashion leading to very long timelines. Our primary focus in this paper is on constructing a real-world applicable ML pipeline that can analyze such massive EHR data at scale and rapidly accelerate the required timeline. Figure 1a describes a typical study time line and how we managed to accelerate these steps. In this paper, we focus on a class of problem related to analyzing EHR data for identification of hitherto unknown and weakly correlated risk factors for unwanted outcomes of diseases. These are of prime importance to pharmaceutical companies, medical care providers, and in general the overall health system of countries. Early warnings generated from such factors can lead to effective interventions, lower the lifetime costs associated with diseases, and more importantly, improve quality of life [9, 11, 12]. This can also provide recommendation for treatments and guidance for further research and discovery of new treatments within specific patient sub-populations.

Most studies on EHR datasets have focused on factors identified by experts [7]. There has been some work on combining expert guidance with data-driven factors to create stable disease models [15] which shows great promise but are not entirely suited to massive EHR datasets due to dimensional complexities. Deep learning methods [14] currently lack the interpretability desired for clinical hypothesis generation. EHR datasets are inherently noisy and are typically characterized by missing information about single encounters or missing entire encounters. Disease models based on EHR thus need to be robust and stable with respect to the data quality. Traditionally, risk factors have been analyzed via survival models that can capture the temporal risk associated with extreme

KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330718



(a) Comparison of Study Timelines

(b) SaaS Architecture

(c) Risk Analysis framework

Figure 1: Proposed End-to-end ML driven SaaS Framework to analyze EHR dataset for risk factors associated with disease outcomes: (a) shows the acceleration of a typical study using our framework over the traditional timeline, (b) provides an overview of the different components of the SaaS framework, and (c) illustrates the machine learning module that shows the flow of data and the modeling steps leading up to final hypothesis generation.

outcomes and compare factors with respect to such risks. Its worth mentioning that modeling the outcome as a classification problem and comparing the factors w.r.t. the classification accuracy are *not valid* as such methods fail to capture the competing temporal risks and hence unsuitable towards usage such as interventions. Typically, health information covers both categorical and continuous factors and Cox Proportional Hazard models (CoxPH) are popular in traditional studies providing interpretability and simplicity via its linear formulation [3]. Extension to classic models have been proposed that can handle several hundreds of factors based on sparse regularization [10, 17]. Non-linear survival models based on Survival Support Vector Machines (SSVM) are also increasingly gaining popularity due to their robustness and efficacy [13, 18].

In this paper, we present an end-to-end machine learning Software as a Service (SaaS) framework to identify risk factors for unwanted outcomes that supports our goal of accelerated timeline. We accomplish this by relaxing the limitations of classical approaches that require practitioners to limit the scope via focusing on expert factors and/or the population coverage. The framework aims to provide guidance to medical researchers and data scientists while formulating downstream clinical studies via hypothesis generation. The framework is robust, stable and generalizable to allow for repeated studies at scale. It also inherently supports interpretability.

We applied our framework on proprietary EHR dataset covering ∼55 million individual patients and over ∼20 billion data points collected from participating hospitals and care networks. Since the use cases are mostly retrospective, albeit with accelerated time lines, the data is refreshed only once a week. The overall system architecture is shown in Figure 1b. There are multiple components in the overall system for record linkage, statistical de-identification and normalization. However, we will focus on the framework beyond the data layer in this work.

Figure 1c illustrates the basic components and the data flow within the framework. Broadly, the framework starts by readily and interactively being able to generate a subset (cohort) of patients and associated longitudinal characteristics (including temporal progression of diseases) based on the disease outcome of interest. With simple configuration, cohorts can be frozen in time or be allowed to refresh with the weekly updates. Downstream data analysis including risk modeling can be performed on the selected cohort to describe the probability of important outcomes and the factors that best capture the risk associated with the same.

For example, our proposed framework identifies relevant factors by robust temporal risk modeling of outcomes using a regularized Survival Support Vector Machine with a clinical kernel (SKSVM) [4, 13]. We also aim to identify the most relevant factors affecting the outcomes without sacrificing on stability and interpretability.

We compare the proposed ML pipeline against several baseline models with respect to efficacy (concordance), stability and generalizability.

The main quantifiable contributions of this paper are as follows:

- Accelerating study timelines by completing a (traditionally) $12 - 36$ months study in 1 month, in a repeatable and generalized manner
- Generating interpretable insights using ML methodology that are acceptable by our healthcare practitioner user base
- Being robust to sparse, censored and noisy data
- Exhibit the flexibility of the ML components by performing repeated studies at scale.

In order to support the above contributions, we organize the content broadly into the following sections:

- We present a robust end-to-end SaaS framework to enable modeling of risk factors for disease related outcomes from complete EHR records of ∼55 million patients and validate the same using out-of-bounds prediction for outcome.
- We present the observational study of our framework applied to type 2 diabetes mellitus patients (T2DM) from the EHR dataset to model three of the most important outcomes: (a) renal failure, (b) diabetic retinopathy and (c) peripheral neuropathy.
- We validate the stability of the framework towards risk modeling as well as validation via out-of-bounds prediction.
- We briefly describe a published study performing comparative effectiveness of two drugs available in the US market in a highly accelerated time frame, using the SaaS architecture.

2 METHOD

Figure 1 provides an overview of the end-to-end SaaS framework. In this section, we focus on the machine learning layer and the use case. extracted risk factors.

2.1 Outcome-Driven risk analysis

Outcome-driven risk analysis, has been studied over many decades in various disciplines. Survival analysis[3] are a class of algorithms which can capture temporal risks associated with outcomes over simple classification analysis. Survival analysis is used extensively in health informatics where the time-to-event for outcome(s) of interest may include death, disease progression, and treatment response. Typically, such models are used to analyze extreme outcomes of interest in a pre-specified cohorts of patients based on underlying (and often pre-specified) factors. Statistical comparisons of significance for such factors are also generally investigated to interpret the model results. In this paper, we applied survival analysis to capture the temporal risk of outcomes from EHR datasets covering various facets of patient interactions such as diagnosis, medical encounters and lab observations.

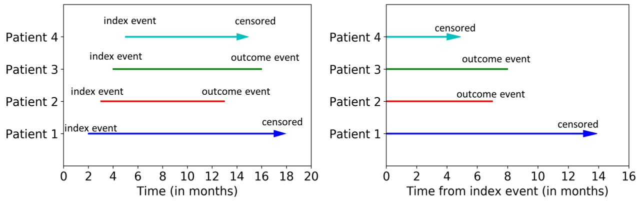


Figure 2: Basic setup of survival analysis. (left) Patients are admitted into studies over various time-points. Patients can either encounter the event of interest or are censored when no event happens within study period. (right) Patient timelines are normalized using index date.

Formally, for a set of patients $i = 1, 2, \cdots, N$ observed over time points t from an index event (for example, start of study or first diagnosis of diseases), survival analysis is concerned with modeling the probability of time of event $T_i, \forall i$ based on associated factors X_i. The survival function $S(t)$ is defined as the probability of survival beyond time t and is defined as:

$$S(t) = P(T > t)$$

Figure 2 illustrates the basic setup of survival analysis individuals are included in the study at various time points. Individuals may or may not be censored depending on whether they encounter the event within study period. The risk associated with such outcomes can also be captured by hazard rate $\lambda(t)$ defined as the event rate at time t conditional on survival until time t or later (that is, $T \geq t$). It can be mathematically expressed as:

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt)}{dt \cdot S(t)} = -\frac{S'(t)}{S(t)}. \quad (1)$$

Hazard rate can also be interpreted as a measure of the instantaneous risk of outcome. For more details see [3].

In the simplest case of categorical factors, Kaplan-Meir (K-M) curves can be used to compare the categorical factor to compare the associated hazards. For example, the importance of gender in renal failure can be analyzed using K-M curves to compare the hazards associated with male patients w.r.t female patients. However, these methods don't allow for continuous or mixed covariates. In such situations, regression based models such as Cox Proportional Hazard (CoxPH) models are more popular.

Cox-Proportional Hazard Models (CoxPH): These methods admit both continuous and categorical values and models the logarithmic hazard rate for an individual i at time point t as a weighted linear combination of the associated factors as given below:

$$\lambda(t|X_i) = \lambda_0(t) \exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) = \lambda_0(t) \exp(X_i \cdot \beta) \quad (2)$$

where, $\lambda_0(t)$ denotes the baseline hazard at time t. Baseline hazard is designed to be independent of patient covariates and only dependent on time. CoxPH model are generally fitted via partial log-likelihood ratios (which removes baseline hazards during fitting). Also depending on the the parameter fitting process, CoxPH can be used in a semi-parametric manner. The log-likelihood ratio can be expressed as:

$$\ell(\beta) = \sum_{i:C_i=1} \left(X_i \cdot \beta - \log \sum_{j:T_j \geq T_i} \theta_j \right) \quad (3)$$

where, $C_i = 0$ indicates that the patient was censored whereas $C_i = 1$ indicates an actual event.

Survival Support Vector Machines (SSVM): CoxPH models are readily interpretable and, due to the linear formulation of the log-likelihood (see equation 3), easily scalable. However, in this paper, we aim to find all possible relevant factors influencing outcome from our complete EHR. The linear assumption of factors for the complete set of factors is not likely to hold. Conversely, Support Vector Machines (SVM) are ideal for non-separable factors and have been successfully applied across several domains for problem such as regression, classification and anomaly detection. More recently, Van Belle et al. [18] postulated a formulation of SVM that can be used for survival analysis (SSVM).

SSVM admits data from N patients in the form of triplets $(\mathbf{X}_i, T_i, \delta_i)$ where $\delta_i \in \{0, 1\}$ is the binary event indicator. SSVM aims to minimize the following objective function:

$$\arg\min_{\mathbf{w}, b} \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \tfrac{\alpha}{2}[r \textstyle\sum_{i,j \in \mathcal{P}} \max(0, 1 - (\mathbf{w}^T\mathbf{X}_i - \mathbf{w}^T\mathbf{X}_j))^2$$
$$+(1 - r) \textstyle\sum_{i=0}^{n} \left(\zeta_{\mathbf{w}, b}(T_i, X_i, \delta_i)\right)^2]$$

where,

$$\zeta_{\mathbf{w}, b}(T_i, \mathbf{X}_i, \delta_i) = \begin{cases} \max(0, T_i - \mathbf{w}^T\mathbf{X}_i - b) & \text{if } \delta_i = 0, \\ T_i - \mathbf{w}^T\mathbf{X}_i - b & \text{if } \delta_i = 1 \end{cases}$$

$$\mathcal{P} = \{(i, j) \mid T_i > T_j \wedge \delta_j = 1\}_{i,j=1,\ldots,n}$$

The hyper-parameter $\alpha > 0$ determines the amount of regularization to apply with an increasing value signifies increased regularization.

Survival Support Vector Machines with clinical Kernel (SKSVM): SSVM relaxes the linear assumptions associated with standard CoxPH models. However, standard SSVM kernel doesn't disassociate between categorical and continuous factors leading to a poor ordinal fit of the data for survival methods where within factor comparisons are of interest. Daemen et al.[4] postulated clinical kernels that aims to address such data formulations and models clinical data using a min-max variant of kernel referred to as clinical kernel. Formally, the kernel can be expressed as:

$$k_X(i,j) = \frac{(max - min) - |X_i - X_j|}{max - min} \tag{4}$$

where *max* and *min* are the maximal and minimal value for variable X defined on the training data set. Pölsterl et al. [13] postulated a fast training algorithm for Kernel Support Vector machines which have been found to produce stable performance across many datasets. *For our framework, we thus chose SKSVM to admit possibility of non-linear factor separability while still retaining the modes of the survival dataset to model the temporal risks associated with outcomes.* For a complete comparison of these methods please see Section 3

2.2 Identification of relevant factors

Classical methods for survival analysis capture the risks over a pre-determined set of factors or covariates. These models, can be readily used to compare the effects, within- and between the factors systematically. Various tests of statistical significance, such as log likelihood tests can be performed to declare these significance measures. Our goal is to identify the risk factors without pre-specifications and/or expert guidance and thus the framework is formulated to admit all all possible observational factors from the EHR dataset. The classical models are thus not ideal for our use case, especially when such factors/covariates are possibly correlated. Following the success of LASSO or L1 regularization [16] in various problem domains, significant research on applying such algorithms to survival analysis using CoxPH models have been proposed [5, 7, 17]. Formally, the log-likelihood under L1 penalty is given as:

$$\begin{aligned}\ell(\beta) = {} & \textstyle\sum_j (\sum_{i \in H_j} X_i \cdot \beta \\ & - \textstyle\sum_{\ell=0}^{m-1} \log \left(\sum_{i:T_i \geq t_j} \theta_i - \frac{\ell}{m} \sum_{i \in H_j} \theta_i \right)) + \phi \|\beta\|_1\end{aligned} \tag{5}$$

where, ϕ is the L1 penalty factor.

However, from initial experimentation we observed that while L1-regularization leads to relevance of features, the model fits are not stable and fail to optimally converge. Sparsity of information about each individual factor in a single encounter as well as significant levels of censoring (about 95%) for the outcomes of interest is commonplace in the real world. We postulate that the 'missingness' of encounters/outcomes renders the survival likelihoods constructed with all factors ill-formed and leads to pre-mature convergence. Feature selection procedures have been successfully applied under similar conditions in various domains [6]. Figure 3 shows a comparison of concordance measures using three strategies:

- L1 regularized CoxPH models
- CoxPH models fitted on pre-processed set of factors selected using Non-Negative Matrix Factorization (NMF) [6].
- CoxPH models fitted on pre-processed set of top k features selected using Chi-square distances [6] of factors to event outcome.

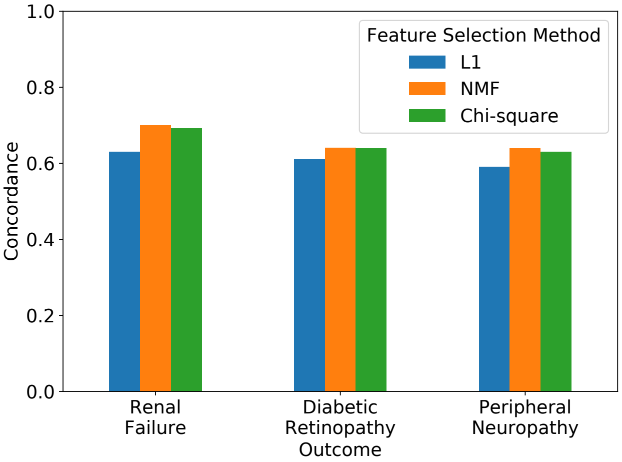


Figure 3: Comparison of Feature Selection Strategies using CoxPH model using concordance as accuracy measure. Chi-square distance-based factors perform at a similar level with NMR projected factor sets.

Our result indicates that the feature selection processes reduces the sparsity, retains the most relevant factors and leads to better concordance. Although NMF often leads to the optimal concordance scores, it transposes the factors to a low dimensional representation where the original factors are lost. As such we opted for Chi-square strategy where the selected factors are subsets of the original factors and interpretable with respect to risk analysis with comparable results. For other possible strategies see [6]. *For the framework, we chose to use Chi-square strategy to maintain stability as well as interpretability of our models.*

2.3 Validation of factors

Factors identified from EHR datasets require a rigorous clinical validation. However, its impractical to compare the observational cohorts against tightly controlled cohorts from registries or clinical trials as the observational cohorts are not guaranteed to capture all confounding factors. Thus, we aim to extract factors towards hypothesis generation which can then be verified in clinical settings and the framework instead includes a validation step to accurately quantify the importance of such factors as supported by EHR data. More specifically, we validate the factors by predicting for an occurrence of outcomes among 'at-risk' patients from the cohort over various observation windows based on their concurrent medical history. We use a Support Vector regression on the said factors and quantify the importance of the factors using their relative feature importance for prediction. For more details see Section 3.2.2.

3 EXPERIMENTS

3.1 Experimental Setup

Figure 1c illustrates the data flow and the components. In the first step, dataset(s) can be interactively created for risk analysis by selecting the cohort of interest and subsequently extracting and normalizing all relevant factors based on therapeutic definitions.

Specifically, for the illness under consideration we employ the following steps:

- Identify cohort of interest by selecting patients using therapeutic definitions (e.g.'snomed concept codes' for diagnosis and 'LOINC ids' for observation)
- Extract all relevant encounters along with all lab observations for such encounters, for the selected cohort
- For each events of interest, extract all relevant outcomes for event identification
- Resolve and normalize observations and outcomes w.r.t. encounters based on closest encounters

EHR records are in general noisy, sparse, and may contain 'incomplete' medical encounters. Thus, some key data processing challenges involve curating relatively synchronized and complete medical histories, while admitting maximal patient coverage. We admit patients with mostly complete medical histories and resolve encounters/observations, which usually feature a 'many-to-many' mapping on a similar temporal scale by associating observation with the closest historical encounter.

The encounter-observation information extracted out of this process is sparse w.r.t. factors as all medical encounters don't necessarily lead to observation of all relevant metrics such as HbA1c level. We filter the factors to top 95% cumulative coverage in our complete dataset. For each individual patients, we subsequently fill the factors by linearly interpolating the patient specific attributes. It is to be noted that survival methods compare for importance of factors based on their relative change. Thus, without loss of generality, we fill in the remaining missing factors with the cohort mean.

Finally, as illustrated in Figure 2, we use the first diagnosis date of each individual patients as the corresponding index data and normalize the time scale for subsequent risk analysis. It is to be noted that the final patient attributes can encompass both static information such as 'age at first diagnosis' and time varying information such as weight that changes throughout the patient medical history. *Thus we employ our survival analysis using time-varying covariates by treating the dataset as an interval censored dataset. For more details see [3].*

3.1.1 Case Studies: 3 common complications of Type 2 Diabetes. We applied our framework to model risk factors among Type 2 Diabetes Mellitus (T2DM) patients. T2DM affects approximately 30 million patients worldwide and is often associated with multiple severe outcomes. We select the cohort of patients affected by T2DM using a three step definition as follows:

(1) **Diagnostic:** Patient must have been diagnosed with T2DMs (identified via 'snomed' codes)
(2) **Observational:** Patients must have had abnormally high HbA1c levels (> 5.7) in their observations
(3) **Support:** Patients must have had at least 3 diabetic related encounters

These definitions ensure that the patients have been identified as diabetic both from procedural and diagnostic side, thus reducing false positives. This also helps to remove left censored patients. As a result, we can assume the first diabetic encounters of patients to be the start time and proceed with normal survival analysis. Using these definitions we find a diabetic cohort spanning approximately 4 million patients. We analyze 3 of the most common complications among T2DM patients:

• **Renal Failure** (RF): We considered both critical and non-critical renal failures defined by ICD-9 codes 584.∗ and 585.∗. 93682 T2DM patients had renal failures.
• **Diabetic Retinopathy** (RT): We considered the following ICD-9 codes = 362.01, 362.02, 362.03, 362.04, 362.05, 362.06. 19587 T2DM patients had Diabetic Retinopathy.
• **Peripheral Neuropathy** (PN): We considered the following ICD-9 codes - 250.6, 357.2, 337.1. 47915 T2DM patients had Peripheral Neuropathy.

We analyzed each outcome separately and marked each encounters with positive/negative occurrence of the event of interest. It is to be noted, the dataset is extremely skewed with majority of the encounters correlated with no outcome events. Figure 4 shows the demographic distribution of patients in our final cohort at their first renal failure. The figures indicate that 'age' could be one of the key risk factors along with race and gender. We investigate such apparent as well as possible hidden factors and present our analysis in the following section.

3.2 Experimental evaluation

We evaluate our framework on the aforementioned 3 separate outcomes against the candidate methods outlined in Section 2. Furthermore, to evaluate the importance of the extracted factors we consider the following factors identified by experts as possible indicators: (a) Age at first renal failure, (b) weight, (c) gender, (d) race, and (c) HbA1C level in blood.

For the subsequent analysis, we treat these 5 factors as our 'expert factor' and benchmark the data-driven factors against the same.

3.2.1 Factor Identifications. We applied our framework on each of the three outcomes and extracted the most relevant factors from final set of 567 factors following the methodology outlined in Section 3.1. Following the discussion presented in Section 2, we consider the following methods as candidates:

- CoxPH model on expert factors (M1)
- SSVM model on expert factors (M2)
- SKSVM model using clinical kernel on expert factors (M3)
- CoxPH model jointly trained with Chi-square selected factors (M4)
- SSVM model jointly trained with Chi-square selected factors (M5)
- SKSVM model using clinical kernel, jointly trained with Chi-square selected factors (M6)

To evaluate the model accuracy, we compare the predicted survival outcome of encounters using the standard concordance metric. Concordance can be thought of as a rank correlation metric (range [0, 1]) where the predicted order of outcomes are compared against actual outcome outcomes such that a higher score indicates a better model fit. For more details see [3]. For our purpose, when the predicted survival time is the quantity of interest, the concordance or c-index can be given as:

$$c = \frac{1}{N} \sum_{i:\delta_i=1} \sum_{j:T_i<T_j} \mathcal{I}[S\left(\hat{T}_j|X_j\right) > S\left(\hat{T}_i|X_i\right)] \qquad (6)$$

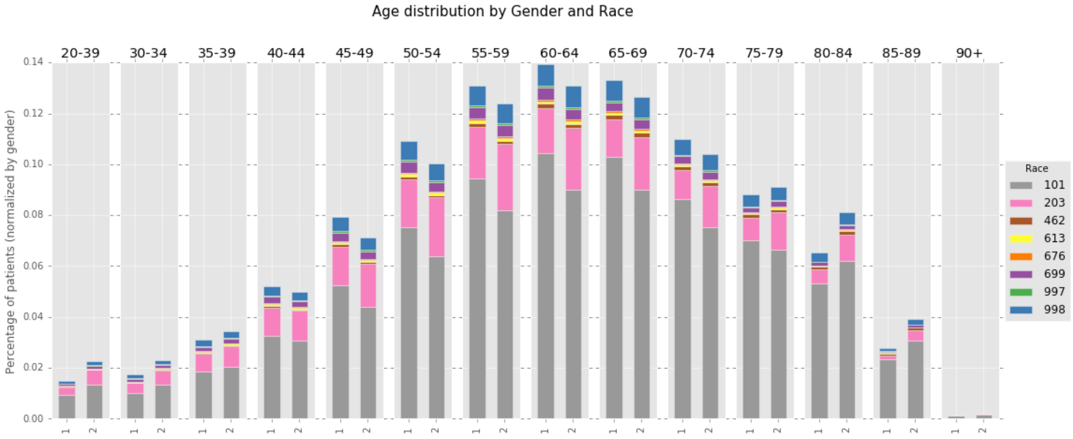


Figure 4: Demographic distribution of T2DM patients at first diagnosis.

Table 1: Accuracy comparison of models M1 - M6 for 3 separate case studies. Concordance is used to evaluate the model fitness.

Factors	Model	Diabetic Retinopathy	Peripheral Neuropathy	Renal Failure
expert	CoxPH	0.5423	0.5404	0.5542
	SSVM	0.5457	0.5409	0.5568
	SKSVM+clinical Kernel	0.6149	0.6208	0.6551
data-driven	CoxPH	0.6395	0.6307	0.6923
	SSVM	0.6489	0.6364	0.6811
	SKSVM+clinical Kernel	**0.8186**	**0.8185**	**0.8446**

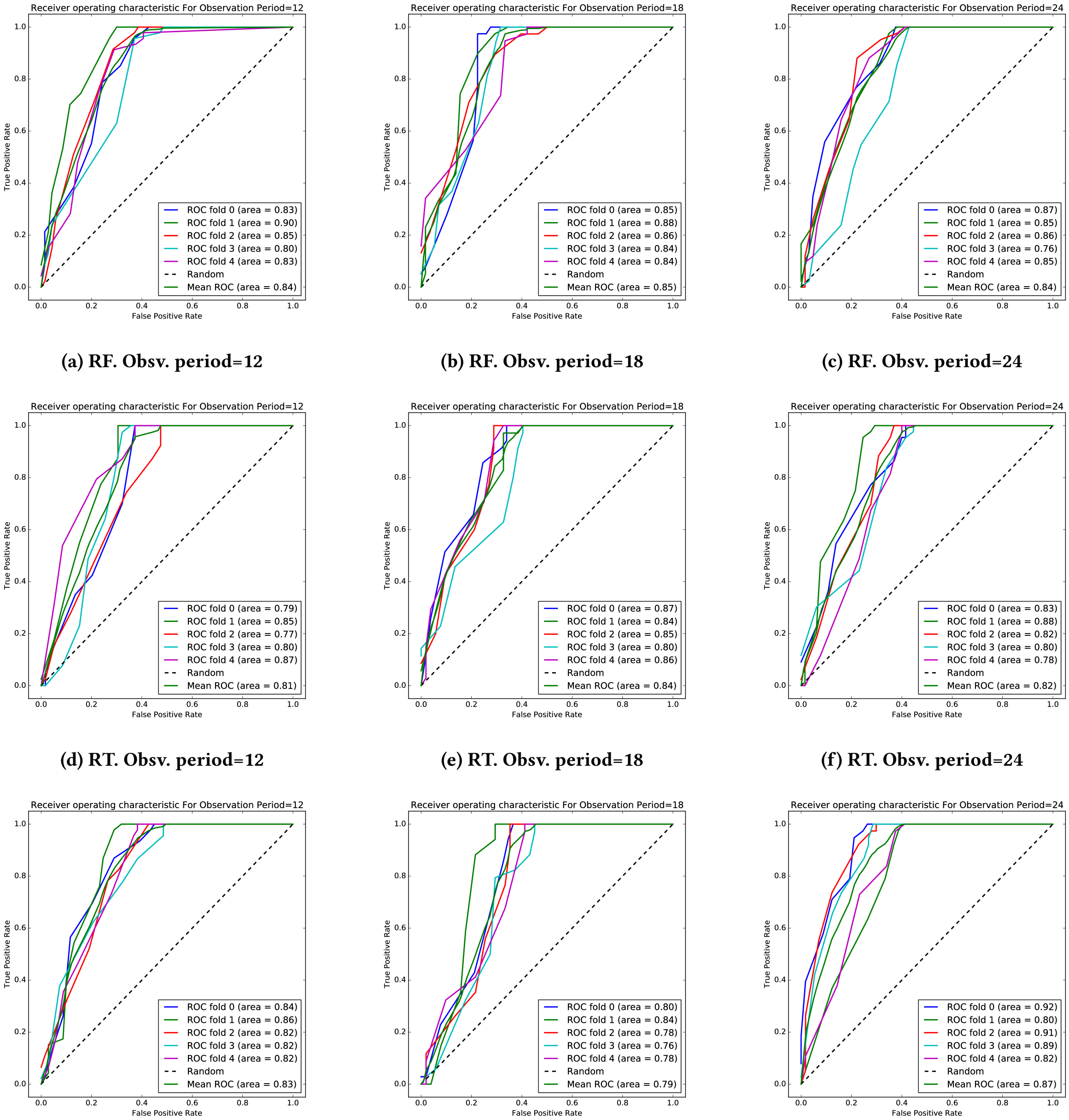
Table 2: Validation of importance of extracted data-driven factors against expert factors for 3 separate case studies. Out-of-bounds prediction performance over multiple observation period is used to quantify factor validity.

		Diabetic Retinopathy		Peripheral Neuropathy		Renal Failure	
Obsv. Period	Metric	data-driven	expert	data-driven	expert	data-driven	expert
12	Accuracy	0.719563	0.691053	0.738503	0.680790	0.755902	0.718057
	F1	0.652231	0.599672	0.681468	0.578900	0.703365	0.651233
	Precision	0.675897	0.598846	0.706570	0.553237	0.720814	0.656522
	Recall	0.639296	0.611859	0.669253	0.612360	0.690043	0.647177
18	Accuracy	0.733189	0.692044	0.704882	0.735091	0.741491	0.716642
	F1	0.688748	0.623886	0.641899	0.657090	0.685340	0.637730
	Precision	0.741176	0.637983	0.666387	0.648235	0.714170	0.624831
	Recall	0.643792	0.610881	0.625252	0.674860	0.666846	0.654102
24	Accuracy	0.721305	0.710160	0.775285	0.718150	0.741807	0.705830
	F1	0.664836	0.638786	0.733218	0.636547	0.669804	0.628075
	Precision	0.691121	0.648837	0.762304	0.624324	0.658250	0.625249
	Recall	0.643597	0.632442	0.716624	0.656825	0.683981	0.635019

where, $\mathcal{I}$ is the indicator function, $S(\hat{T}_i|X_i)$ is the predicted survival probability for the i^{th} data point, and *num* is the number of comparable pairs.

Table 1 reports the accuracy results for the 6 candidate models. As can be seen, SKSVM jointly trained with feature selection step using Chi-square strategy exhibits the best performance across all 3 of the studies indicating the effectiveness as well as the generalizability of the proposed step. *This factor identification model is used to evaluate and analyze the subsequent steps of the framework.*

3.2.2 *Prediction of disease outcomes.* To validate the extracted factors, we use the predictive component of the framework to predict for each outcome individually based on the past historical



(a) RF. Obsv. period=12 **(b) RF. Obsv. period=18** **(c) RF. Obsv. period=24**

(d) RT. Obsv. period=12 **(e) RT. Obsv. period=18** **(f) RT. Obsv. period=24**

(g) PN. Obsv. period=12 **(h) PN. Obsv. period=18** **(i) PN. Obsv. period=24**

Figure 5: Area under ROC curves for outcome prediction using data-driven factors. (a)-(c) ROC curves for Renal Failure (RF), (d)-(f) ROC curves for Diabetic Retinopathy (RT), and (g)-(i) ROC curves for Peripheral Neuropathy (PN).

records. For each individual outcome, we collect all the factors over a pre-specified observation period from the first diagnosis date and predict the possibly of occurrence of the relevant outcome within the next 6 months. To accurately test the sensitivity of the factors, we vary the observation period over 12, 18, and 24 months respectively. For each observation period, we only consider those patients who are in the 'risk set' at the observation boundary i.e. haven't encountered the relevant outcome within the observation period. We compare the prediction performance using only expert factors and the factors identified from the previous step and report our 5 fold cross-validation performance in Table 2. The data-driven factors are found to produce the best performance with respect to all the categories and case studies. Furthermore, it can be seen that while accuracy measures are high, F1-scores are found to be a function of the total coverage of outcomes for both factors.

4 DISCUSSIONS

As discussed earlier, our goal was to propose a robust, generalizable and data-driven framework to extract risk factor for outcomes. In this section, we discuss some of the more interesting findings towards the stated goal based on our results from Section 3.

Stability of Factor Identification: To analyze the stability of the factor identification model, we plot the error distribution over 5 fold cross-validation fit of SKSVM jointly trained with Chi-square distribution for the 3 case studies in Figure 6. The error distribution indicates that the framework is stable across the different fold sets with similar levels of error spread with respect to concordance. Narrower spread of error is observed for Renal Failure and Diabetic Retinopathy compared to Peripheral Neuropathy. This can be argued to be a function of outcome prevalence but could also be a function of comorbidities among affected patients.

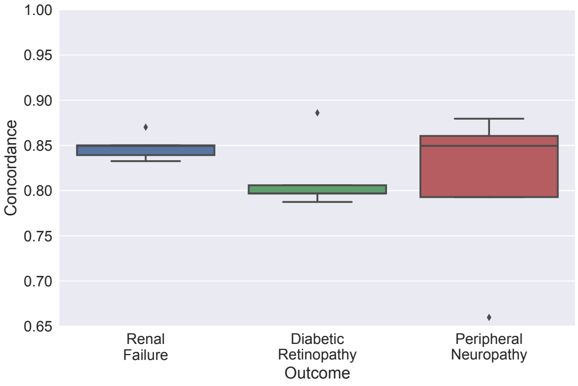


Figure 6: Stability Analysis of Risk Modeling. Error distribution of 5 fold cross validation for Renal Failure.

Sensitivity of factor validation : To analyze the sensitivity of the factor validation step, we compare the area under ROC curves over 5 fold cross validation for all the three studies over the observation periods in Figure 5. It is interesting to note that the mean AUC metric is not found to vary with observation period showing slight deviations across the 3 case studies. While it can be considered as a proof of the stability of the factors, this could also be an artifact of an ever-changing risk set such that the information gain in terms of past history for a longer observation period is accompanied by an increased skewness of the model with an ever decreasing positive occurrence in the training set. Nevertheless, the mean AUC's indicate that the data-driven factors are relevant and could be considered for further clinical studies.

Data-driven vs Expert Risk Factors : We present the top 20 data-driven factors extracted from each of the 3 studies in Tables 3. The data-driven factors identified via Section 3.2.1 were validated via Section 3.2.2 and ranked with respect to the feature importance for predictive task. It is interesting to note that while several of the expert factors such as some race and age groups are found to be important, several non-expert observational factors appear prominently for the case studies. Significantly, several blood pressure, heart rate, and blood platelet measures are found to be significant and could be useful for further clinical studies.

Acceleration in Study Times: In order to establish time savings we conducted another real world study [20]. This time we wanted to compare the effectiveness of two drugs available in the US market in patients at risk of stroke or systemic embolic events (SEE). We compared a blood thinning medication Edoxaban introduced into the US market after FDA approval in 2015. The incumbent drug of choice in the market at that time was Warfarin. During the clinical trial Edobaxan was found to be non-inferior to Warfarin in a controlled study. In our study supported by real world data, we concluded that Edobaxan is non-inferior to Warfarin in prevention of stroke but reduces the risk of embolism was compared to Warfarin. Using our framework, this study was completed from problem statement to submission for publication, including review and approval by subject matter experts in less than 1 month with a very small number of human resources. The average time frame for these type of studies is > 1 yr. These long time frames were also validated by many of our internal market and subject matter experts. Assuming a blended rate of $\$100/hr$ and an average of 3 resources for the year, the total savings in this study would amount to $\sim 600k$. Given a medium scale pharmaceutical company that has 10 drugs in the market, it would translate to a $USD6M$ saving. Worth noting is that these savings do not represent the total impact on the health system that can result due to the accelerated timelines. The total impact, beyond the scope of quantifying in this paper, would be significantly higher.

5 CONCLUSION

In this paper, we proposed an end-to-end ML driven SaaS framework to analyze massive EHR datasets. We illustrated a scalable, generalized and interpretable ML framework to uncover risk factors for multiple disease outcomes as well as to understand the comparative effectiveness of drugs.

The results indicate that the system exhibits robustness in identifying risk factors from sparse event outcomes without expert supervision. The system is able to gracefully handle the often incomplete and censored data that regularly turn out to be the Achilles heal of the previous methods utilized in the field.

This system exhibits stability in identifying the relevant factors and generalizability in terms of consistent performance across all case studies. In our experience this is the largest and the most generic system to perform such studies with highly accelerated timelines. The system also supports interpretability which is a key to the acceptance of the results and the insights generated by such systems without having to go through manually intensive processes, data gathering, hypotheses generation and review cycles.

We evaluated our framework against 5 candidate models that includes the traditional models used for risk analysis for 3 separate case studies. Furthermore, we validate the importance of the extracted factors by using these to predict for the corresponding outcomes over multiple observation windows and compare against the expert factors. The factors identified by the framework can be used for further clinical studies and are useful for hypothesis generation.

Table 3: Top data-driven risk factors for Renal Failure among T2DM patients as identified by the framework.

Peripheral Neuropathy			Renal Failure			Diabetic Retionpathy		
Factors	Score	Description	Factors	Score	Description	Factors	Score	Description
AGE: (64, 69]	0.082		AGE: (69, 74]	0.090		AGE: (64, 69]	0.132	
ET3111-5	0.041		RACE: 203.0	0.035		FEMALE	0.037	
8277-6	0.036	BSA	ET3111-5	0.033		RACE: 203.0	0.034	
8867-4	0.022	Heart rate	3094-0	0.028	BUN SerPl-mCnc	8302-2	0.020	Bdy height
29463-7	0.021	Weight	AGE: (79, 84]	0.025		8462-4	0.019	BP dias
39156-5	0.019	BMI	8302-2	0.023	Bdy height	39156-5	0.018	BMI
8462-4	0.019	BP dias	8480-6	0.021	BP sys	26453-1	0.018	RBC # Bld
FEMALE	0.017		8462-4	0.021	BP dias	8480-6	0.017	BP sys
2339-0	0.017	Glucose Bld-mCnc	8867-4	0.021	Heart rate	ET3111-5	0.016	
8480-6	0.016	BP sys	FEMALE	0.019		2965-2	0.014	Sp Gr Ur
8310-5	0.015	Body temperature	26453-1	0.019	RBC # Bld	8277-6	0.01	BSA
8302-2	0.015	Bdy height	8310-5	0.017	Body temperature	29463-7	0.014	Weight
17861-6	0.015	Calcium SerPl-mCnc	29463-7	0.017	Weight	8867-4	0.013	Heart rate
1975-2	0.015	Bilirub SerPl-mCnc	39156-5	0.017	BMI	3094-0	0.013	BUN SerPl-mCnc
2160-0	0.014	Creat SerPl-mCnc	8277-6	0.015	BSA	2951-2	0.013	Sodium SerPl-sCnc
26511-6	0.014	Neutrophils NFr Bld	4548-4	0.014	Hgb A1c MFr Bld	2161-8	0.013	Creat Ur-mCnc
4548-4	0.014	Hgb A1c MFr Bld	2028-9	0.013	CO2 SerPl-sCnc	2339-0	0.012	Glucose Bld-mCnc
2075-0	0.013	Chloride SerPl-sCnc	2160-0	0.012	Creat SerPl-mCnc	17861-6	0.012	Calcium SerPl-mCnc
2085-9	0.013	HDLc SerPl-mCnc	26515-7	0.012	Platelet # Bld	9279-1	0.01	Resp rate
RACE: 203.0	0.012		1742-6	0.012	ALT SerPl-cCnc	2028-9	0.011	CO2 SerPl-sCnc

We illustrate the impact of a carefully constructed ML framework in critical problems such as clinical hypothesis generation by reducing the timelines of studies from an average of 12 – 36 months to ~1 month. This massive improvement can lead to millions, potentially billions of dollars of impact, both in terms of cost savings as well as accelerating discovery, research and future investments.

Our future research involves improving the system both in terms of ML methodologies and robustness to clinical use cases. For the former, we are currently focusing efforts on newer deep learning methods. Whereas these methods are shown to be superior in accuracy, our user base is completely un-accepting of black box models. As such, we are exploring more interpretable deep architectures to potentially bridge that gap. For clinical robustness, we are focusing on modeling risk factors for competing outcomes to better model the complexities of individual patient trajectories.

REFERENCES

[1] Dustin Charles, Meghan Gabriel, and Michael F Furukawa. 2013. Adoption of Electronic Health Record Systems among US Non-Federal Acute Care Hospitals: 2008-2012. *ONC data brief* 9 (2013), 1–9.

[2] Yu-Yi Chen, Jun-Chao Lu, and Jinn-Ke Jan. 2012. A Secure EHR System based on Hybrid Clouds. *Journal of medical systems* 36, 5 (2012), 3375–3384.

[3] David R Cox. 1992. Regression Models and Life-tables. In *Breakthroughs in statistics*. Springer, 527–541.

[4] Anneleen Daemen and Bart De Moor. 2009. Development of a kernel function for clinical data. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 5913–5917.

[5] J. J. Goeman. 2010. L1 Penalized Estimation in The Cox Proportional Hazards Model. *Biometrical Journal* 52 (2010), –14.

[6] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.

[7] Yolanda Hagar, David Albers, Rimma Pivovarov, Herbert Chase, Vanja Dukic, and Noémie Elhadad. 2014. Survival analysis with electronic health record data: Experiments with chronic kidney disease. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 7, 5 (2014), 385–403.

[8] Chun-Ju Hsiao, Esther Hing, Thomas C Socey, Bill Cai, et al. 2010. Electronic Medical Record/Electronic Health Record Systems of Office-based Physicians: United States, 2009 and Preliminary 2010 State estimates. *National Center for Health Statistics* (2010).

[9] Peter B Jensen, Lars J Jensen, and Søren Brunak. 2012. Mining Electronic Health Records: Towards Better Research Applications and Clinical Care. *Nature Reviews Genetics* 13, 6 (2012), 395–405.

[10] Cheng Liu, Yong Liang, Xin-Ze Luan, Kwong-Sak Leung, Tak-Ming Chan, Zong-Ben Xu, and Hai Zhang. 2014. The L1/2 Regularization Method for Variable Selection in the Cox Model. *Applied Soft Computing* 14 (2014), 498–503.

[11] Gou Masuda and Norihiro Sakamoto. 2002. A Framework for Dynamic Evidence based Medicine using Data Mining. In *Computer-Based Medical Systems, 2002.(CBMS 2002). Proceedings of the 15th IEEE Symposium on*. IEEE, 117–122.

[12] Lucila Ohno-Machado et al. 2015. Mining Electronic Health Record Data: Finding the Gold Nuggets. *Journal of the American Medical Informatics Association* 22, 5 (2015), 937–937.

[13] Sebastian Pölsterl, Nassir Navab, and Amin Katouzian. 2015. Fast training of support vector machines for survival analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 243–259.

[14] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. 2017. Deep learning for health informatics. *IEEE journal of biomedical and health informatics* 21, 1 (2017), 4–21.

[15] Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Edabollahi, Steven E Steinhubl, Zahra Daar, and Walter F Stewart. 2012. Combining knowledge and data driven insights for identifying risk factors using electronic health records. In *AMIA Annual Symposium Proceedings*, Vol. 2012. American Medical Informatics Association, 901.

[16] Robert Tibshirani. 1996. Regression Shrinkage and Selection via The Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[17] Robert Tibshirani et al. 1997. The Lasso Method for Variable Selection in The Cox Model. *Statistics in medicine* 16, 4 (1997), 385–395.

[18] Vanya Van Belle, Kristiaan Pelckmans, JAK Suykens, and Sabine Van Huffel. 2007. Support vector machines for survival analysis. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*. 1–8.

[19] Steven Wallace, Keith Maxey, and Lakshmi S Iyer. 2014. A Multi-case Investigation of Electronic Health Record Implementation in Small-and Medium-size Physician Practices. *Journal of Information Technology Case and Application Research* 16, 1 (2014), 27–48.

[20] L Zhou, R Zhang, P Chakraborty, F Farooq, and S Hensley Alford. 2018. Comparative Effectiveness Of Edoxaban And Warfarin In Prevention Of Stroke And Systemic Embolism In Non-Valvular Atrial Fibrillation Using Observational Healthcare Data. *Value in Health* 21 (2018), S56.