Inferring the Interactions of Risk Factors from EHRs

Travis Goodwin, Sanda M. Harabagiu, PhD
University of Texas at Dallas, Richardson, TX, USA

Abstract

The wealth of clinical information provided by the advent of electronic health records offers an exciting opportunity to improve the quality of patient care. Of particular importance are the risk factors, which indicate possible diagnoses, and the medications which treat them. By analysing which risk factors and medications were mentioned at different times in patients' EHRs, we are able to construct a patient's clinical chronology. This chronology enables us to not only predict how new patient's risk factors may progress, but also to discover patterns of interactions between risk factors and medications. We present a novel probabilistic model of patients' clinical chronologies and demonstrate how this model can be used to (1) predict the way a new patient's risk factors may evolve over time, (2) identify patients with irregular chronologies, and (3) discovering the interactions between pairs of risk factors, and between risk factors and medications over time. Moreover, the model proposed in this paper does not rely on (nor specify) any prior knowledge about any interactions between the risk factors and medications it represents. Thus, our model can be easily applied to any arbitrary set of risk factors and medications derived from a new dataset.

Introduction

As defined by the World Health Organization (WHO), a risk factor is any attribute, characteristic, or exposure of an individual that increases the likelihood of developing a disease. Because risk factors are such powerful indicators of the likelihood of a patient developing a disease, they play a critical role in the management and care of individual patients. Naturally, risk factors are frequently explicitly documented in the Electronic Health Record[1] (EHR) associated with a patient. However, as revealed by consultations conducted by the Informatics for Integrating Biology at the Bedside (i2b2) and University of Texas Health and Sciences center (UTHealth) with clinicians, many risk factors are not explicitly diagnosed; rather, they are merely implied through natural language text in the EHR [1]. For example, an EHR may omit an explicit diagnosis of *diabetes*, instead stating an abnormally high blood glucose measurement indicative of the disease. For this reason, it is important to consider both the explicitly mentioned risk factors as well as the textual indicators that suggest them. In addition to risk factors, EHRs also document other elements of the patient's care, such as the medications prescribed to the patient. The prescription of medications and the presence of risk factors play complementary roles in the management and care of a patient: risk factors increase the likelihood of a patient having or developing a disease, while medications decrease the likelihood of the disease presenting in the future. Unfortunately, the exact relationship between individual medications and the risk factors they are targeting is rarely stated in EHRs. Moreover, many medications which target a particular risk factor can interact with the other risk factors associated with a patient. These interactions are difficult to anticipate without elaborate clinical trials and analysis, particularly for uncommon combinations of risk factors. To make matters worse, there are a variety of complex interactions between multiple risk factors (i.e. a patient with high blood pressure who also smokes is more likely to develop coronary artery disease than a patient with only high blood pressure). However, by exploiting the fact that EHRs document the risk factors and the medications given to patients at different times during their clinical care, it is possible to construct a chronological model of how the risk factors and medications interact over time. In this paper, we define a novel data-driven probabilistic model of the interactions between risk factors and medications which uses statistical trends discovered across a large set of EHRs. We also show how this model can be used to (1) predict the presence or absence of certain risk factors in a patient's future, to (2) discover the relationships between individual risk factors and medications, and to (3) identify patients with irregular or unusual progressions of risk factors and medications.

In order to evaluate our model, we utilized the set of longitudinal EHRs provided by the organizers of the *Challenges in Language Processing for Clinical Data* shared task sponsored by the 2014 Informatics for Integrating Biology and the Beside (I2B2) and University of Texas Health Science Center (UTHealth). These EHRs document the progression of heart disease for a population of diabetic patients, and are particularly well-suited for our model because they were manually annotated by physicians to denote the presence of risk factors and medications relevant to diagnosing heart disease.

The remainder of this paper is organized as follows. Section 2 reviews related work and provides background information on the theory supporting our model. Section 3 describes the dataset used both to motivate and evaluate our model, Section 4 details our approach, while Section 5 presents our results which are discussed in Section 6. Finally, Section 7 summarizes our conclusions.

[1] In this paper, we consider the EHR of a patient to be the set of all chronological electronic medical records associated with a patient. Thus, each patient is associated with a single EHR which consists of multiple individual records, such as discharge summaries generated on different dates.

Related Work and Background

Historically, temporal models for clinical prediction use established criteria specific to an individual disease and do not often generalize well to new diseases. For example, a regression model capable of selecting patients who may become at risk for heart disease was developed in [2], while a variety of different prediction models were analysed based on their ability to screen for individual types of cancer based on known antigen relationships in [3]. An automatic system based entirely on narrative content was constructed in [4] and evaluated for its ability to identified patients with pneumonia based on past mentions of the disease. More recent models have focused on modelling multiple types of diseases jointly. A disease-subtype prediction model was developed in [5] which relies on mixture modeling and a joint-disease risk prediction model using logistic regression was described in [6].

However, these models cannot account for variations in the amount of time between successive disease observations. Moreover, the more generalized models do not account for the common semantics associated with diseases and medications (namely, that disease can predict disease, and that medications can prevent disease). In order to advance predictive modelling past both of these barriers, we developed a general multiple risk factor and medication prediction model based on recent advances in statistical modeling. Specifically, we rely on a powerful probabilistic framework known as Probabilistic Graphical Models (PGMs) [7] which can be viewed as a generalization of both mixture and regression modelling. Graphical models are able to not only encode knowledge about multiple risk factors and medications at particular times, but can also directly represent the inter-actions between these different points in time.

In this work, we leverage both the sequential modelling and probabilistic inference capabilities of PGMs by defining a model of patient's chronologies which is general in the sense that it does not rely on pre-specified knowledge about the relationships between risk factors and medications. Our model is able to recover these relationships from a large body of EHRs, enabling us to not only predict the way risk factors may progress for patients, but to discover the latent interactions between risk factors and disease.

Materials

When conducting our experiments, we used a collection of EHRs associated with 178 diabetic patients, provided by the organizers of the shared-tasks on *Challenges in Language Processing for Clinical Data*[2] sponsored by the 2014 Informatics for Integrating Biology and the Beside[3] (i2b2) and the University of Texas Health Science Center at Houston[4] (UTHealth). These EHRs document chronological information about the progression of heart disease for diabetic patients and thus each EHR contains between three to five individual reports (in the form of discharge summaries) generated at different times. In total, the EHRs in this collection contain 790 de-identified discharge summaries. The discharge summaries associated with each EHR contain (1) a patient code which uniquely identifies the patient associated with the discharge summary, (2) a timestamp indicating the approximate creation time of the summary, as well as (3) a large body of narrative text. Note that in order to follow HIPAA guidelines and to protect patients' privacy, the patient information in these records was de-identified, meaning that patients' names and, more importantly, the timestamps associated with each individual discharge summary are obfuscated. Fortunately, the timestamps were obfuscated in a way that preserved the relative elapsed time between successive discharge summaries for the same patient. That is, the de-identification procedure merely adjusted all timestamps for a patient by a fixed amount, so that although the exact date of each discharge summary cannot be recovered, the relative elapsed time between successive discharge summaries is unchanged. The 2014 i2b2/UTHealth dataset was well suited for our experiments because it contains gold-standard annotations explicitly documenting the presence of risk factors and medications associated with hearth disease. A total of 7 risk factors were considered, with each risk factor having a variety of indicators:

- Diabetes was indicated by (1) an explicit diagnosis of type 1 or type 2 diabetes, (2) an A1c test over 6.5, or (3) mentions of two fasting blood glucose measures above 126.
- Coronary artery disease (CAD) was indicated by (1) an explicit diagnosis of coronary artery disease, (2) a mention of myocardial infarction, (3) description of revascularization, cardiac arrest, or ischemic cardiomypathy, (4) a stress test showing ischemia, (5) abnormal cardiac catherization showing coronary stenoses, or (6) chest pain consistent with angina.
- Hyperlipidemia was indicated by (1) an explicit diagnosis of hyperlipidemia or hypercholesterolemia, (2) a total cholesterol measurement above 240, or (3) an low-density lipoprotein (LDL) measurement of over 100 mg/dL.

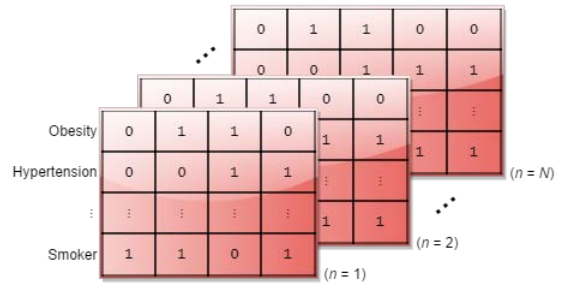
[2]Information on these shared-tasks is available at `https://www.i2b2.org/NLP/HeartDisease/`.
[3]I2B2 is a NIH-funded National Center for Biomedical Computing. Additional information is available at `https://www.i2b2.org/index.html`.
[4]THE UT Health Science Center is part of the University of Texas System. More information is available at `https://www.uth.edu/`.

- Hypertension was indicated by (1) an explicit diagnosis of hypertension, or (2) a blood pressure measurement above 140/90 mm/hg.
- Obesity was indicated by (1) a description of the patient being obese, (2) a body mass index (BMI) above 30, or (3) a waist circumference above 40 in for males or 35 in for females.
- Family history of premature CAD was indicated by a description of a first-degree relative (i.e. parent, sibling or child) who diagnosed prematurely (i.e. below the age of 55 for males and 65 for females) with CAD.
- Smoking was indicated by a mention of patient having smoked within the past year.

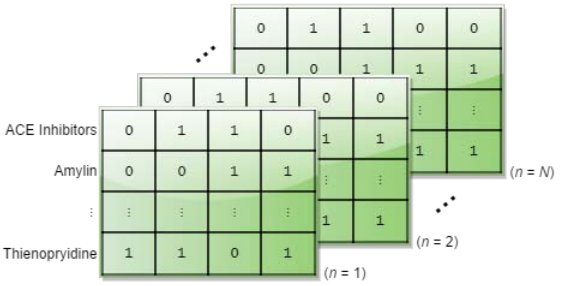
In addition to the seven risk factors, the discharge summaries were also annotated with medications prescribed for each patient which were related to diabetes, CAD, hyperlipidemia, hypertension, or obesity. The exact risk factors targeted by each medication were not explicitly annotated. In total, 22 medications and medication types were considered: (1) ACE inhibitors, (2) Amylin, (3) anti-diabetes medications, (4) ARB, (5) Aspirin, (6) beta blockers, (7) calcium channel blockers, (8) diuretics, (9) DPP4 inhibitors, (10) Ezetimibe, (11) Fibrates, (12) GLP1 agonis, (13) Insulin, (14) Meglitinides, (15) Metformin, (16) Niacin, (17) Nitrates, (18) anti-obesity medications, (19) Statin, (20) Sulfonylureas, (21) Thiazolidinedione, and (22) Thienopryidine.

Each risk factor and medication was associated with a particular *temporal signal*, indicating whether the risk factor or medication was *present* at the creation time of the discharge summary, *after* the creation time of the summary, *before* the creation time of the summary, or *during* the entire duration of the summary. As reported in [8], 89% of all risk factors and medications annotated were labeled with both *present* and *during* temporal signals. For this reason, we discarded any risk factors and medications annotated as occurring only *after* or *before* the timestamp of the discharge summary.

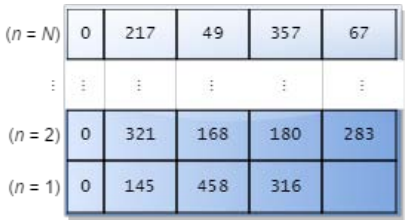
In this work, we considered the gold-standard annotations of risk factor. However, the i2b2/UTHealth shared task created for these annotations result in 49 submissions by 20 teams aiming to automatically recognize and classify which portions of the text correspond to these particular risk factors and medications [1]. Thus, our experiments can be easily replicated on new data-sets by relying on the automatic risk factor and medication recognition systems developed for the task, such as the first-place system developed by the National Library of Medicine [9], the second-place system developed by the Harbin Institute of Technology Shenzhen Graduate School [10], or third-place Kaiser Permanente system [11].



(a) The risk tensor $\mathcal{R}$



(b) The medication tensor $\mathcal{M}$



(c) The elapsed time matrix $\mathcal{E}$

Figure 1: Visualization of the (a) Risk Factor Tensor $\mathcal{R}$, (b) Medication Tensor $\mathcal{M}$ and (c) Elapsed Time matrix $\mathcal{E}$ with slices shown for the individual patients 1, 2, and N. In $\mathcal{R}$ and $\mathcal{M}$, each slice corresponds to a patient (n), each row corresponds to a risk factor or medication (respectively), and each column refers to the index of the corresponding discharge summary (i). In $\mathcal{E}$, each row refers to a patient (n), and each column refers to the index of the associated discharge summary (i).

Methods

In order to automatically model the interactions between the risk factors and medications in EHRs, we define a probabilistic model. This model operates by discovering latent *trends* in the way in which risk factors and medications changed in successive discharge summaries from a collection of patient EHRs. Because this model relies on the trends present in a particular dataset, we will first describe how to pre-process a collection of

longitudinal EHRs by extracting the clinical chronologies and encoding them into mathematical structures. Then, we will describe a probabilistic model over these data structures and demonstrate how the model can be used to (1) apply these latent trends to the chronology of a new patient in order to predict how his or her risk factors might progress, to (2) infer the interactions between pairs of risk factors, or between risk factors and medications over time, and to (3) identify patients whose risk factors and medications have an irregular progression.

Representing Clinical Chronologies

In order to model a collection of EHRs, we first define the following parameters which characterize the data:
$N =$ the number of patients in the EHR collection,
$L_n =$ the number of discharge summaries associated with patient n in the data, i.e., the *length* of the patient's chronology,
$V =$ the number of possible risk factors our model should consider, i.e., the size of the risk factor vocabulary,
$U =$ the number of possible medications and medication types our model should consider, i.e., the size of the medication lexicon.

In using the 2014 i2b2/UTHealth dataset, a total of $N = 128$ patients were used to train our model. Each of these patients was associated with $L_n \in [3,5]$ discharge summaries which were chronically ordered according to their timestamps. In our experiments, we considered the annotated risk factors and medications described in the previous section; thus, $V = 7$ and $U = 22$.

Given these parameters, we were able to represent the clinical chronologies of all patients in the data by defining three mathematical structures, which for each patient n, encode the set of risk factors and medications which were indicated during the i-th discharge summary (and i ranges from 1 to L_n for each patient):

$$\mathcal{R} = \left\{ R_{n,v,i} \in \{0,1\}^{N \times V \times L_n} \right\} \quad \mathcal{M} = \left\{ M_{n,u,i} \in \{0,1\}^{N \times U \times L_n} \right\} \quad \mathcal{E} = \left\{ E_{n,i} \in \mathbb{R}^{+N \times L_n} \right\}$$

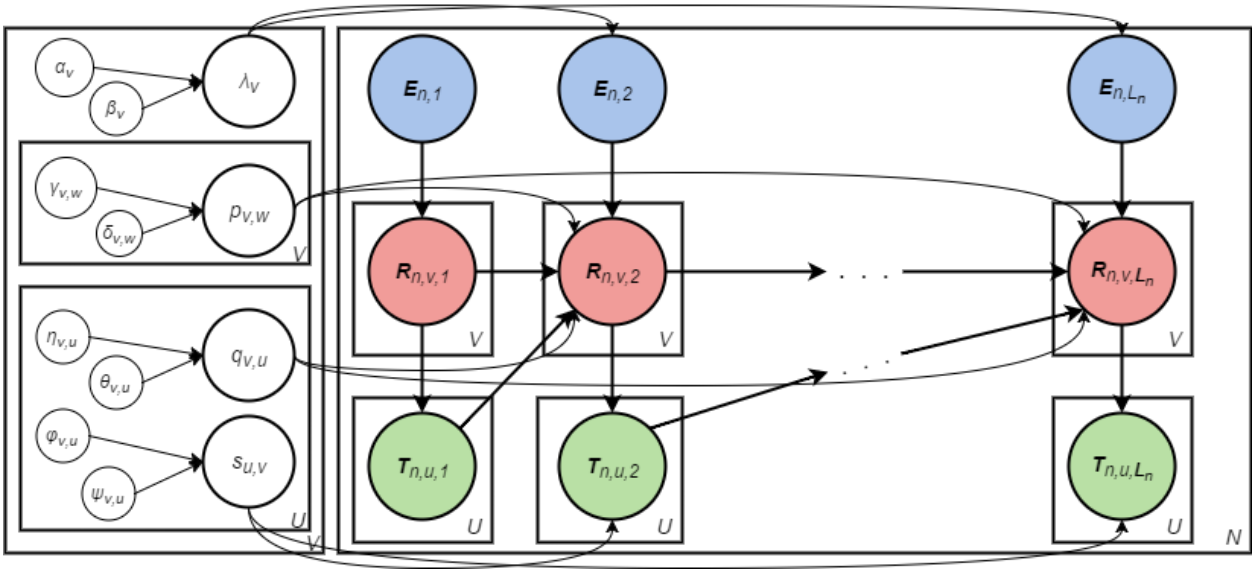
where $R_{n,v,i}$ is an entry in the 3rd-order risk factor tensor[5] $\mathcal{R}$ which indicates whether the v-th risk factor was mentioned in the i-th discharge summary for patient n (we assigned a value of 1 when the v-th risk factor was mentioned, and 0 otherwise); $M_{n,u,i}$ is an entry in the 3rd-order medication tensor $\mathcal{M}$ which indicates whether the u-th medication was mentioned in the i-th discharge summary for patient n (we assigned a value of 1 when the u-th medication or medication type was mentioned, and 0 otherwise); and $E_{n,i}$ is an entry in the elapsed time matrix $\mathcal{E}$ which stores the number of days elapsed between discharge summary i and the previous discharge summary, $i-1$, for patient n. Note that the elapsed time for the first discharge summary for each patient is defined as zero, i.e., $E_{n,0} = 0$. Figure 1 illustrates slices from the risk factor tensor $\mathcal{R}$, the medication tensor $\mathcal{M}$ and rows from the elapsed time matrix $\mathcal{E}$ which show the clinical chronology for individual patients. As illustrated, $\mathcal{R}$, $\mathcal{M}$ and $\mathcal{E}$ are all *jagged* structures, meaning that the number of discharge summaries associated with each patient (n) may vary according to the value of L_n. In this way, not only have we accounted for the de-identification of EHR timestamps, but we can directly discover temporal patterns based on the relative time elapsed between successive discharge summaries for each patient.

Modelling Chronological Interactions

Using the mathematical representation of patient chronologies obtained from a particular dataset, we would like to discover patterns in how the risk factors and medications interacted over time. We accomplished this by constructing a probabilistic graphical model (PGM) [7], which can be viewed as a generalization of a traditional mixture model which allows us to directly encode the *dependencies* between risk factors and medications over time. Probabilistic Graphical Models, like mixture models, operate on a set of statistical random variables and allow us to efficiently compute the joint distribution of these variables, from which any desired probability can be derived (e.g. conditional probabilities, prior probabilities, etc). In order to exploit the latent statistical information present in the data, our model bust be able to encode any arbitrary patient's clinical chronology. To do this, we define a binary random variable for each entry in the risk factor tensor $R_{n,v,i}$ and each entry in the medication tensor $M_{n,u,i}$, as well as a continuous random variable for each entry in the elapsed time matrix $E_{n,i}$. Thus, the joint distribution over these variables captures the likelihood of observing any possible clinical chronology which may be associated with a patient. Figure 2 illustrates this model using standard *plate notation*, wherein each shaded circle denotes an observable variable, each edge represents a statistical dependency, and plate (rectangular box) indicates that all variables contained in the box are copied or duplicated as many times as indicated by the quantity in the bottom-right of the plate For example, the binary variable indicated by

[5] A k-th order tensor is the k-dimensional analogue of a mathematical vector.

Figure 2: A Probabilistic Graphical Model of Patient Chronologies.



$R_{n,v,1}$ is duplicated for each risk factor $v \in [1,V]$ and each patient $n \in [1,N]$. The opaque variables in the left plate correspond to latent statistical parameters which will be inferred from the data, while each shaded column captures the elapsed time, risk factors, and medications which were present and absent in each discharge summary. As shown, the risk factors in each discharge summary are influenced by (1) the time elapsed since the previous discharge summary, (2) the risk factors present in the previous discharge summary, and well as (3) the medications mentioned in the previous discharge summary, while the medications depend only on the risk factors observed in the same discharge summary. In order to define the full joint distribution, we must formally define each of these four dependencies probabilistically.

We encode the fact that the presence of a particular risk factor $v \in [1,V]$ in discharge summary i for patient n is likely depend on the amount of time elapsed $E_{n,i}$ since the previous discharge summary, by defining an Exponential distribution for each possible risk factor:

$$P(R_{n,v,i}|E_{n,i}) \approx Exponential(E_{n,i};\lambda_v) = \lambda_v e^{-\lambda_v E_{n,i}} \tag{1}$$

where $\lambda_v \sim Gamma(\lambda_v;\alpha_v,\beta_v)$ is the parameter of the exponential distribution over elapsed times associated with risk factor v; α_v is the number of patient chronologies with v; and β_v is the sum of elapsed times associated with discharge summaries mentioning v. Thus, Equation 1 states that the likelihood of a particular risk factor given an arbitrary elapsed time follows an Exponential distribution unique to that particular risk factor.

In addition to the elapsed time, the risk factors in discharge summary i are influenced by the risk factors in the previous discharge summary $i-1$. For example, if a patient is diagnosed with diabetes in discharge summary i, it is very likely that diabetes will be observed in the $(i+1)$-th discharge summary. To represent this type of positive correlation, we define a *Noisy-Or* distribution for each risk factor, v. This enforces the semantics that risk factor v could have been triggered by each risk factor w observed in the previous discharge summary with some probability $(p_{v,w})$. Moreover, the Noisy-Or distribution states that the likelihood of a risk factor being present increases with each additional risk fact which was present during the previous discharge summary. Thus, the likelihood of risk factor v being present in discharge summary i given the presence or absence of each risk factor in the previous discharge summary $(i-1)$ is:

$$P(R_{n,v,i}|R_{n,1,i-1},...R_{n,V,i-1};p_{v,1},...p_{v,V},k_v) = 1 - \prod_{w=1}^{V} \begin{cases} 1-p_{v,w}, & \text{if } R_{n,w,i-1}=0 \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

Equation 2 states that the likelihood of risk factor $R_{n,v,i}$ given previous risk factors $R_{n,w,i-1}$ for $w \in [1,V]$ follows the Noisy-Or distribution parametrized by $p_{v,w}$ encoding the likelihood that the presence of risk factor w in the previous discharge summary can predict the presence of risk factor v in the current discharge summary. We can calculate the value $p_{v,w} \sim Beta(\gamma_{v,w},\delta_{v,w})$ by defining v as the number of patient chronologies wherein risk factor v was *present* in a discharge summary immediately following a discharge in which the risk factor w was present and $\delta_{v,w}$ as the number of patient chronologies wherein risk factor v

was *absent* in a discharge summary immediately following a discharge in which the risk factor w was present.

The final indicator for the presence of a risk factor v in discharge summary i for patient n is the set of medications prescribed to the patient in the previous discharge summary. This follows the intuition that the previous prescription of a medication can prevent the presence of targeted risk factors, and captures the negative correlation between medications and risk factors. To model this, we utilize an inverted (i.e. $(1-p)$) Noisy-Or distribution for each risk factor v which states that *absence* of risk factor v can be predicted based on the presence of each medication u in the previous discharge summary with some probability $q_{v,u}$:

$$P(R_{n,v,i}|M_{n,1,i-1},...M_{n,U,i-1};q_{v,1},...q_{v,U},a_v) = \prod_{u=1}^{U} \begin{cases} 1-q_{v,u}, & \text{if } M_{n,u,i-1}=0 \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

Equation 3 defines the probability of observing risk factor $R_{n,v,i}$ despite each medication prescribed in the previous discharge summary. We can estimate the probability $q_{v,u} \sim Beta(\eta_{v,u},\theta_{v,u})$ by defining $\eta_{v,u}$ as the number of patient chronologies in which risk factor $R_{n,v,i}$ was present following a discharge summary in which medication u was prescribed, and $\theta_{v,u}$ as the number of discharge summaries in which $R_{n,v,i}$ was absent following a discharge summary in which medication u was prescribed.

Together, Equations 1 through 3 capture the statistical dependencies governing the presence or absence of each risk factor. However, the presence of a risk factor can also influence the set of medications which are prescribed during the same discharge summary. Consider, for example, the intuition that many medications are only prescribed after certain risk factors have been diagnosed. To represent this type of interaction, we employ a *Noisy-And* distribution for each medication u which assumes the presence of each medication mentioned in a discharge summary depends on one or more risk factors being mentioned in the same discharge summary. Moreover, the Noisy-And distribution states that as the number of diagnosed risk factors decreases, so must the probability of each medication. Mathematically, this has the form:

$$P(M_{n,u,i}|R_{n,1,i},...R_{n,V,i};s_{u,1},...s_{u,V},b_u) = \prod_{v=1}^{V} \begin{cases} 1-s_{u,v}, & \text{if } R_{n,v,i}=0 \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

where $s_{u,v}$ indicates the probability that medication u requires risk factor v to be diagnosed. As with Equations 2 and 3, we estimate the probabilities $s_{u,v} \sim Beta(\phi_{u,v},\psi_{u,v})$ for $u \in [1,U]$ and $v \in [1,V]$ by defining $\phi_{u,v}$ as the number of patient chronologies in which medication u was prescribed in the same discharge summary in which risk factor v was present, and $\psi_{u,v}$ as the number of patient chronologies in which medication u was not prescribed in the same discharge summary in which risk factor v was present.

Using these four equations, we can define the joint probability of observing any possible patient chronology as:

$$\begin{aligned} P(\mathcal{E},\mathcal{R},\mathcal{M};\Theta) = & \prod_{n=1}^{N}\prod_{v=1}^{V} k_v \prod_{u=1}^{U} P(M_{n,u,1}|R_{n,1,1}...R_{n,V,1}) \\ & \prod_{i=2}^{L_n}\prod_{v=1}^{V} P(R_{n,v,i}|R_{n,1,i-1}...R_{n,V,1}) P(R_{n,v,i}|E_{n,i}) \prod_{u=1}^{U} P(M_{n,u,i}|R_{n,1,i}...R_{n,V,i}) \end{aligned} \tag{5}$$

where Θ refers all the latent variables in our model, i.e., λ_v, α_v, β_v, $\eta_{v,w}$, $\theta_{v,w}$, $\gamma_{v,u}$, $\delta_{v,u}$, $\psi_{u,v}$, and $\phi_{u,v}$. Thus, Equation 5 represents the joint distribution in terms of products of the previously defined condition distributions given in Equations 1 to 4, and allows us to determine any arbitrary probability involving these variables by appealing to the basic laws of probability.

Discovering the Latent Interactions of Risk Factors and Medications

The probabilistic model representing Equation 5 encodes multiple types of interactions between risk factors and medications. The first latent interaction, as characterised by Equation 2, shows the positive correlation or causal relationship between each pair of risk factors, v, and w in successive discharge summaries through the latent variable $p_{v,w}$. The second interaction, defined through Equation 3, captures the negative correlation or inhibiting relationship between each medication, u, and each risk factor v in the latent variable $q_{v,u}$. The final interaction, embodied in Equation 4, captures the associative strength or enabling relationship between each risk factor v and each medication u with the latent variable $s_{u,v}$. We learned these variables by applying a straight-forward collapsed Gibbs sampler, as described in [12, 13] using the definitions provided in [14] and in [15].

Predicting Patient Outcomes from their Histories

After discovering the latent interactions implied by the latent variables in our model, we are able to predict the clinical outcomes (risk factors) for a new patient by determining the likelihood of each possible risk factor $v \in [1..V]$. To enable such a prediction, we most perform three steps: (1) encode the patient's history using binary random variables so that we can leverage our probabilistic model, (2) use the joint probability to predict how the patient's observations may progress.

We can encode the clinical chronology for a new patient $\hat{p}$ in a similar manner to the way we represented the clinical chronologies pertaining to the original set of patients in our dataset. Let $\hat{L}$ represent the number of longitudinal discharge summaries for patient $\hat{p}$. This allows us to define $\hat{\mathcal{R}}_{v,i} \in \{0,1\}^{V \times \hat{L}}$ to be the risk factor matrix, $\hat{\mathcal{M}}_{u,i} \in \{0,1\}^{U \times \hat{L}}$ to be the medication matrix, and $\hat{\mathcal{E}}_i \in R^{\hat{L}}$ to be the elapsed time vector. After sorting the discharge summaries for the patient in ascending chronological ordering (according to their timestamps), we can set the value of $\hat{\mathcal{R}}_{v,i}$ to 1 when risk factor v was mentioned in the i-th discharge summary, and 0 otherwise. Likewise, we can set $\hat{\mathcal{M}}_{u,i}$ to 1 when medication u was mentioned in the i-th discharge summary, and 0 otherwise. Finally, we assign to $\hat{\mathcal{E}}_i$ the elapsed time in days between discharge summary i and the previous discharge summary, $i-1$ where $\hat{\mathcal{E}}_1$ is set to 0. In this way, we have defined the risk factor and medication matrices as well as the elapsed time vector in the same way that we defined each slice of the risk factor and medication tensor and each row of the elapsed time matrix generated for our original dataset.

This representation allows us to predict clinical outcomes for the patient by constructing latent variables $x_1,...,x_V$ indicating the presence or absence of each risk factor $v \in [1,V]$ and by defining y to be the time elapsed from the last discharge summary in the patients chronology. To accomplish this, we compute the maximum a posteriori (MAP) assignment for each variable x_v:

$$\begin{aligned} \hat{x}_v &= \underset{x' \in \{0,1\}}{\operatorname{argmax}} \frac{P(\hat{\mathcal{R}}, \hat{\mathcal{E}}, \hat{\mathcal{M}}, x_1, ..., x_V, y; \Theta)}{P(\hat{\mathcal{R}}, \hat{\mathcal{E}}, \hat{\mathcal{M}}; \Theta)} \\ &= \underset{x' \in \{0,1\}}{\operatorname{argmax}} \prod_{v=1}^{V} P(\hat{x}_v = x' | P(\hat{\mathcal{R}}_{1,\hat{L}}, ... \hat{\mathcal{R}}_{V,\hat{L}}; p_{v,1} ... p_{v,V}) \end{aligned} \tag{6}$$

In this way, Equation 6 allows us to predict whether each risk factor $v \in [1..V]$ will be present or absent given the clinical chronology for the patient according to the latent interaction variables (Θ) discovered for our dataset. This technique could also be easily extended to predict the presence or absence of observations between discharge summaries – for example during long gaps in the patient's history.

Identifying Irregular Patients

Another potential application of the model arises when one wants to identify patients whose clinical chronologies are unlike a particular patient population. This allows for down stream clinical decision support systems to monitor patients who may present with unusual risk factors or disease progressions. To identify such patients, the model must be first initialized on some dataset which does not already include the patient (that is, the patient's EHR must be removed or ignored in the dataset when learning the latent parameters). Then, let $\hat{L}$ represent the number of longitudinal discharge summaries for the target patient $\hat{p}$. This allows us to define $\hat{\mathcal{R}}_{v,i} \in \{0,1\}^{V \times \hat{L}}$ to be the risk factor matrix, $\hat{\mathcal{M}}_{u,i} \in \{0,1\}^{U \times \hat{L}}$ to be the medication matrix, and $\hat{\mathcal{E}}_i \in R^{\hat{L}}$ to be the elapsed time vector. After sorting the discharge summaries for the patient in ascending chronological ordering (according to their timestamps), we can set the value of $\hat{\mathcal{R}}_{v,i}$ to 1 when risk factor v was mentioned in the i-th discharge summary, and 0 otherwise. Likewise, we can set $\hat{\mathcal{M}}_{u,i}$ to 1 when medication u was mentioned in the i-th discharge summary, and 0 otherwise. Finally, we assign to $\hat{\mathcal{E}}_i$ the elapsed time in days between discharge summary i and the previous discharge summary, $i-1$ where $\hat{\mathcal{E}}_1$ is set to 0. This allows us to determine how likely patient $\hat{p}$'s chronology is by simply computing the joint probability of that patient's chronology using Equation 5, based on the latent variables (Θ) discovered from the training corpus.

Results

In our experiments, we relied on the collection of annotated longitudinal EHRs provided in the 2014 shared task on *Challenges in Language Processing on Clinical Data* sponsored by the 2014 Informatics for Integrating Biology and the Bedside (i2b2) and the University of Texas Health Science Center at Houston (UTHealth) created with the purpose of fostering the development of automatic systems for detecting clinical findings, medications, and temporal signals. We re-purpose this data in order to learn and evaluate our model of clinical histories. That said, for the sake of consistence and reproducibility, we report our performance using

Table 1: Predictive performance individual risk factors, as well as the micro-average over all risk factors.[6]

Risk Factor	ACC	PPV	FNR	FPR	TNR	TPR	F1	TP	FP	FN	TN
Obesity	0.864	1.0	0.941	0.0	1.0	0.058	0.111	1	0	16	101
Hypertension	0.958	0.958	0.0	1.0	0.0	1.0	0.978	113	5	0	0
Diabetes	0.788	0.812	0.115	0.4	0.6	0.885	0.847	69	16	9	24
Hyperlipidemia	0.729	0.663	0.0	0.582	0.482	1.0	0.797	63	32	0	23
CAD	0.746	0.485	0.448	0.191	0.809	0.551	0.516	16	17	13	72
Micro-average	0.794	0.617	0.172	0.221	0.779	0.828	0.707	735	456	153	1606

the same training and testing partitions given by the i2b2/UTHealth organizers. Note: that was also evaluated our model using 10-fold cross validation; for the sake of brevity, these results are not reported in this paper because the difference in performance was statistically insignificant ($p=0.04$). Using this partitioning, our training set consisted of EHRs documenting the progression of heart disease for 178 patients, and our testing set consisted of EHRs for 118 patients.

In order to evaluate the predictions enabled by our model, we cast the problem of predicting the presence or absence of risk factors as a *binary classification* problem. However, our evaluation had to consider that each discharge summary was associated with multiple risk factors. Thus, we leveraged the experimental methodology used for evaluating multi-label classification problems in the machine learning community [16]. After training our the latent variables in our model using the clinical chronologies extracted on the 168 patients in the training set, we evaluated the accuracy of our model in predicting the risk factors present and absent in the last discharge summary, given all the preceding discharge summaries for each patient. Specifically, for each patient n with an EHR containing L_n chronologically ordered discharge summaries, we used our trained model to predict the presence or absence of each risk factor $v \in [1,V]$ given the chronology in the first (L_n-1) discharge summaries as well as the amount of time elapsed since the $(L-1)$-th discharge summary (i.e. $\mathcal{E}_{n,L_n}$). Then, we compared the predicted presence or absence of each risk against the actual values extracted from the L-th discharge summary. Formally, we considered a predicted risk factor as a *true positive* (**TP**) if it was predicted by the model and was present in the discharge summary, as a *false positive* (**FP**) if it was predicted by the model but was absent in the discharge summary, as a *false negative* (**FN**) if it was not predicted by the model but was present in the discharge summary, and as a *true negative* (**TN**) if was not predicted by the model and was absent in the discharge summary. This allowed us to compute a variety of performance measures, such as the Accuracy (**Acc.**,the Positive Predictive Value (**PPV**), the False Negative Rate (**FNR**), the False Positive Rate (**FPR**), the True Negative Rate (**TNR**), the True Positive Rate (**TPR**), and the F_1 Measure ($\frac{2TP}{2TP+FP+FN}$). Table 1 presents these results[6]. Overall performance was high, although certain classes (such as Hyperlipidemia) proved more difficult than others (e.g. Hypertension). Note that because the F_1-measure considers only true positive (and not true negative) labels, the performance of the overwhelmingly absent risk factor *Obesity* is better assessed by the accuracy measure. Interestingly, despite the entire patient cohort having a diagnosis of Diabetes, a number of discharge summaries did not contain diagnoses of the disease, suggesting that either the condition was managed, or not of primary interest to the physician. We additionally compared our approach against a previously developed system. The baseline system, reported in [8] does not represent medications nor the elapsed time between successive discharge summaries which achieved a micro-average predictive accuracy of only 54.3%. The superior performance achieved by the model outlined in this paper demonstrates the importance of encoding the semantics in the types of interactions present in patient's clinical chronologies.

Discussion

In addition to the predictive performance, we also explored the latent interactions recovered by our model. Table 2 presents the likelihood that a risk factor w in a discharge summary i will positively predict the present of risk factor v in discharge summary $(i+1)$ (this corresponds to the latent variable $p_{v,w}$ in Equation 3). As shown by the probabilities in the diagonal, each risk factor is unsurprisingly potent at predict a recurrence of itself. More interestingly, the presence of any of the heart-disease-related risk factors is a strong predictor for the presence of Hypertension. The difference between the micro-average predictive performance, and the and correlations shown in Table 2 suggests that while individual risk factors may not be valuable for predicting other risk factors, but instead the complete set of risk factors present (and absent) at each discharge summary must be considered.

We also investigated the role each medication has in preventing each risk factor in the immediately following discharge summary, as shown in Table 3 (corresponding to the variable $q_{v,u}$ from Equation 4). Note, these results do not distinguish between situations in which the risk factor was absent in both adjacent discharge summaries and situations in which the risk factor was resolved.

[6]We have omitted the predictive performance for the risk factors *Smokes* and *Family History* because they rarely change and thus unduly inflate the micro-average performance of our model.

Table 2: Likelihood that each risk factors in a (current) discharge summary will positively predict each risk factor in a future discharge summary.

		Future				
		Obesity	**Hypertension**	**Hyperlipidemia**	**Diabetes**	**CAD**
Current	**Obesity**	99.624	98.496	78.947	84.586	62.782
	Hypertension	43.522	99.834	76.744	86.711	64.950
	Hyperlipidemia	44.776	98.507	99.787	86.994	65.245
	Diabetes	42.056	97.570	76.262	99.813	64.486
	CAD	41.646	97.506	76.309	86.035	99.751
	FamilyHistory	43.160	97.883	76.221	86.971	65.147
	Smoker	43.160	97.883	76.221	86.971	65.147

Table 3: Likelihood of each medication preventing each risk factor in the immediately following discharge summary.

		Risk Factor				
		Obesity	**Hypertension**	**Hyperlipidemia**	**Diabetes**	**CAD**
Medication	**ARB**	56.627	0.602	18.072	9.036	33.133
	beta_blocker	56.794	1.742	22.997	12.718	33.449
	metformin	46.642	1.493	16.418	1.493	40.672
	diuretic	42.188	0.521	19.792	23.438	44.792
	aspirin	55.109	1.277	22.445	14.051	31.204
	statin	54.696	2.394	19.705	12.707	34.991
	sulfonylureas	53.394	1.810	20.362	1.810	38.009
	thienopyridine	63.492	3.704	22.222	12.698	22.751
	calcium_channel_blocker	52.222	0.370	21.111	9.259	36.667
	ACE_inhibitor	54.265	1.659	20.853	12.322	35.071
	insulin	54.276	4.276	28.618	0.329	39.145
	nitrate	59.917	1.653	21.074	16.529	12.397
	thiazolidinedione	37.975	1.266	10.127	1.266	31.646
	fibrate	40.909	2.273	2.273	2.273	36.364
	niacin	94.118	5.882	5.882	52.941	5.882
	ezetimibe	75.000	5.000	5.000	5.000	20.000
	anti_diabetes	16.667	16.667	16.667	16.667	16.667

Table 4: Likelihood of each medication being prescribed for each risk factor in the same discharge summary

		Risk Factor				
		Obesity	**Hypertension**	**Hyperlipidemia**	**Diabetes**	**CAD**
Medication	**ARB**	0.433	1.000	0.819	0.914	0.671
	beta_blocker	0.431	0.984	0.770	0.874	0.665
	metformin	0.531	0.988	0.840	0.988	0.592
	diuretic	0.576	1.000	0.804	0.767	0.551
	aspirin	0.447	0.989	0.776	0.861	0.688
	statin	0.452	0.977	0.803	0.874	0.649
	sulfonylureas	0.463	0.986	0.799	0.986	0.618
	thienopyridine	0.364	0.967	0.781	0.876	0.773
	calcium_channel_blocker	0.477	1.000	0.791	0.910	0.634
	ACE_inhibitor	0.457	0.985	0.793	0.880	0.649
	insulin	0.456	0.959	0.714	1.000	0.608
	nitrate	0.399	0.987	0.789	0.838	0.880
	thiazolidinedione	0.616	1.000	0.909	1.000	0.687
	fibrate	0.593	1.000	1.000	1.000	0.648
	niacin	0.000	1.000	1.000	0.474	1.000
	ezetimibe	0.217	1.000	1.000	1.000	0.826
	anti_diabetes	1.000	1.000	1.000	1.000	1.000

As observed, Niacin and Ezetimibe were the best predictors of the absence of Obesity. This shows that modelling the set of medications can improve the ability of the model to negatively predict future risk factors.

Finally, we analysed the association between each medication and each risk factor (corresponding to $s_{u,v}$ from Equation 5); that is, we present the strength of the recovered probability that medication u would be prescribed to a patient presenting with risk factor v. These results are listed in Table 4. Unsurprisingly, given that our dataset is a cohort of diabetic patients, each patient was taking at least one anti-diabetes medication. More interestingly, our model was able to recover that fibrates, niacins, and ezetimibes are used to treat Hyperlipidemia. Moreover, by normalizing these values according to the average for each risk factor, and the average for each medication (i.e. by calculating the point-wise mutual information), we observed that our model was able to recover additional interactions, such as Aspirin being prescribed for CAD, and Metfornin being associated with Diabetes.

Conclusion

We designed a data-driven probabilistic graphical model for risk factors and medications interact through patient's clinical chronologies. This model operates by first learning the latent interactions between successive pairs of risk factors and medications using semantically motivated probability distributions. These latent variables, in-turn, allow us to (1) predict the way a new patient's clinical chronology might progress as well as to (2) identify patients who clinical chronology is progressing unusually given some cohort of similar patients. We evaluated the individual risk factor and micro-average performance when predicting how a patient's risk factors progressed, compared to the actual risk factors mentioned in their EHR. Experiments demonstrated an accuracy up to 95.8% for a single class, and a micro-average accuracy of 81.6%, illustrating the potential of our model for predicting personalized patient outcomes from longitudinal EHRs. Moreover, we presented and analysed the interactions discovered from the 2014 i2b2/UTHealth collection of diabetic patients' EHRs. Future performance may be improved by (1) normalizing the statistical information informing each latent variable in the model, (2) leveraging larger EHR collections, and (3) employing more sophisticated inference techniques.

Acknowledgements

This work was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Stubbs A, Kotfila C, Xu H, Uzuner O. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. Journal of biomedical informatics. 2015;.

[2] Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Medical care. 2010;48(11):981–988.

[3] Vickers AJ. Prediction models in cancer care. CA: a cancer journal for clinicians. 2011;61(5):315–326.

[4] Bejan CA, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Assessing Pneumonia Identification from Time-Ordered Narrative Reports. AMIA Annual Symposium Proceedings. 2012 Nov;2012:1119–1128. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540463/.

[5] Huopaniemi I, Nadkarni GN, Nadukuru R, Ellis SB, Gottesman O, Bottinger E. Disease progression subtype discovery from longitudinal EMR data with a majority of missing values and unknown initial time points. In: AMIA Annual Symposium Proceedings. vol. 2014; 2014. .

[6] Wang X, Wang F, Hu J, Sorrentino R. Exploring joint disease risk prediction. AMIA Annual Symposium proceedings. 2014;2014:1180–1187.

[7] Koller D, Friedman N. Probabilistic graphical models: principles and techniques. MIT press; 2009.

[8] Goodwin T, Harabagiu SM. A Probabilistic Reasoning Method for Predicting the Progression of Clinical Findings from Electronic Medical Records. In: AMIA Joint Summit proceedings. vol. 2015. San Francisco, California.; 2015. .

[9] Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. Journal of Biomedical Informatics. 2015;58:S111–S119.

[10] Chen Q, Li H, Tang B, Liu X, Liu Z, Liu S, et al. Identifying risk factors for heart disease over time–HITSZ's system for track 2 of the 2014 i2b2 NLP challenge. In: Seventh i2b2 Shared Task and Workshop: Challenges in Natural Language Processing for Clinical Data; 2014. .

[11] Torii M, Fan Jw, Yang Wl, Lee T, Wiley MT, Zisook DS, et al. Risk factor detection for heart disease by applying text analytics in electronic medical records. Journal of biomedical informatics. 2015;.

[12] Liu JS. The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem. Journal of the American Statistical Association. 1994 Sep;89(427):958–966.

[13] Porteous I, Newman D, Ihler A, Asuncion A, Smyth P, Welling M. Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '08. New York, NY, USA: ACM; 2008. p. 569–577.

[14] Pearl J. Probabilistic reasoning in intelligent systems: Networks of plausible reasoning. Morgan Kaufmann Publishers, Los Altos; 1988.

[15] Friedman N, Murphy K, Russell S. Learning the structure of dynamic probabilistic networks. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc.; 1998. p. 139–147.

[16] Tsoumakas G, Katakis I. Multi-label classification: An overview. Int J Data Warehousing and Mining. 2007;2007:1–13.