Constructing Disease Network and Temporal Progression Model via Context-Sensitive Hawkes Process

Edward Choi, Nan Du, Robert Chen, Le Song and Jimeng Sun
School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, USA
Email: {mp2893, dunan, rchen87}@gatech.edu, {lsong, jsun}@cc.gatech.edu

***Abstract*—Modeling disease relationships and temporal progression are two key problems in health analytics, which have not been studied together due to data and technical challenges. Thanks to the increasing adoption of Electronic Health Records (EHR), rich patient information is being collected over time. Using EHR data as input, we propose a multivariate context-sensitive Hawkes process or *cHawkes*, which simultaneously infers the disease relationship network and models temporal progression of patients. Besides learning disease network and temporal progression model, *cHawkes* is able to predict when a specific patient might have other related diseases in future given the patient history, which in turn can have many potential applications in predictive health analytics, public health policy development and customized patient care. Extensive experiments on real EHR data demonstrate that *cHawkes* not only can uncover meaningful disease relations and model accurate temporal progression of patients, but also has significantly better predictive performance compared to several baseline models.**

***Keywords*—Point Process; Hawkes Process; Health Analytics; Disease Relation; Disease Prediction; EHR;**

I. Introduction

Applying automatic computational/statistical approaches to medical fields has attracted much attention from the communities of both academia and industry in the era of Big Data [1]. Such popularity has been spurred by the introduction of Electronic Health Records (EHR). EHR contains temporal event sequences such as admission time, discharge time, sex, ethnicity, age, weight, diagnoses, procedures, and medications. Recently, there have been a number of studies that tried to utilize such data [2]–[7] for different purposes, such as disease progression modeling [7], phenotyping [3], [5] and mortality modeling [6].

There are two key problems in health analytics that are particularly challenging, namely,

- **disease relation discovery**: What are the temporal relationships between diseases?
- **temporal progression model**: How do different diseases progress over time for each individual patient?

In order to model temporal relations among diseases for a diverse patient population, we propose context-sensitive Hawkes Process model *cHawkes*. The classical Hawkes Process [8], a point process to model a finite set of temporal events, considers the relations between all past events and the current event, namely, how past events will affect the chance of the current event happening. In our setting, Hawkes Process can capture the fact that a person who recently visited a hospital for hypertension has a higher chance of having heart failure than a person who has never suffered from hypertension. However, the classical Hawkes Process does not consider the context of each patient's specific diagnosis. For instance, in our setting, the relation between hypertension and heart failure is applied to all people regardless of their physical differences. This is far from being realistic, since a person's age, weight and many other factors might affect the chance of having hypertension or heart failure. It is plausible to think that heart failure is more likely to follow hypertension for an obese person than an average person. Therefore, *cHawkes* identifies disease relations and temporal progression while still capturing the personal physical differences of patients using multivariate context-sensitive Hawkes process.

The contributions of the paper include the following :

- We propose *cHawkes*, a context-sensitive Hawkes Process to simultaneously model disease relationship network and temporal progression using EHR data.
- *cHawkes* captures both global interacting relations among diseases and how the characteristic of individual patients affect the occurrence of diseases.
- *cHawkes* generates sparse and interpretable models through regularization.
- We discover clinically meaningful disease relationship network by applying *cHawkes* on real EHR datasets and demonstrate accurate risk predictions for individual patients using the proposed model.

The rest of the paper is organized as follows: After we survey the related work in section II, we briefly review the general Hawkes Process in section III. We then describe in section IV, our proposed Context-Sensitive Hawkes Process and parameter learning, including regularization. In section V, we conduct extensive qualitative and quantitive evaluations on a real-world dataset, MIMIC II. We conclude in section VI with future research directions.

II. Related Work

Disease Relationship Discovery Studies have been conducted on disease relations discovery [9]–[11] through analyzing how past diseases can affect the occurrence of current diseases. Most of the studies, however, has limitations in dealing with time dimension. Leiva-Murillo et al. [9] applied continuous-time Hidden Markov Model (HMM) to capture disease relations. However, their model only captures the influence from the most recent event, as they use a first-order HMM. Savova et al. [10] used natural language processing algorithms to extract temporal relations between disease occurrences. Their work, however, uses free-text which cannot capture the exact duration between the events. A recent study by Zhao et al. [11] tries to learn the triggering kernels of the Hawkes Process in order to study the disease relations, and also proposes a metric termed 'Individual Physique' to represent a person's natural fitness. The major difference to our work is that we utilize concrete

features of patients that change over time (e.g. weight, age) rather than represent the natural fitness of a person as a single constant value. Moreover, our method uncovers the general latent disease network by appropriate regularizations, which produces clinically meaningful sparse structures as verified in the experiments.

Temporal Progression Model A number of recent studies tried to model temporal aspect of patients and their diseases [7], [12], [13]. Most of the studies, however, focus on predicting the progression of a specific disease. Zhou et al. [12] focuses on modeling the progression of Alzheimer's disease using biomarkers. Liu et al. [13] captured the functional and structural degeneration in the glaucoma process by using the 2-dimensional continuous-time Hidden Markov Model. Wang et al. [7] proposed a more general approach to model the progression of an arbitrary disease. They used unsupervised learning to analyze the comorbidities of chronic obstructive pulmonary disease (COPD) patients and predicts the progression of COPD. Since they model the progression of COPD through various stages along with its comorbidities, the performance of their work depends on the prior knowledge of the target disease and its comorbidities.

Network Diffusion Modeling Uncovering the relation between diseases shares similarities with network diffusion modeling, which has been actively studied recently [14]–[19]. Just as a patient experiencing multiple diseases forms a cascade, a tweet being retweeted by Twitter users forms a cascade. CONNIE [14] and NETINF [15] tried to infer the network connectivity with fixed transmission rates. NETRATE [16] and KernelCascade [17] employ a survival analysis approach for learning probabilistic transmission rates. More recently, MoNET [18] and TopicCascade [19] respectively used the features of nodes and the features of events to infer the transmission rates. TopicCascade, although similar to *cHawkes*, assumes the characteristics of events do not change over time. This assumption does not apply to patient modeling, as patient features do change over time. While there are similarities between *cHawkes* and network diffusion models, there are two big differences: 1. Network diffusion models are mainly interested in uncovering the hidden network structure while *cHawkes* performs disease relations, patient context, and risk prediction. 2. The transmission process in network diffusion models is influenced only by the most recent event. *cHawkes*, on the other hand, tries to capture the relation between diseases that did not occur consecutively.

III. Hawkes Process

Hawkes Process is a type of point processes for modeling temporal event sequences. The intuition behind Hawkes Process is *self-excitation*, meaning that the past occurrences of events make the future event more probable. More formally, a general joint likelihood of observing a single sequence of events $\mathcal{T} = \{t_1, \dots, t_n\}$($t_i$ is the i-th occurrence of the event) within the time window starting from $t = 0$ and ending at $T \geq t_n$, can be given as follows:

$$\mathcal{L}(\mathcal{T}) = \prod_{t_i \in \mathcal{T}} \lambda(t_i) \cdot \exp\left(-\int_0^T \lambda(\tau) d\tau\right), \tag{1}$$

The conditional intensity function $\lambda(t)$ in Equation 1 describes the behavior of a point process. For one-dimensional Hawkes Process, its conditional intensity function is given by

$$\lambda(t) = \mu + \alpha \sum_{t_i < t} g(t - t_i),$$

where μ is the base intensity rate capturing the spontaneous rate to generate new events, $g(t - t_i)$ is the triggering kernel quantifying the influence of a past event on the occurrence of the new event, and α measures the amount of influence from past events on the current event. The limitation of one-dimensional Hawkes Process is that it can only model a single type of event. In our setting, we want to capture the interacting processes of different event types (or diseases, more specifically). This is when the multi-dimensional Hawkes Process comes into play.

Multi-dimensional Hawkes Process models not only the self-excitation of a single type of event but also captures the mutual excitations among different types of events. Formally, we denote $\mathcal{T} = \{(t_i, d_i)\}_{i=1}^n$ an event sequence of time t_i associated with event type d_i. The conditional intensity function for each event type d is thus given by

$$\lambda_d(t) = \mu_d + \sum_{t_i < t} \alpha_{d,d_i} g(t - t_i) \tag{2}$$

where μ_d is the base intensity rate of event type d, and α_{d,d_i} is the strength of influence event type d_i has over event type d. Then the log-likelihood of observing $\mathcal{T}$ is the following :

$$\ell(\mathcal{T}) = \sum_{d=1}^{D} \left\{ \sum_{(t_i, d_i = d) \in \mathcal{T}} \log \lambda_d(t_i) - \int_0^T \lambda_d(\tau) d\tau \right\}, \tag{3}$$

where D is the total number of diseases.

So far, at the first glance, the multi-dimensional Hawkes Process seems able to model the inter-relations among diseases. However, the direct application to our setting will incur two major issues. First, the model is constructed for each disease, which ignores the physical difference of the patients. As a result, we cannot make any patient-specific risk predictions. Second, because the number of unknown parameters $\left\{\alpha_{d_j,d_i}\right\}_{i,j}$ grows quadratically as the number of diseases increases, even on moderate size of EHR data, the model will incur huge computation cost and is often overfitted.

IV. Context-Sensitive Hawkes Process

From EHR data, there are three modeling insights, namely, 1) patient's context-sensitive disease risk, 2) the relationship between various diseases as to how they influence the occurrence of one another, and 3) the temporal dynamics of diseases. To capture these insights from EHR, we propose Context-sensitive Hawkes Processes *cHawkes*.

A. Context-Sensitive Hawkes Process

We first denote by $\mathcal{T}^i = \left\{(t_j^i, d_j^i, \boldsymbol{f}_j^i)\right\}_{i=1}^n$ the sequence of clinical visits of patient i, where t_j^i is the time, d_j^i the type of disease, $\boldsymbol{f}_j^i$ the physical features(*e.g* age, weight, etc) of patient i at visit j. Each $\mathcal{T}^i$ is referred to as a *cascade* in the sense that for patient i, his (or her) current disease might trigger other related symptoms and diseases in the future. Overall EHR data are modeled as a collection $\mathcal{C}$ of *i.i.d.* cascades $\left\{\mathcal{T}^1, \dots, \mathcal{T}^{|\mathcal{C}|}\right\}$, one from each patient.

The general multi-dimensional Hawkes Process assumes that the spontaneous intensity rate μ_d and the mutual-excitation

rate $\alpha_{d,d'}$ are the same for all cascades. Since different patients have different sets of physical features, such as age, weight and height, it would be unrealistic to assume that the same μ_d and $\alpha_{dd'}$ can be applied to all patients.

To incorporate such patient contexts, we introduce a feature vector $\boldsymbol{f}_j^i$ for patient i at visit j, which can be parameterized based on information available in EHR data. For example, patient features can include discrete values such as ethnicity and gender, as well as real values such as age and weight. In our experiments, for simplicity all features are converted to discrete values as can be seen from the age and weight in Figure 2[1]. We now modify the conditional intensity function for the dth disease given patient i as follows.

$$\lambda_d^i(t) = \underbrace{\boldsymbol{\mu}_d^\top \boldsymbol{f}_j^i}_{\substack{\text{patient}\\ \text{context}}} + \sum_{t_j^i < t} \underbrace{\alpha_{d,d_j^i}}_{\substack{\text{disease}\\ \text{network}}} \underbrace{g(t - t_j^i)}_{\substack{\text{temporal}\\ \text{dynamics}}} \tag{4}$$

Note that what used to be $\lambda_d(t)$ in multi-dimensional Hawkes Process is now $\lambda_d^i(t)$, which is a function of disease type d given patient i. Essentially, we are learning intensity functions for individual patients. As shown in Equation 4, the model consists of the following three key components:

- **Patient context**: We formulate the spontaneous occurrence strength $\boldsymbol{\mu}_d$ as a linear combination of patient-specific, time-variant features $\boldsymbol{f}_j^i$. As a result, the conditional intensity function can now capture the heterogeneous evolving process of each specific patient.
- **Disease network:** We learn mutual excitation $\left\{\alpha_{d_j,d_i}\right\}$ variables for any pair of diseases to construct the disease relationship network. Although one might want to learn a patient specific disease network, the reality is that the available information from a single patient is often too limited to reliably learn all the parameters. As a result, we choose to learn a global disease relation network for all patients.
- **Temporal dynamics:** The triggering kernel $g(t)$ controls the aspect of temporal dynamics of diseases. Without loss of generality, in this work, we use the exponential decay kernel $g(t) = \lambda e^{-\lambda t}$, which is commonly used in many fields for simplicity. [2]

B. Parameter Estimation

By Equation 3, the log-likelihood $\ell(\mathcal{T}^i)$ of observing a single cascade $\mathcal{T}^i \in \mathcal{C}$ for patient i is given by

$$\ell(\mathcal{T}^i) = \sum_{d=1}^{D} \left\{ \sum_{(t_j^i, d_j^i = d, \boldsymbol{f}_j^i) \in \mathcal{T}^i} \left(\log \lambda_d^i(t_j^i) - \int_{t_{j-1}^i}^{t_j^i} \lambda_d^i(\tau) d\tau \right) - \int_{t_{n,d}^i}^{T} \lambda_d^i(\tau) d\tau \right\},$$

where $t_{n,d}^i$ is the last event on the dimension d. Then, the joint log-likelihood of observing all the cascades $\mathcal{C} = \left\{\mathcal{T}^1, \ldots, \mathcal{T}^{|\mathcal{C}|}\right\}$ is simply derived as

$$\ell(\mathcal{C}|\boldsymbol{A}; \{\boldsymbol{\mu}_d\}_{d=1}^D) = \sum_{\mathcal{T}^i \in \mathcal{C}} \ell(\mathcal{T}^i),$$

where $\boldsymbol{A} = \left\{\alpha_{d_j,d_i}\right\}_{d_i,d_j}$ is the D-by-D matrix and D is the total number of diseases. A desirable characteristic of this log-likelihood function is that it is concave in the arguments $\boldsymbol{A}$ and $\{\boldsymbol{\mu}_d\}$, which will allow us to find the global maximum solution efficiently using various convex optimization tools. Moreover, we want to induce a sparse network structure from the diseases and avoid overfitting. If the mutual excitation rates $\alpha_{d_j,d_i} = 0$, then there is no edge (or direct transmission) from the disease d_i to d_j. For this purpose, we impose L_1 type of regularization on the parameters $\left\{\alpha_{d_j,d_i}\right\}$ so that we can obtain a sparse disease network structure. As a consequence, the sparse disease network structure is reflected in the non-zero patterns of the final matrix $\boldsymbol{A}$. Similarly, we impose L_2 regularization on the parameters $\{\boldsymbol{\mu}_d\}$ so that we can obtain robust estimates of the parameters over patient features. Finally, we have the following optimization problem:

$$\min \left\{ -\ell\left(\mathcal{C}|\boldsymbol{A}; \{\boldsymbol{\mu}_d\}_{d=1}^D\right) + \lambda_1 \|\boldsymbol{A}\|_1 + \frac{\lambda_2}{2} \sum_{d=1}^{D} \|\boldsymbol{\mu}_d\|_2^2 \right\}$$
$$\text{subject to } \boldsymbol{A} \geqslant 0,\ \{\boldsymbol{\mu}_d\}_{d=1}^D \geqslant 0 \tag{5}$$

After we learned the network structure with the L_1 regularization, we then refit the nonzero parameters $\left\{\alpha_{d_j,d_i}\right\}$ to achieve better estimations of those parameters without L_1 regularization.

V. Experiments

A. Dataset

Our experiments used the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II) clinical database [20]. MIMIC II is a collection of de-identified clinical visit records of Intensive Care Unit patients between 2001 and 2008 from a single tertiary teaching hospital. At each visit, a patient is diagnosed with the ICD-9 code system. A patient could receive more than one diagnosis at a single visit, one of which is assigned as the primary diagnosis. MIMIC II also includes information regarding the patient such as gender, birth date, weight, medication and various lab test results. After preprocessing the data(extracting patients with at least two hospital visits who had their weights checked at every visit, grouping the 5-digit ICD9 codes to 3-digit ICD9 codes, filtering out non-primary diagnoses), we are left with 593 patients, 186 disease types and total 1,269 visits. The features we used, age and weight, are shown in the axes of Figure 2. We tried to include more features such as blood pressure. However, due to the sparse nature of EHR data, weight and age were the most suitable candidates.

B. Hyper Parameter Setting

cHawkes uses three hyper parameters: Regularization parameters λ_1, λ_2 in Equation 5 and the exponential decay kernel parameter λ in $g(t) = \lambda e^{-\lambda t}$. For λ_1 and λ_2 we tested values of 0, 10, 100 and 1000. For the exponential decay kernel parameter we tested 0.2, 0.4, 0.6, 0.8, 1.0. After iterations of rigorous experiments using a machine equipped with an Intel Xeon E5-2630 (24 cores) and 132GB memory, we chose $\lambda_1 = \lambda_2 = 10$ and 0.2 for the decay kernel parameter. The criteria for choosing the optimal value was the model's predictive performance, which will be discussed in section V.D.

[1] We also tried using real-valued features directly, with interpolation applied between visits. But such an approach exhibited inferior performance.

[2] Rayleigh kernel was also tested, but yielded slightly a slightly inferior performance

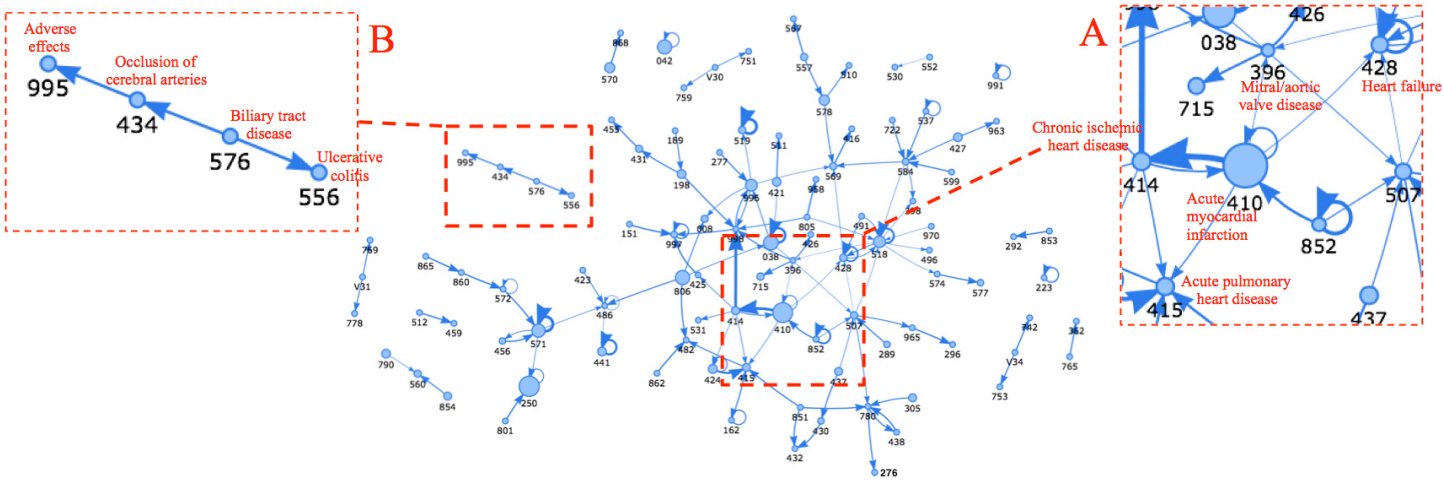

Fig. 1. Disease network of a 65kg, 25-year-old, built with *cHawkes*. Each node represents a type of disease, under which the number is its ICD9 code. The size of each node represents the strength of its spontaneous occurrence(for a person of age 25, weight 65kg). Edges between the nodes represent the direction of influence. Thicker edges mean stronger influence. Diseases that have no connection or very weak connections with other diseases were filtered out in the generation process for succinct representation of information

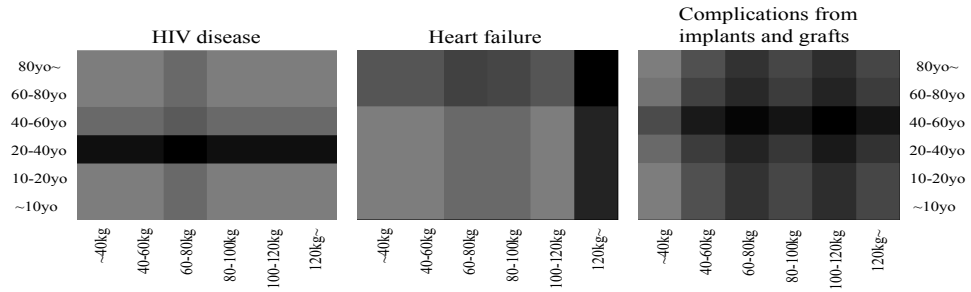

Fig. 2. Heat map of three different diseases: HIV, heart failure, complications from implants and grafts. The Y-axis corresponds to ages: <10, [10, 20), [20, 40), [40, 60), [60, 80), and >80. The X-axis consists of weights: $<$40kg, [40kg, 60kg), [60kg, 80kg), [80kg, 100kg), [100kg,120kg), and $>$120kg. The darker regions represent stronger activity of the disease.

C. Disease Relation, Context Sensitivity and Temporal Dynamics

In this section, we present the disease network to show the relations between diseases, explain how change of context(patient features) affects disease occurrence, and also show how occurrence intensities of diseases change over time.

1) Disease Network: Figure 1 is the disease network constructed using *cHawkes*, specifically for a patient of age 25, weight 65kg. The qualitative interpretation of the networks was provided by a current medical student who also has broad experience in medical data mining. The network was confirmed to be clinically meaningful with connections that represent possible real-world scenarios where certain diseases may precede others. In Figure 1, we provided two specific examples where relations between nodes are explained.

In Figure 1, subgraph A, the largest node corresponds to acute myocardial infarction(AMI). The most influential edge connects AMI to other forms of *chronic ischemic heart disease*, which is a relationship that is commonly seen in real life. Furthermore, AMI has an edge that points to itself. This is also clinically significant because some patients may experience successive episodes of re-infarction after the first MI event [21]. The *heart failure* node connected to AMI represents the fact that heart failure may follow MI in about 29% of patients [22]. The *acute pulmonary heart disease* node is connected to AMI, which is clinically meaningful, as cardiogenic pulmonary edema can occur as a complication of systolic heart failure [23, Chapter 11]. The *diseases of the mitral and aortic values* node may represent complications of myocardial infarction such as mitral valve prolapse [23, Chapter 11].

Another set of disease relations worth mentioning is shown in Figure 1, subgraph B. The root of this subgraph, *biliary tract disease*(BTD) is connected to *ulcerative colitis*. This relationship is fundamentally important because there is a known medical association between primary sclerosing cholangitis(PCS), a type of biliary disease, and ulcerative colitis(UC) [24]. BTD also points to *occlusion of cerebral arteries*. One explanation for this is that both biliary disease and occlusion of cerebral arteries (a possible result of thromboembolism) may result due to side effects of oral contraceptives such as drospirenone/ethinylestradiol [25] [23, Chapter 19]. *Occlusion of cerebral arteries* points to *certain adverse effects not classified elsewhere*. This particular node makes sense to be linked to cerebral artery occlusion, because ischemic blood loss that results from the occlusion may result in a number of musculoskeletal, spatiovisual, or cognitive deficits.

2) Context-Sensitivity of Diseases: As we have optimized $\boldsymbol{\mu}$ vectors of different disease, we can now plug in specific age and weight to analyze how diseases behave under different contexts. Figure 2 is the heat map of three different

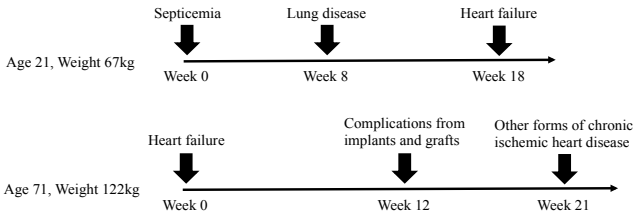

Fig. 3. Disease history of two people with different physical features

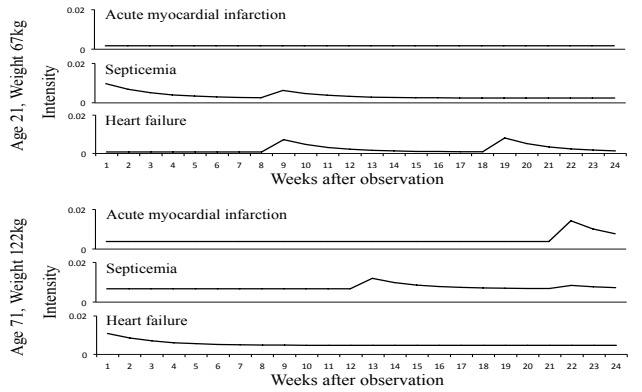

Fig. 4. Disease intensities of two people with different physical features. We can see that intensity trajectories differ for different patients even though the diseases being plotted are the same.

diseases, *Human immunodeficiency virus disease*(042), *Heart failure*(428) and *complications from implants and grafts*(996) calculated using a range of values for ages and weights. The heat maps are based on $\boldsymbol{\mu}$ variables of Equation (4).

It can be seen that diseases behave quite differently in different contexts. For example, HIV disease acts especially strongly for people between the age 20 and 40, which correctly reflects reality[3].This characteristic could be due to the fact that younger people are more sexually active, rendering them more susceptible to sexually transmitted diseases than older people. *Heart failure*, on the other hand, is heavily affected by a person's weight. It is also correlated with higher age, but the correlation is stronger with obesity. This is consistent with the common knowledge that obese people are more susceptible to heart diseases[4]. *Complications from implants and grafts* seem to be ubiquitous to various groups of people except for patients with a very light weight. This is equivalent to saying that this disease can generally occur to anyone except for very young children who weigh under 40kg. This makes sense as young children are less likely to receive implants or grafts such as cardiac devices or prosthetic joints.

3) Temporal Dynamics: Here we present two groups of disease intensities, one for a young, average weight person(age 21, weight 67kg), and another for an old, overweight person(age 71, 122kg). Each person suffers different sets of diseases throughout 4 months as shown in Figure 3. The former patient has suffered *septicemia*, *other diseases of lung* and *heart failure*. The latter patient has suffered *heart failure*, *complications from implants and grafts* and *other forms of chronic ischemic heart disease*. Based on theses records, we plotted the intensities of three well-known diseases for each patient using Equation 4.

It can be seen from Figure 4 that the intensity of a disease not only depends on the types of diseases the patient had and how much time has passed after each disease, but also on the physical features of the patient. For example, the intensity of acute myocardial infarction(AMI) for the younger patient maintains its default strength since the diseases he suffered have little influence on AMI. The intensity of septicemia, on the other hand, spikes after he suffers septicemia and lung disease, as they both influence the occurrence of septicemia. You can see, however, that after a period of time, the intensity of septicemia drops down to its default strength. The intensity of heart failure maintains very weak default strength except for the two times the patient suffers lung disease and heart failure. For the older patient, after being influenced by the initial instance of heart disease, the heart failure intensity maintains relatively high throughout the whole observation compared to the younger patient. This is due to the difference of physical features of the two patients. A higher heart failure intensity for the older, overweight patient is consistent with what we have presented in section V.C.2, where we have shown that heart failure acts more strongly when combined with obesity.

D. Disease Prediction

Next we provide quantitative evaluation of *cHawkes* by performing disease risk prediction. Given a disease history of a patient, we try to predict the most likely disease he/she will have in a certain future time window by calculating the conditional cumulative distribution of each disease.

Figure 5 is the result of 10-fold cross validation of risk prediction by various methods including Hawkes Process and other well-known methods. We tried predicting diseases that occurred in four different three months windows, while varying the number of predictions p. If one of p predictions is correct, we consider it an accurate prediction. For *cHawkes* and Hawkes Process, we choose p diseases with the highest conditional cumulative probability. We also tested homogeneous Poisson Process which is equivalent to removing the α variables from the intensity function (2) of Hawkes Process. For Poisson Process, we also used conditional cumulative probability for prediction. Linear regression models were trained to predict the time of the next occurrence of the disease, given all past diseases and the most recent age, weight information. We picked all diseases that occurred in the target time window. The average number of the selected diseases for the different time windows are 48, 14, 11 and 11. We put a single dot for linear regression, as a type of baseline. For multinomial logistic regression, we used the same features as linear regression. Again, p diseases with the highest probability were chosen. Although logistic regression cannot be directly compared with other temporal models, we plot its performance in the first figure so that it can serve as a reference.

It is readily visible from Figure 5, that *cHawkes* outperforms other well-known methods except for the first three month window. In that particular window, *cHawkes* was having trouble correctly predicting diseases concerned with newborn infants, especially *Other and ill-defined conditions originating in the perinatal period*(779). *cHawkes*, however, was predicting *Disorders relating to short gestation and low birthweight*(765) instead, which is very similar to 779. In all other situations, *cHawkes* exhibits superior performance. Also it is worth noting that *cHawkes* is particularly robust in making

[3]http://www.cdc.gov/hiv/risk/age/olderamericans/

[4]http://www.cdc.gov/healthyweight/effects/

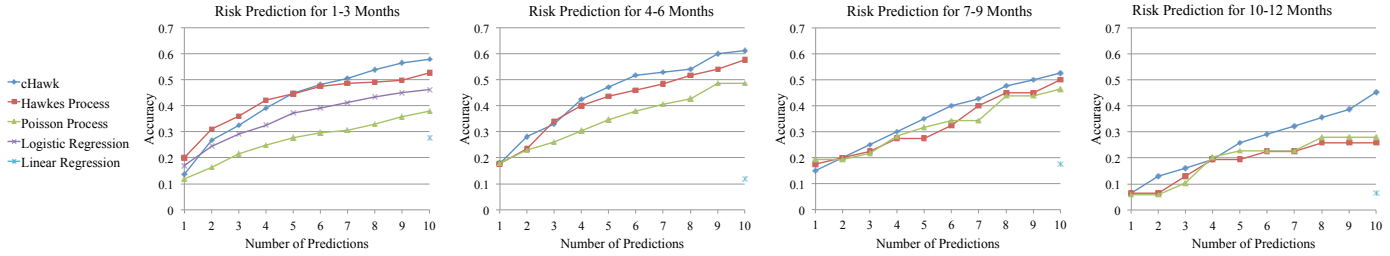

Fig. 5. Prediction performance comparison for different time windows. We used $\lambda_1 = 10, \lambda_2 = 10$ for *cHawkes* and Hawkes Process. Regularization was ineffective for Poisson Process. For the exponential decay parameter, we used 0.2 for *cHawkes*, Hawkes Process and Poisson Process. We used Python Scikit-Learn for Logistic Regression and Linear Regression.

a longer prediction, which can be attributed to its taking into consideration patient feature.

VI. Conclusions

In this paper we proposed *cHawkes* to capture the three aspects of EHR, namely disease relations, context-sensitivity of diseases and temporal dynamics of diseases. We showed a detailed derivation of the model. In the experiments, we applied *cHawkes* to MIMIC II, a real-world EHR comprised of ICU patients, to build disease networks, context-sensitive heat maps and show temporal dynamics of disease intensities. Risk predictions were performed for quantitative evaluation, which showed that *cHawkes* is able to predict future diseases more accurately than the original Hawkes Process and other traditional methods.

In the future, we plan to apply *cHawkes* to several EHR datasets of varying patient demographics and diverse sets of diagnoses. Such application will uncover different sets of meaningful disease relations. We also plan to use a larger dataset with numerous patient features, so that we can identify useful relations between various patient features and the risk of disease occurrence.

Acknowledgements

This work was supported by the National Science Foundation, award IIS- #1418511 and CCF-#1533768, Children's Healthcare of Atlanta, CDC I-SMILE project, Google Faculty Award, AWS Research Award, Microsoft Azure Research Award, UCB, NSF/NIH BIGDATA 1R01GM108341, NSF IIS-1116886, NSF CAREER IIS-1350983, ONR N000141512340 and Samsung Scholarship

References

[1] L. Ohno-Machado, "Nih's big data to knowledge initiative and the advancement of biomedical informatics," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 193–193, 2014.

[2] F. Wang, P. Zhang, B. Qian, X. Wang, and I. Davidson, "Clinical risk prediction with multilinear sparse logistic regression," in *KDD*, 2014, pp. 145–154.

[3] J. Zhou, F. Wang, J. Hu, and J. Ye, "From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records," in *KDD*, 2014, pp. 135–144.

[4] Y. Park and J. Ghosh, "Ludia: An aggregate-constrained low-rank reconstruction algorithm to leverage publicly released health data," in *KDD*, 2014, pp. 55–64.

[5] J. C. Ho, J. Ghosh, and J. Sun, "Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization," in *KDD*, 2014, pp. 115–124.

[6] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding physiological state: Mortality modelling in intensive care units," in *KDD*, 2014, pp. 75–84.

[7] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *KDD*, 2014, pp. 85–94.

[8] A. G. Hawkes and D. Oakes, "A cluster process representation of a self-exciting process," *Journal of Applied Probability*, pp. 493–503, 1974.

[9] J. Leiva-Murillo, A. Artés-Rodríguez, and E. Baca-García, "Visualization and prediction of disease interactions with continuous-time hidden markov models," in *NIPS 2011 Workshop on Personalized Medicine*, 2011.

[10] G. Savova, S. Bethard, W. Styler, J. Martin, M. Palmer, J. Masanz, and W. Ward, "Towards temporal relation discovery from the clinical narrative," in *AMIA*, 2009, pp. 568–572.

[11] Y. Zhao, X. Qi, Z. Liu, Y. Zhang, and T. Zheng, "Mining medical records with a klipi multi-dimensional hawkes model," in *KDD 2014 Workshop on Health Informatics*, 2014.

[12] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *KDD*, 2012, pp. 1095–1103.

[13] Y.-Y. Liu, H. Ishikawa, M. Chen, G. Wollstein, J. S. Schuman, and J. M. Rehg, "Longitudinal modeling of glaucoma progression using 2-dimensional continuous-time hidden markov model," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, 2013, pp. 444–451.

[14] S. Myers and J. Leskovec, "On the convexity of latent social network inference," in *NIPS*, 2010, pp. 1741–1749.

[15] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *KDD*. ACM, 2010, pp. 1019–1028.

[16] M. Gomez Rodriguez, D. Balduzzi, B. Schölkopf, G. T. Scheffer *et al.*, "Uncovering the temporal dynamics of diffusion networks," in *ICML*, 2011, pp. 561–568.

[17] N. Du, L. Song, M. Yuan, and A. J. Smola, "Learning networks of heterogeneous influence," in *NIPS*, 2012, pp. 2789–2797.

[18] L. Wang, S. Ermon, and J. E. Hopcroft, "Feature-enhanced probabilistic models for diffusion network inference," in *ECML PKDD*, 2012, pp. 499–514.

[19] N. Du, L. Song, H. Woo, and H. Zha, "Uncover topic-sensitive information diffusion networks," in *AISTATS*, 2013, pp. 229–237.

[20] M. Saeed *et al.*, "Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database," *Critical care medicine*, vol. 39, no. 5, p. 952, 2011.

[21] S. J. Kernis, K. J. Harjai, G. W. Stone, L. L. Grines, J. A. Boura, M. W. Yerkey, W. O'Neill, and C. L. Grines, "The incidence, predictors, and outcomes of early reinfarction after primary angioplasty for acute myocardial infarction," *Journal of the American College of Cardiology*, vol. 42, no. 7, pp. 1173–1177, 2003.

[22] P. S. Jhund and J. J. McMurray, "Heart failure after acute myocardial infarction a lost battle in the war on heart failure?" *Circulation*, vol. 118, no. 20, pp. 2019–2021, 2008.

[23] E. F. Goljan, *Rapid Review Pathology*, 12nd ed. Elsevier Saunders, 2014.

[24] R. Chapman, Z. Varghese, R. Gaul, G. Patel, N. Kokinon, and S. Sherlock, "Association of primary sclerosing cholangitis with hla-b8." *Gut*, vol. 24, no. 1, pp. 38–41, 1983.

[25] M. A. Goldstein, *The MassGeneral Hospital for Children Adolescent Medicine Handbook*. Springer Science & Business Media, 2010.