Interpreting Comorbidity Groups via Risk Trajectories in the Health Record

Bharat Srikishan
Columbia University
bs2982@columbia.edu

Rajesh Ranganath
Princeton University
rajeshr@cs.princeton.edu

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

Abstract

In this paper, we aim to explore patient trajectories in time that evolve according to their risk of developing comorbidities. For our analysis, we use a probabilistic latent variable model, which aligns patients through time according to their risk of developing conditions within discovered comorbidity groups. We report our findings on a large population of patients and their in-hospital admissions through time (300,000 patients, over a span of 24 years overall).

1 Introduction

We are interested in exploring the trajectories of patients in a hospital according to their risk of developing conditions. There is a wide mixture of problems that causes each admission to a hospital, ranging from acute problems to chronic ones drawn from a large range of conditions [1]. Probabilistic latent variable models enable the automated discovery of groups of conditions that are coherent with each other [2, 4]. Temporal probabilistic latent variable models can align patients in time, thus allowing for the exploration of how risk evolves for patients in different comorbidity groups.

In our work, we consider the survival filter [5]. Given a patient population and their longitudinal series of healthcare visits and corresponding diagnosis codes, the survival filter learns groups of comorbidities. Further for each patient, it infers a latent temporal representation that captures the risk of an event occurring from each group of comorbidities across time.

We apply the survival filter to a large cohort of patient records (300,000) which consists of in-hospital admissions only and varies over a large space of diagnosis codes (over 9,000). We then explore the trajectory of risk across time for high-risk patients within each co-morbidity group.

This paper makes the following contributions: (1) we describe a set of visualizations to explore the way the burden of disease evolves though time for patients; (2) we explore the burden of disease for patients at high risk of developing new complications within a given comorbidity group (within-group analysis); and (3) we explore the burden of disease for patients at high risk of developing complications across comorbidity groups according to dimensions of acute vs. chronic, healthy vs. sick, and comorbid vs. single-issue. For instance, some comorbidity groups like the discovered pregnancy group have generally low-risk on every other comorbidity group (i.e., pregnant patients are typically healthy and their interaction with the healthcare institution is limited to their pregnancy). In contrast, the infection comorbidity group shows trajectories for patients that are acute and well-delimited in time according to risk, the patients are also sicker than general with associated high risk for a mixture of comorbidity groups including heart disease, renal disease, and metabolic disorders.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Table 1: Sample Learned Groups

Group 2	Group 3	Group 7	Group 18	Group 19
Gynecological examination	Congestive heart failure	Depressive disorder	Hypertension	Mother with single liveborn
Other screening mammogram	Essential hypertension	Anxiety state	Pure hypercholesterolemia	HIV counseling
Lump or mass in breast	Atrial fibrillation	Alcohol abuse	Type II diabetes mellitus	Supervised normal pregnancy
Vaginitis	Pneumonia	Tobacco use	Hyperlipidemia	Pregnant state
Leiomyoma of uterus	Coronary atherosclerosis	Major depressive disorder	Osteoarthrosis	Counseling on contraceptives

The paper is organized as follows. We first give an overview of the survival filter and detail scalable inference for it. We provide a description of our dataset, and describe our findings.

Model Survival analysis studies the time to an event. Traditionally, models for survival analysis focus on a single event. Ranganath et al. [5] developed a model for simultaneous survival problems called the survival filter. The survival filter models each patient with a latent trajectory that interacts with factors shared across patients to produce the risk a patient has for an event (in our case diagnostic codes) at a particular time.

Our model has a few important variables, the data: $x_{p,t,c}$; the latent risk associated with a patient acquiring a new code at time t: $z_{p,t}$; and finally the weights for each factor/grouping of ICD9 codes: W. The observation $x_{p,t,c}$ is one if code c occurs at visit t for patient p and zero otherwise.

Let σ and μ denote hyperparameters and D be a distribution over the positive reals. The generative process of the model is

$$
\begin{aligned}
W &\sim D \\
z_{p,1} &\sim \text{Normal}(\mu, \sigma^2_{z_0}) \\
z_{p,t} &\sim \text{Normal}(z_{p,t-1}, \sigma^2_z) \\
x_{p,t,c} &\sim \text{Bernoulli}(1 - \exp(-W_c^\top \exp(z_{p,t}))).
\end{aligned}
$$

The weights W shared across data group the diagnostic codes into groups of comorbidities. Given these groups of comorbidities, the z correspond to per-patient risks for the groups of comorbidities. Both W and z are inferred via variational inference [3]. (For more on variational inference for the survival filter, see Ranganath et al. [5].) Because the survival filter models both risk over time and comorbidities, we are able to directly examine patients over time in comorbidity groups.

To accelerate inference, we parallelized inference across multiple machines using a parent-child node structure. This was made possible by the conditional independence structure in the model. That is, given W each patient is independent. Thus, inference can proceed in parallel across multiple patients, while the parent maintains the current approximation on W.

Data Our dataset contains ICD9 codes per visit for each patient. We have a total of 304,941 patients with 8,562 ICD9 codes. This data was collected at a large metropolitan hospital.

2 Results

For our experiments, we had the model learn $K = 25$ code groups with our positive real distribution D being the Log Normal distribution.

Learned Groups. We list a sample of some of the learned groups in Table 1. As is shown, many of the groups have codes that occur within the same category of conditions. Group 2 contain OB/GYN conditions, Group 3 encompasses heart conditions, and Group 7 includes depression, anxiety, and other psychological conditions, Group 18 has chronic lifestyle conditions like hypertension, high cholesterol, diabetes, Group 19 is the pregnancy group containing the normal birth codes.

Patient Risk within Groups. $z_{p,t}$ can be interpreted as a measure of the risk of developing new conditions within a comorbidity group for a given patient. This can be seen in the model generative process as $x_{p,t,c} \sim \text{Bernoulli}(1 - \exp(-W_c^\top \exp(z_{p,t})))$. As $z_{p,t}$ increases, the code c is more likely to occur for patient p.

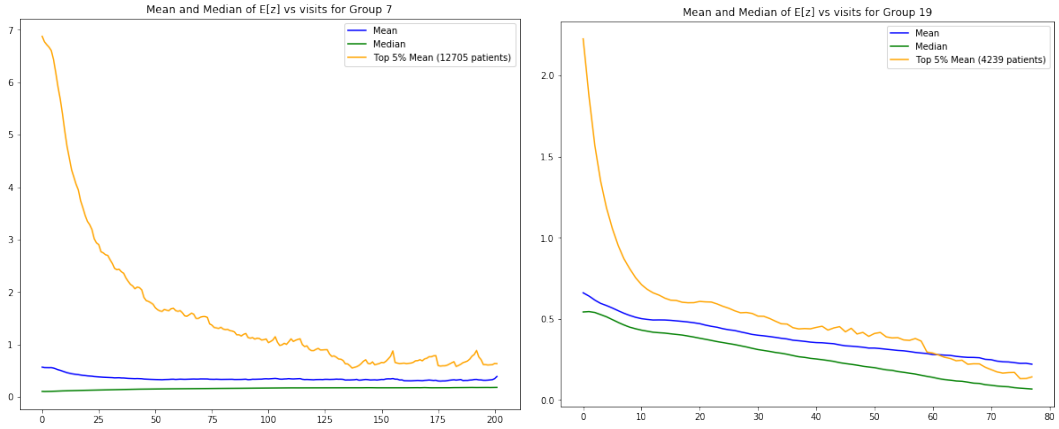
High risk patients begin with a high disease burden represented in the model by the z latent variable. To determine how high risk patients evolve over time, we plot the mean, median, and top 5% of

patient's expected z values in each group over time. We limit the time to the domain where there are at least 50 patients remaining in the top 5% so as to reduce noise.

In general, the same graph profile presents across the groups. High risk patients (top 5%) begin with a large expected z value and fall over time. In some groups, this risk regresses to the mean risk of the population. However, in the majority of groups, the risk falls but remains above the population mean risk for the high risk patients. This suggests that high risk patients accumulate conditions early in the course of their hospital visits, and as they accumulate conditions, their survival risk falls within their comorbidity group.

For example, a patient with high expected z in the mental health group is likely to develop depressive disorder, acquires it early in their timeline while still being likely to aquire other conditions in the same group such as alcohol abuse or anxiety. As this patient progresses through future visits, the model implies that they are likely to acquire these other conditions (alcohol abuse, anxiety) and their survival risk as determined by expected z will fall.

Figure 1: Risk Within Groups



(a) Mental Health Group

(b) Pregnancy Group

Figure 1a shows the typical risk profile in most comorbidity groups. This group includes mental health conditions (depression, anxiety, and drug use) and implies that high risk patients are more likely than the general population to acquire other codes from this comorbidity group.

Figure 1b displays a unique group risk profile as the high risk patients fall below the mean population risk. This is the pregnancy comorbidity group which contains codes such as normal pregnancy, pregnant state, normal delivery. The risk profile displayed is expected with a normal birth, as it is an acute condition which resolves in a usually predictable way with low risk of reentering the same comorbidity group for some time.

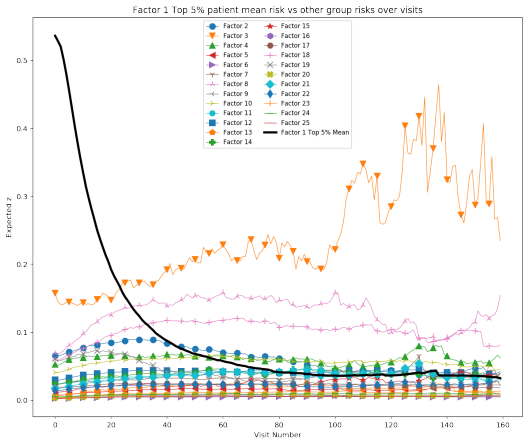
Many groups have similar behavior to the mental health comorbidity group. We hypothesize that this behavior occurs because as high risk patients progress through the group in time, they accumulate conditions within the group, and the risk/hazard as defined by the model falls.

Patient Risk Across Groups. We now examine patient risk across comorbidity groups. We plot the top 5% of high risk patients within a group along with those patient's risk profiles in all other groups.

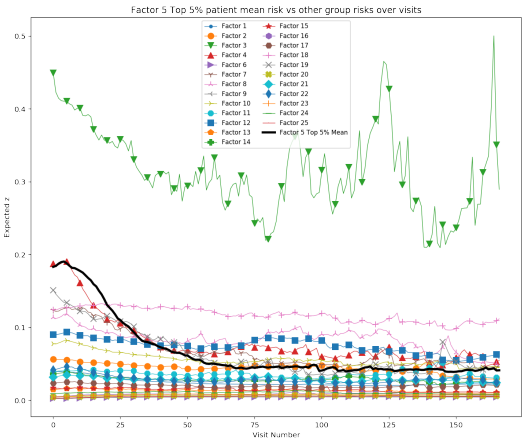
In Figure 2c we show the normal pregnancy group of patients. The model shows that this comorbidity group does not correlate with risk in other groups. As expected, normal pregnancy and birth would not result in higher risk in other comorbidity groups.

Figure 2a represents general medical care such as laboratory examination, counseling, and general medical examination. The high risk patients in this group exhibit a high risk in many other comorbidity groups as well, however this is an example of how the measurement process is predictive of what will happen. Medical examination often leads to diagnosis of a new condition and as a result, the model estimates high risk in other comorbidity groups.

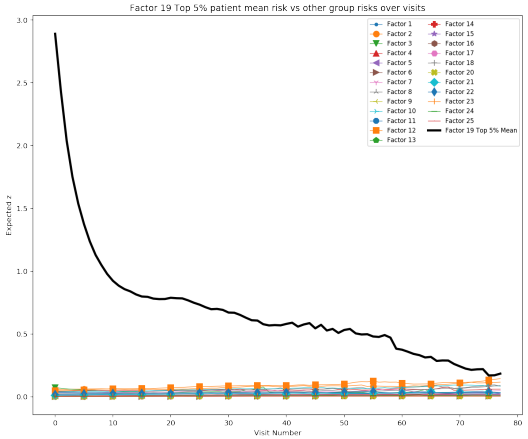
Figure 2: Cross Group Risk



(a) General Medical Care



(b) Hospital Acquired Conditions Group



(c) Pregnancy Group

Table 2: Model Performance vs Baselines

	Log Likelihood	Rank	Recall
Mean Disease Risk Baseline	-4854920	616.881	0.0891799
Person Disease Risk Baseline	-5954510	4246.29	0.00225894
Survival Filter	**-4579830**	**450.946**	**0.137399**

Figure 2b visualizes the group of conditions acquired during long term hospital stays such as cellulitis, abscess, edema, and local infection of skin. Here we see that group 3, the heart conditions comorbidity, has varying risk over time for this same cohort of patients. Additionally, other groups in the figure also seem to have independent trends from our base group. This suggests that our patients risk in other comorbidity groups is independent of them having high risk in the hospital acquired conditions group.

Baselines Similar to Ranganath et al. [5], the survival filter model performed better than baseline measures in terms of log likelihood, rank, and recall. See Table 2.

3 Conclusion

We used a probabilistic latent variable model to learn groups of comorbidities and how patients express those groups over time. We used the inferences from the model to explore how the burden

of disease in comorbidity groups evolves and study how groups of comorbidities interact. We find groups that vary along various dimensions such as acute vs. chronic, and that while most comorbidities do not interact, some such as pregnancy dampen the likelihood of other groups of conditions. The use of a model to align and find groups of comorbidites made it possible to develop human evaluations.

References

[1] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2012.

[2] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[3] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[4] Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous ehr data. *Journal of biomedical informatics*, 58:156–165, 2015.

[5] Rajesh Ranganath, Adler J Perotte, Noémie Elhadad, and David M Blei. The survival filter: Joint survival analysis with a latent time series. In *UAI*, pages 742–751, 2015.