國立臺灣大學管理學院資訊管理研究所

碩士論文

Department of Information Management
College of Management
National Taiwan University

Master Thesis

用非監督學習建立疾病進展模型

Models of Disease Progression: An Unsupervised Learning Approach

林鈺嫻

Yu-Hsien Lin

指導教授：盧信銘 博士

Advisor: Shin-Min Lu, Ph.D.

中華民國 104 年 9 月

September, 2015

誌謝

在寫論文的路上,對於所有曾經交流過的人們我都非常感謝,但無法一一感謝。以下對於那些貴人書寫些文字表達感謝。沒有這些貴人，就沒有今天的我。

能完成這篇論文，首先要謝謝無條件支持我的父母與家人。沒有他們的各種無論是心靈上還是物質上的強力支援，我不可能會進入台大資管所，然後還能夠從台大資管所畢業。只能督促自己未來要更加努力，才能對得起父母的栽培養育之恩。也要謝謝姐姐的各種協助，姐姐的各種威能與榜樣也總是讓我景仰的。

這篇論文能夠誕生且孵化成碩士論文,都必須感謝指導教授盧信銘老師的教誨。如果沒有老師的指導，我自己是不可能完成這篇論文的。盧老師除了對我在學業上的指導，對於其他方面像是未來職涯等也都很關心，這些學生都謹記在心。

假如缺少好友的支持，我寫起論文一定無法得心應手。雖然好友無法幫我寫論文，但如果沒有常常跟好友一起講些五四三的話題讓我心情愉悅的話，我一定無法寫得如此順利。所以在此鄭重感謝大學好友老馮的陪伴，讓我得到能量，並且能夠支持我寫完論文。在此也祝福他外國留學一切順利。

感謝兩位研究室的夥伴,首先要感謝博士班的大學長提供了我許多寶貴的意見,總是讓我超級敬佩大學長的學識淵博與歷練。以及超有氣質的小花同學也常常告訴孤陋寡聞如我許多重要的資訊，讓我可以節省很多自己去查的精力。這兩位夥伴的幫助常常都是及時雨，點滴在我的心頭，讓我沒齒難忘。

寫論文的時候如果沒有適當的放鬆，很容易就把自己耗盡能源、精疲力盡。因此我想在這邊感謝那些陪伴我一起寫論文的娛樂們，雖然他們並不會知道。如果我沒有喜歡上寶塚，我不會知道那些每天都努力試著實現自己夢想的人們（就算後來夢想沒有成真），也不會因而鞭策自己更努力。非常感謝有機會認識寶塚。

期許未來的自己能夠有面對挑戰努力堅持的勇氣！

摘要

由許多事件組合而成的連續序列資料中，可能隱含一些像是病況的嚴重程度，或是病情加重的速度等資訊。舉例來說，像是在醫院中的病人的看診資料，如果把每次看診的疾病代碼當作是一個事件，每位病人的病例就能夠看成是一序列資料。而隨時間累積而成的序列有兩個獨特的特性。第一個特性是序列能夠根據不同的演化速度或事件組合而被區分成不同的類別（class），另一個特性是可以根據序列中發生的事件順序來推算目前此序列所在的階段（stage）。而因為序列之間會有不同的序列長度和演化速度，因此要來預測序列的類別與階段是有難度的。以慢性病病人來說，他們的病歷資料就會有上述的那些特性。因此我們能利用建立模型來分析資料，藉以從模型的產出找到有用的資訊來預測或是預防疾病的發展。從前人的研究得到的結果可以知道，利用疾病發展模型不只可以預測疾病的發展，連疾病的共病和藥物作用都可以被預測，也因此這些研究的結果能夠廣泛的運用在醫療之中。

在本研究中，我們使用資料導向（data-driven）的方法來分析健保資料中的糖尿病患者病歷，並且從資料中擷取出疾病發展的不同階段。本研究的結果顯示，用於研究中的模型預測能力跟前人的研究結果類似，而從分析模型產出的類別也發現，從這些類別中可以推測出兩種不同類別的糖尿病患。這些從本研究中得到的資訊可以讓我們更加了解糖尿病患的發展模式與其中的差異，並且可以提供給後人參考。

關鍵字：事件序列、疾病發展模型、時間序列、非監督式學習、資料導向

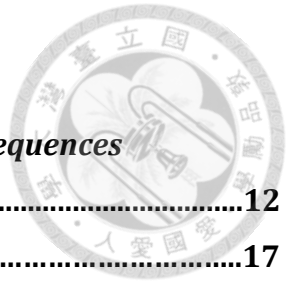
Abstract

A series of events, such as a patient's medical records, have two natural features, class and stage that are not easy to find. Since each event sequence may have different length, and different progression speed. Especially for chronic diseases patients, they may suffer with these diseases for a long time. The development of model to estimate the disease progression can help to provide information for them. From previous findings, their results suggested that by modeling disease progression, not only disease progression rate can be predicted but disease's comorbidities and drug effect.

In this paper, we present a data-driven approach to analyze the health insurance claim records to extract the disease progression stages of diabetic patients. Our experiments suggested that our model's performance is consistent with previous finding. And the progression classes learned from our model have revealed different types of diabetic patients.

Keywords: event sequence, disease progression model, time series, unsupervised learning, data-driven

Contents



List of Figures

List of Tables

1 Introduction

A series of events, such as a patient's medical records, may contain implicit information, for instance, disease severity, progressing rate and progressing pattern. Theses information can be detected through two distinct features in event sequence: stage and class. Progression stages are the phases that may change overtime. For example from McAuley et al. (2013), it's known that as users acquire more products, their tastes will change. As for classes, different patients with different genders or ages may progress through stages at different speed. (Kramer et al. (2003)) The goal of our study is to model these progression patterns.

Chronic diseases are one of the most influential disease types worldwide. Thus the treatments and interventions for chronic diseases are crucial for doctors to find. According to World Health Organization's (WHO) statistics, 4 of top 10 causes to death around the world are noncommunicable diseases, including cardiovascular diseases, cancers, diabetes and chronic lung diseases. Moreover, in Taiwan, diabetes is one of the top 10 causes to death in 2012. Chronic diseases like diabetes may cause patients to suffer for a long period of lifetime. Thus, it is needed to attend more useful information to intervene or remedy chronic diseases. The development of a model to estimate the disease progression can help provide new insights.

Disease progression modeling (DPM) is an important methodology to model disease process over time through mathematical and systematic approach. (Mould et al. 2012) DPM can get helpful information for doctors to adjust or identify the development of a

disease to intervene the symptoms for chronic diseases. By using DPM, analyzing disease process can improve the prediction of disease states, drug development, and treatment design.

In this paper, we present a data-driven approach to analyze the health insurance claim records to extract the disease progression stages of diabetic patients.

2 Literature Review

Existing works on disease progression modeling have revealed useful information in the medical field, such as the target disease's treatment redesign, disease progression rate of change, drug development and early intervention. In the following, we will discuss the related works in two parts according to either having domain knowledge before modeling or not. After discussing previous findings, we have listed different models with their features and other information in table 1 for comparison.

2.1 Progression models with domain knowledge

Previous studies of modeling with domain knowledge have found relationship between diseases and diseases' indicators. For example, Doody et al. (2010) have proposed that there are correlations between a pre-progression rate and disease severity. Pre-progression rate is calculated through combining clinician's standardized assessment of symptom duration and the baseline Mini Mental Status Exam (MMSE) score which is a commonly used cognitive score to measure dementia severity. After pre-progression rate has been calculated, a regression model is made to find the relationship between the pre-progression rate and subsequent rate of decline on cognitive and functional measures of Alzheimer's disease. In other word, the pre-progression rate can estimate the cognitive value of Alzheimer's disease in the following years. In addition, the average survival time for each progression group was estimated through Cox survival analysis and the analysis showed that the slow progress group has longer survival years. The results suggested that the pre-progression rate is better at predicting slow and fast progress

patient, and the slower the progress is the longer the patient will live. On the other hand, Raj et al. (2012) modeled disease progression using longitudinal magnetic resonance imaging (MRI) images which are graphs scanned from human bodies such as brain. After preprocessing the images, they presented a model to predict dementia diseases' evolution. By using correlation analysis, they found convincible evidence that the degeneration processes of dementia disease matches latest reports of dissociated brain networks.

However, with only one progression model could gain little knowledge. To have a more comprehensive perspective of the disease progression, Mould (2012) have used meta-analysis-based modeling to combine multiple models' results for treatment adjustment and testing trial revision. Meta-analysis-based modeling has gathered multiple findings' results to increase the individual model's capability and robustness. This study has presented that through combining multi models' information from progression modeling can help not only redesign the clinical trials but also gain evaluations of new treatment combinations.

To gain useful knowledge more than only on disease progression, Postet et al. (2005) have proposed a progression model, which can estimate both disease progression and drug effect. According to their findings, not only the disease's process can be monitored through progression modeling, but also the treatment's efficacy can be recognized.

2.2 Progression models without domain knowledge

Besides modeling with domain information, many studies have been made through data mining techniques. For instance, Yang et al, (2014) proposed a generative model in which they gather event sequences, including patients' records, web pages view history, product reviews and textual sentences together to identify event sequence's class for different evolving pattern, and identify each event into different stage at the same time. With the results they have for classes and stages, they have found some interesting pattern along with different dataset. This study has revealed that stages and classes can be detected through combining all time points together to be viewed as a sequence. Furthermore, Fonteijn et al, (2012) proposed a Markov Chain Monte Carlo algorithm to model disease progression by treating each patient's state change as an event. In this study, the disease progress is treated as time-varying events sequence. The model has revealed that through event-based progression modeling, the disease process can be mined for Alzheimer and Huntington's diseases. In addition, Zhou et al. (2011) have presented a model by using Multi-Task regression for predicting disease progression. They treated a prediction at a time point as a task, there can be more than one predictions at a time point. Their model, comparing to previous related data, can predict the progression process of Alzheimer disease. Likewise, to use an adjusted modeling method on the same target disease, Zhou et al. (2012) has proposed a model by using Fused Sparse Group Lasso to cut down features and to predict disease progression of Alzheimer disease. The presented model, which included known biomarkers, not only can

predict disease progression, but also can reveal pattern of biomarkers of Alzheimer disease.

Nonetheless, progression modeling can provide more information than predicting disease states. Some studies have indicated that through disease progress modeling can help detecting different stages of disease and its comorbidities. For example, Wang et al, (2014) have presented a probabilistic model through Markov Jump process. They used clinical records to discover Chronic Obstructive Pulmonary Disease (COPD)'s progression stage as well as COPD's comorbidities. Their results suggested that even without professional knowledge of medical background, disease stage and comorbidities can be discovered by their model.

Besides modeling for disease, the patient's state can also be captured through modeling. Cohen et al. (2010) have proposed a clustering approach by using hierarchical clustering to model patient's stages in intensive care units (ICUs). The result have showed that there are 10 clusters for all patients. Each cluster represents a patient stage, such as infection, multiple organ failure or mortality. Patient's state in ICU may jump from one state to another. From this study, we can obtain that by using modeling techniques, we can gain new insights of progression process about patient's states to help hospitals to manage the usage of ICUs.

Table 1. Comparison between different models from previous findings.

Author	Proposed model	Model features	Target disease
Doody et al. (2010)	Regression analysis	Cox survival analysis	Alzheimer's
Raj et al. (2012)	Correlation analysis	MRI graphs	Dementia diseases
Mould et al. (2012)	Meta-analysis based	Combined several studies' results	Alzheimer's & Crohn's disease
Yang et al. (2014)	Generative model	Stage and class.	Chronic kidney disease
Fonteijin et al. (2012)	Event-based model	Stage change as event	Alzheimer's & Huntington's disease
Zhou et al. (2011)	Multi-task regression	Predict multiple scores at one time point	Alzheimer's
Zhou et al. (2012)	Multi-task regression	Fused sparse group Lasso	Alzheimer's
Wang et al. (2014)	Continuous-time Markov model	Onset comorbidities	COPD

3 Dataset and Model

In this chapter, the dataset and the model will be presented and explained. We will first describe our dataset and the filtering rules for diabetic patients from National Health Insurance Research Database (NHIRD). In the dataset, diagnoses were coded according to the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). As for drug prescriptions, we use Anatomical Therapeutic Chemical code (ATC), which is a pharmaceutical coding system controlled by WHO, to apply to our filtering rules. After data description has been made, we will propose our model.

3.1 Dataset

The data used in this study are extracted from NHIRD, which is an administrative medical database. The original data records comprise of patients' records on medical treatments and diagnosis for inpatients and outpatients. To make our filtering rules to be more robust, we compare the rules previous studies have used in their studies for Diabetes type 2 disease. We have gathered three studies, and the rules they proposed are arranged in table 2. We explain each rules in the following paragraph.

Table 2. Selecting Diabetic Patients Rules from Previous Studies

Disease	Filtering rules	Source
Diabetes type 2 disease	Type 2 diabetes diagnosis.	Yah et al, (2012)
	Diabetes and one or more oral hypoglycaemic agent or insulin prescription.	Lin et al, (2013)
	1. Diabetes and one antidiabetic drugs prescription. 2. One or more oral antidiabetic agents' prescription and diabetes-related illness ambulatory visit. 3. Four or more ambulatory visits for diabetes-related illness within 1 year. Exclusions: 1. Diagnosis of type 1 diabetes. 2. Younger than 30 years of age and doesn't have any oral anridiabetes drugs records. 3. Have insulin therapy for the first year after diabetes diagnosis.	Hsu et al, (2012)

Table 2 has shown the three different filtering rules. For selecting patients with type 2 diabetes disease, Yah et al, (2012) provided the method to select patients with type 2 diabetes (codes 250.x0, 250.x2) diagnosis. Another rules was proposed by Lin et al, (2013), which picked patients with diagnoses as diabetes (code 250) with one or more oral hypoglycaemic agent or insulin prescription (code 251.0, 251.1, and 251.2). The other rules were provided by Hsu et al, (2012). They have a more comprehensive rule of selecting diabetic patients. The selected the patients who have to fit one of the three conditions below: 1) Patients who diagnosed with diabetes (code 250.xx) and had one antidiabetic drugs prescription. Or 2) Patients have one

or more oral antidiabetic agents' prescription and diabetes-related illness ambulatory visit. Or 3) Patients have four or more ambulatory visits for diabetes-related illness within 1 year. And they excluded patients with these conditions below: 1) Patients who Diagnosed of type 1 diabetes (code 250.x1, 205.x3). Or 2) Patients who were younger than 30 years of age and don't have any oral anridiabetes drugs records. Or 3) Patients have insulin therapy for the first year after diabetes diagnosis.

After comparing previous works' filtering rules and consulting with domain experts, we use the rule mixed with the rule Lin et al, (2013) presented. Patients with the following ICD codes 250 or 250.0 or 250.00 and has prescribed one of blood glucose lowering drug or insulin or insulin analogues will be selected (ATC codes starts with A10A or A10B). Besides ICD code and drug prescription, we rule out patients with less than 85 medical records. As for ICD codes, we ruled out codes with appearances less than 100 times. After the whole process of filtering, the diabetic patient data set has 47,486 patients and 3,892 distinct ICD codes.

Table 3. Description of NHIRD diabetic patient data set.

	Diabetic patients (Total: 47,486)
Number of distinct ICD codes	3,892
Gender	Female: 24,519
	Male: 22,967
Age	45 and more: 45,775
	Under 45: 1,711
Average records/ month	3.18

3.2 Model

The model presented in this paper was originated by Yang et al. 2014. In the following sections we will define our problem and then explain the model in detail.

3.2.1 Problem definition

The definition of our problem is showed in figure 1. For data preprocessing, we treat each patient's record as a sequence of events, and for each ICD-9-CM code as an event. For each event sequence, we do two procedure, first is to put each sequence in the most likely class, second is to assign each event to a stage.

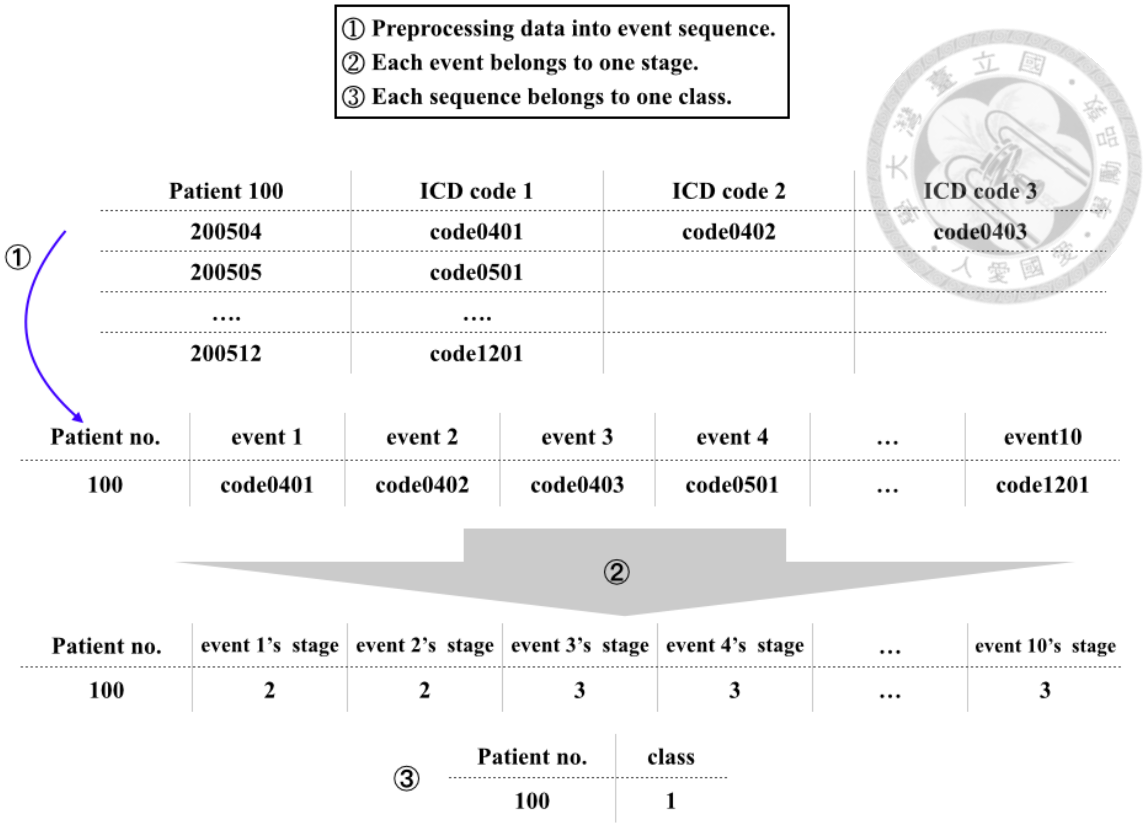


Figure 1. The procedure from data preprocessing to sequences assignment.

Table 4. Variables definition.

Variables	Definition
N	The number of patient sequences.
M	The number of possible ICD codes.
C	The number of classes.
K	The number of stages.
x_i	The i-th patient's sequence.
x_{ij}	The i-th patient j's ICD code.
c_i	The i-th patient's class.
s_{ij}	The i-th patient j's stage.

3.2.2 Model description

In this section, we describe our model in detail. First we define some variables. We also have the variables definition listed in table 4. We define each event sequence as $x_i, i = 1, \dots, N$ where N is the number of sequences, and x_i as patient i. For each $x_{ij} \in \{1, \dots, M\}$ where M is the number of possible ICD codes and x_{ij} as j represent the j-th ICD code of x_i (j is ordered by time). Each sequence belongs to a single class where $c_i \in \{1, \dots, C\}$, C as the number of classes. For each event $x_{ij} \in x_i$, we define $s_{ij} \in \{1, \dots, K\}$ to be the stage of the sequence x_i at time j, K as the number of stages.

We make stage s_{ij} as non-decreasing function. This constrain allow event sequence never go backward while progressing.

$$\forall i, j, k \; j \geq k \Longrightarrow s_{ij} \geq s_{ik} \tag{1}$$

Besides this constrain, we do not have any restriction that any sequence should progress through all stages, which indicates that some sequences may begin from intermediate stages while some sequences may never reach the end of stages.

Each x_{ij} is generated from a multinomial distribution with parameter $\Theta\left(c_i, s_{ij}\right) \in \mathbb{R}^M$. From this assumption, we can ensure that sequences from the same class and stage will have similar set of events.

$$x_{ij} \sim Mutinomial(\Theta(c_i, s_{ij})) \quad (2)$$

Lastly, we generate $\Theta\,(c_i, s_{ij})$ from a uniform Dirichlet distribution with a hyper parameter λ.

$$\Theta\,(c_i, s_{ij}) \sim Dirichlet(\lambda) \quad (3)$$

3.2.3 Model iteration

In this section, we provide the details on learning stages and classes for each event sequence. We have a set of event sequences $\{x_{ij}\}$. And we learn from cross validation for the number of classes C and stages K. Our goal is to find each x_{ij}, the stage s_{ij} for every $x_{ij} \in x_i$ and the class c_i for each sequence x_i. We specify that $\Theta = \{\Theta(\mathrm{p}, \mathrm{q}) | \mathrm{p} = 1, \ldots, \mathrm{C}, \mathrm{q} = 1, \ldots, \mathrm{K}\}$, and we find classes c_i, stages s_{ij} and Θ by maximizing the log likelihood:

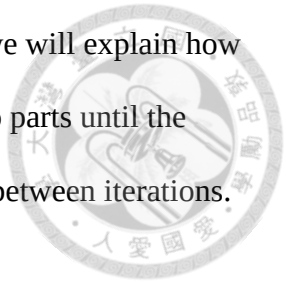
$$log\ P(\{x_i\}|\Theta, \{c_i\}, \{s_{ij}\}) \quad (4)$$

Since $\{x_{ij}\}$ are conditionally independent of each other given $\{c_i\}$, $\{s_{ij}\}$, (4) turns out to be formula (5). And our goal has become solving the optimization problem (6). Where $\{s_{ij}\}\nearrow_j$ is the constrain we specified in Eq. (1).

$$log\ P(\{x_i\}|\Theta, \{c_i\}, \{s_{ij}\}) = \sum_{i,j} logP(x_{ij}|\Theta(c_i, s_{ij})) \quad (5)$$

$$\mathrm{argmax}_{\{c_i\},\{s_{ij}\}\nearrow_j,\Theta} \sum_{i,j} logP(x_{ij}|\Theta(c_i, s_{ij})) \quad (6)$$

To solve (6), we update stages and classes, apart from Θ, iteratively until convergence. In the following sector 3.2.3.1 and 3.2.3.2 we will explain how we update the two parts respectively. We iterate these two parts until the classes and stages learned from the model do not change between iterations.



3.2.3.1 Updating Θ

For fixed stages and classes, we can compute Θ by maximize (6). Since $log\ P(x_{ij}|\Theta(c_i, s_{ij}))$ are conditionally independent of each other given classes and stages, we can separate (6) into:

$$\sum_{i,j} logP(x_{ij}|\Theta(c_i, s_{ij})) = \sum_{p=1}^{C}\sum_{q=1}^{K}\sum_{i,j}\mathbb{I}\{c_i = p \wedge s_{ij} = q\} logP(x_{ij}|\Theta(p,q)) \quad (7)$$

$\mathbb{I}$ is an indicator function. And we can find the optimal value of $\Theta(p,q)$ for each ICD code r in every classes from the probability formula (8) smoothed by Dirichlet parameter λ.

$$\Theta(p,q)_r = \frac{\lambda + \sum_{i,j}\mathbb{I}\{c_i = p \wedge s_{ij} = q \wedge x_{ij} = r\}}{M\lambda + \sum_{i,j}\mathbb{I}\{c_i = p \wedge s_{ij} = q\}} \quad (8)$$

$$r = 1, \ldots, M$$

3.2.3.2 Updating stages and classes

The second step is to update each sequence's stages and classes. This process is done through two steps. First, we assign stages within the event sequence for every class. After all classes have the stages assigned, we compare all classes to get the optimal class for the event sequence. To do the above twos steps for each event sequence, in detail we have to find the maximum of formula (9) for each event sequence.

$$\text{argmax}_{c_i,\{s_{ij}\}\nearrow_j} \sum_{i,j} logP(x_{ij}|\Theta(c_i, s_{ij})) \quad (9)$$

For each class, we pick the class with the highest likelihood, which means that we need to find the stage route for each class through solving (10) first before we solve (9).

$$\max_{\{s_{ij}\}\nearrow_j,\Theta} \sum_j logP(x_{ij}|\Theta(c_i, s_{ij})) \quad (10)$$

To solve (10), we transform finding route problem into the Longest Common Subsequence problem. We explain in detail on how we solve assigning stages in the next paragraph.

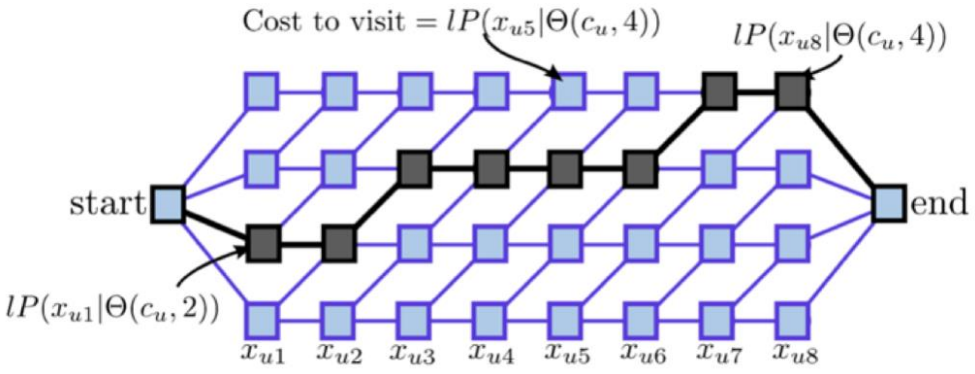


Figure 2. Dynamic programming procedure.
(Figure 2 was illustrated by Yang et al. 2014)

Dynamic programming procedure is an iterative procedure which is showed in figure 2. We consider each event sequence as a path of stages. With given class c_i, we update the stages s_{ij} by finding the black path in figure 2. While calculating cost $g(j,s)$ to reach j-th event at the s-th stage by forward recursion, there can only be two options: $g(j-1,s-1)$ going up or $g(j-1,s)$ staying at the same stage. In here $Cost(j,s) = logP(x_{ij}|\Theta(c_i,s))$.

After stage assigning problem is settled black path has been decided, which means (10) has been computed, we record the path and we go back to formula (9) and pick the class with the maximum value

$$g(j,s) = max\,(g(j-1,s) + Cost(j,s), g(j-1,s-1) + Cost(j,s)) \quad (11)$$

.

3.2.4 Cross validation

In the previous section we mentioned that we use fixed number for class and stage for modeling. Now we explain the procedure to get these number. We do cross validation by splitting each event sequence into 90% training and 10% testing dataset. After splitting the dataset, we trained our model with training data. And then we use the class and stage assigned by training to test data set. We sum up $\sum_j logP(x_{ij}|\Theta(c_i, s_{ij}))$ for testing data set and we look for model with the highest value.

3.2.5 Model initialization

Initially, we randomly assign classes for each event sequence and assign stages uniformly according to each event sequence length. As for the hyper-parameter λ, we choose 1 according to the previous finding.

4 Experiments

In experiments section, we tested our model's performance and inspect the results provided by our model. First, we do cross validation to select the class and stage number, and then we compare our model with two baseline models to check how well our model can predict events. Third, we will look deeper into each class to compare differences between classes. And then we random pick three patients to inspect how they progress between stages. Last, we calculate cross entropy to know if different classes but same stages will have similar diseases.

4.1 Cross validation

To find an optimal class and stage number for our model, we do cross validation to find them. We tried 2, 3 for class number and 1, 3, 5, 7, 9 for stage number for each class. The result gathered in figure 3 shows that model with 2 classes with 5 stages performs best among all others.

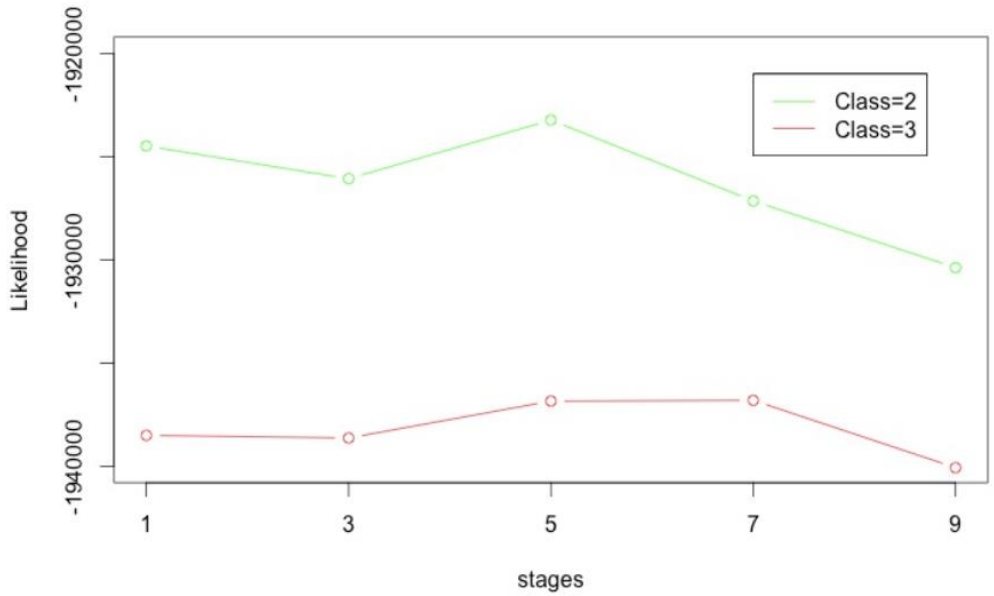


Figure 3. Cross validation for picking class number and stage number.

4.2 Predicting accuracy

We want to know how well our model can predict on missing events or future events. Therefore we construct 2 scenarios to test the accuracy, first is to predict the last visit of each patient record, second is to predict random visit of each patient record. Accuracy is measured using hit rate for top 10. We consider two baseline models to compare the results, Logistic regression and baseline 2.

For our model, we split each patient into training events and testing events. After fitting our model with training set, we use the classes and stages assigned from training to predict testing events. As for Logistic regression, we split the training events into feature events and response event to train the Logistic classifier. And we use the Logistic classifier to predict training events and test events. Baseline 2 is a simpler approach. We count 50 events before testing events and rank frequency to count accuracy. The training and testing method for Logistic regression and baseline 2 was illustrated in figure 4.

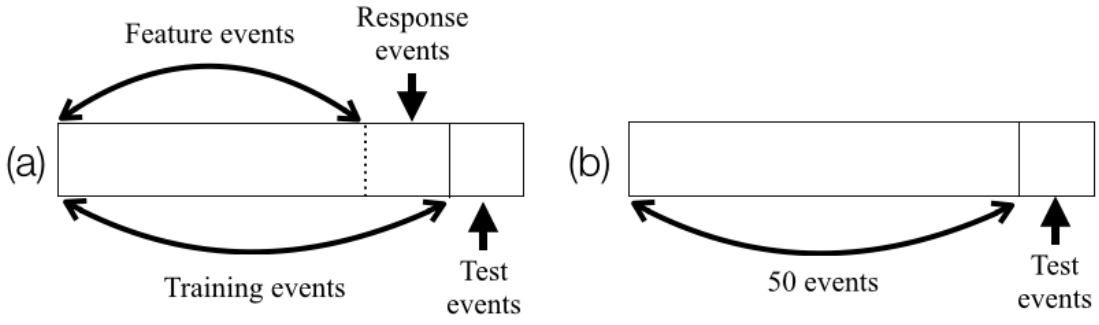


Figure 4. The split of training testing data set. (a) Baseline 1: Logistic regression; (b) Baseline 2.

Table 5. Performance of predicting last visit and random visit.

		Absolute accuracy	Relative to random gussing	Gain over baseline (%)
Last visit	Our model	0.4	1556.8	
	Logistic regression	0.04	155.68	1401.12
	Baseline 2	0.42	1634.64	-77.84
Random visit	Our model	0.35	1362.2	
	Logistic regression	0.05	194.6	1167.6
	Baseline 2	0.36	1401.12	-38.92

The results for three models are showed in table 5. In predicting last visit, our model has a better performance compares to Logistic regression and our model has the accuracy that is close to baseline 2. A similar result can be found in predicting random visit. It is worth noticing that the performance for random visit is generally lower than last visit, it may be that random visit doesn't have enough data to train and predict.

4.3 Theta inspection

In theta inspection, we want to know what are the differences between class 1 and class 2. For each class, we got the top 10 ICD codes that are most possible to get in each stage in table 5 and table 6.

Table 6. Modeling last visit's Class 1 with top 10 ICD codes and probability.

Top 10	Stage1	Stage 2	Stage 3	Stage 4	Stage 5
1	272.7 Lipidoses 脂肪代謝障礙 (0.044)	403 Necrobacillosis 高血壓性腎臟疾病 (0.057)	402 Whipple's disease 高血壓性心臟病 (0.091)	402 Whipple's disease 高血壓性心臟病 (0.061)	402 Whipple's disease 高血壓性心臟病 (0.066)
2	466.0 Acute bronchitis 急性支氣管炎 (0.018)	402 Whipple's disease 高血壓性心臟病 (0.022)	272.7 Lipidoses 脂肪代謝障礙 (0.026)	587 Renal sclerosis, unspecified 腎硬化 (0.031)	587 Renal sclerosis, unspecified 腎硬化 (0.035)
3	241.0 Nontoxic uninodular goiter 非毒性單一結節性甲狀腺腫 (0.015)	401.90 Essential hypertension, unspecified 本態性高血壓 (0.019)	403 Necrobacillosis 高血壓性腎臟疾病 (0.026)	403 Necrobacillosis 高血壓性腎臟疾病 (0.025)	272.7 Lipidoses 脂肪代謝障礙 (0.018)
4	571.49 Other chronic hepatitis 其他慢性肝炎 (0.013)	295.64 Residual schizophrenia, chronic with acute exacerbation 慢性伴有急性發作，殘餘型精神分裂症 (0.017)	587 Renal sclerosis,unspecified 腎硬化 (0.025)	272.7 Lipidoses 脂肪代謝障礙 (0.020)	250.42 Diabetes with renal manifestations, Type II 糖尿病併有腎病表徵之第二型 (0.017)
5	401.1 Benign essential hypertension 良性本態性高血壓 (0.012)	272.7 Lipidoses 脂肪代謝障礙 (0.017)	466.0 Acute bronchitis 急性支氣管炎 (0.017)	250.42 Diabetes with renal manifestations, Type II 糖尿病併有腎病表徵之第二型 (0.016)	466.0 Acute bronchitis 急性支氣管炎 (0.015)
6	414.01 Coronary atherosclerosis of native coronary artery 自體的冠狀動脈粥樣硬化 (0.011)	466.0 Acute bronchitis 急性支氣管炎 (0.014)	250.92 Diabetes with unspecified complication, Type II 糖尿病併有併發症之第二型 (0.014)	466.0 Acute bronchitis 急性支氣管炎 (0.015)	415.1 Pulmonary embolism and infarction 肺栓塞及梗塞 (0.015)
7	402.00 Malignant hypertensive heart disease without congestive heart failure 惡性高血壓性心臟病無充血性心臟衰竭 (0.009)	414.01 Coronary atherosclerosis of native coronary artery 自體的冠狀動脈粥樣硬化 (0.012)	250.62 Diabetes with neurological manifestations, Type II 糖尿病併有神經疾病表徵之第二型 (0.013)	250.62 Diabetes with neurological manifestations, Type II 糖尿病併有神經疾病表徵之第二型 (0.013)	403 Necrobacillosis 高血壓性腎臟疾病 (0.014)
8	272.1 Pure hyperglyceridemia 純高甘油脂血症 (0.008)	414.0 Coronary atherosclerosis 冠狀動脈粥樣硬化 (0.010)	415.1 Pulmonary embolism and infarction 肺栓塞及梗塞 (0.012)	250.92 Diabetes with unspecified complication, Type II 糖尿病併有併發症之第二型 (0.013)	272.4 Other and unspecified hyperlipidemia 其他高脂質血症 (0.013)
9	274.10 Gouty nephropathy, unspecified 痛風性腎病變 (0.007)	434.91 Unspecified cerebral artery occlusion with cerebral infarction 伴有腦梗塞之腦動脈阻塞 (0.010)	250.42 Diabetes with renal manifestations, Type II 糖尿病併有腎病表徵之第二型 (0.012)	564.2 Postgastric surgery syndromes 胃手術後徵候群 (0.011)	250.62 Diabetes with neurological manifestations, Type II 糖尿病併有神經疾病表徵之第二型 (0.013)
10	461 Jakob-Creutzfeldt disease 急性鼻竇炎 (0.006)	780.50 Sleep disturbances, unspecified 睡眠障礙 (0.009)	272.4 Other and unspecified hyperlipidemia 其他高脂質血症 (0.010)	366.14 Posterior subcapsular polar senile cataract 後囊下極部老年性白內障 (0.011)	564.2 Postgastric surgery syndromes 胃手術後徵候群 (0.012)

Table 7. Modeling last visit's Class 2 with top 10 ICD codes and probability.

Top 10	Stage1	Stage 2	Stage 3	Stage 4	Stage 5
1	272.7 Lipidoses 脂肪代謝障礙 (0.074)	401.90 Essential hypertension, unspecified 本態性高血壓 (0.046)	402 Hypertensive heart disease 高血壓性心臟病 (0.090)	402 Hypertensive heart disease 高血壓性心臟病 (0.049)	402 Hypertensive heart disease 高血壓性心臟病 (0.054)
2	466.0 Acute bronchitis 急性支氣管炎 (0.055)	466.0 Acute bronchitis 急性支氣管炎 (0.037)	272.7 Lipidoses 脂肪代謝障礙 (0.034)	466.0 Acute bronchitis 急性支氣管炎 (0.030)	272.7 Lipidoses 脂肪代謝障礙 (0.030)
3	571.49 Other chronic hepatitis 其他慢性肝炎 (0.028)	402 Hypertensive heart disease 高血壓性心臟病 (0.033)	466.0 Acute bronchitis 急性支氣管炎 (0.031)	272.7 Lipidoses 脂肪代謝障礙 (0.028)	466.0 Acute bronchitis 急性支氣管炎 (0.026)
4	521.2 Abrasion 牙齒磨損 (0.018)	272.7 Lipidoses 脂肪代謝障礙 (0.029)	272.4 Other and unspecified hyperlipidemia 其他高脂質血症 (0.031)	272.4 Other and unspecified hyperlipidemia 其他高脂質血症 (0.023)	272.4 Other and unspecified hyperlipidemia 其他高脂質血症 (0.023)
5	461 Jakob-Creutzfeldt disease 急性鼻竇炎 (0.017)	272.4 Other and unspecified hyperlipidemia 其他高脂質血症 (0.018)	466.11 Acute bronchiolitis due to respiratory syncytial virus(RSV) 呼吸道融合病毒引起之急性細支氣管炎 (0.015)	780.50 Sleep disturbances, unspecified 睡眠障礙 (0.016)	415.1 Pulmonary embolism and infarction 肺栓塞及梗塞 (0.015)
6	466.11 Acute bronchiolitis due to respiratory syncytial virus(RSV) 呼吸道融合病毒引起之急性細支氣管炎 (0.015)	272.3 Hyperchylomicronemia 高乳糜微粒血症 (0.018)	401.90 Essential hypertension, unspecified 本態性高血壓 (0.015)	466.11 Acute bronchiolitis due to respiratory syncytial virus(RSV) 呼吸道融合病毒引起之急性細支氣管炎 (0.015)	466.11 Acute bronchiolitis due to respiratory syncytial virus(RSV) 呼吸道融合病毒引起之急性細支氣管炎 (0.014)
7	V70.6 Health examination in population surveys 全人口普查之健康檢查 (0.013)	784.2 Swelling, mass, or lump in head and neck 頸部及頭部之腫脹或腫塊 (0.017)	780.50 Sleep disturbances, unspecified 睡眠障礙 (0.014)	250.92 Diabetes with unspecified complication, Type II 糖尿病併有併發症之第二型 (0.013)	250.42 Diabetes with renal manifestations, Type II 糖尿病併有腎病表徵之第二型 (0.012)
8	523.8 Other specified periodontal diseases 其他特定牙周疾病 (0.010)	571.49 Other chronic hepatitis 其他慢性肝炎 (0.015)	250.42 Diabetes with renal manifestations, Type II 糖尿病併有腎病表徵之第二型 (0.012)	401.90 Essential hypertension, unspecified 本態性高血壓 (0.012)	401.90 Essential hypertension, unspecified 本態性高血壓 (0.012)
9	537.5 Gastroptosis 胃下垂 (0.010)	275.40 Unspecified disorders of calcium metabolism 鈣代謝疾患 (0.012)	250.62 Diabetes with neurological manifestations, Type II 糖尿病併有神經疾病表徵之第二型 (0.012)	250.62 Diabetes with neurological manifestations, Type II 糖尿病併有神經疾病表徵之第二型 (0.012)	250.92 Diabetes with unspecified complication, Type II 糖尿病併有併發症之第二型 (0.012)
10	537.89 Other specified disorders of stomach and duodenum 胃及十二指腸之其他特定疾病 (0.010)	466.11 Acute bronchiolitis due to respiratory syncytial virus(RSV) 呼吸道融合病毒引起之急性細支氣管炎 (0.011)	403 Hypertensive renal disease 高血壓性腎臟疾病 (0.012)	461 Acute sinusitis 急性鼻竇炎 (0.011)	780.50 Sleep disturbances, unspecified 睡眠障礙 (0.011)

After comparing two classes top 10 ICD codes in table 6 and 7, we may infer that class 1 is a type of diabetic patients who has higher possibility to have kidney related diseases, while class 2 is a type of diabetic patients who is more likely to have lung related diseases.

4.4 Patient examples

We are also interesting in how patient evolve from stage to stage. Three patients' data are picked for us to inspect the change of stages and what diseases they have during each stage. We omitted diseases which appear less times.

The first one is patient no. 851, she was 78 years old and was assigned to class 2. According to her record date, she hasn't been going to the hospital for long, and she has been in stage 3 since 200102.

Table 8. Patient 851's diabetes progression status.

Stage 2 (200909~201102)	Stage 3 (201102~201211)
401.1 Benign essential hypertension 24 times 良性本態性高血壓 24次 465.9 Acute upper respiretory infections of unspecified site 20 times 急性上呼吸道感染 20次	465.9 Acute upper respiretory infections of unspecified site 15 times 急性上呼吸道感染 15次 401.1 Benign essential hypertension 12 times 良性本態性高血壓 12次 461.9 Acute sinusitis, unspecified 1 times 急性鼻竇炎 1次

The second patient is patient no. 3306, he was 69 years old and was assign to class 1. He has climbed from stage 1 to stage 5 in only one month, the progressing speed is fast.

Table 9. Patient 3306's diabetes progression status.

Stage 1 (200901~200902)	Stage 5 (200902~201212)
401.9 Essential hypertension, unspecified 2 times 本態性高血壓　2次	401.9 Essential hypertension, unspecified 65 times 本態性高血壓　65次 727.00 Synovitis and tenosynovitis, unspecified 19 times 滑膜炎及腱鞘炎　19次 272.4 Other and unspecified hyperlipidemia 15 times 其他高脂質血症　15次

The last one is patient no. 4569, she was 76 years old and was assigned to class 1. She has the longest record of the three all, and she has turned from stage 1 to stage 3 in two months.

Table 10. Patient 4569's diabetes progression status.

Stage 1 (200201~200201)	Stage 2 (200201~200202)	Stage 3 (200202~201212)
465.9 Acute upper respiretory infections of unspecified site 3 times 急性上呼吸道感染 3次	571.40 Chronic hepatitis, unspecified 3 times 慢性肝炎　3次 401.1 Benign essential hypertension 2 times 良性本態性高血壓 2次	401.9 Essential hypertension, unspecified 122 times 本態性高血壓 122次 274.0 Gouty arthropathy 66 times 痛風性關節病變 66次

After look into the above three patient records, we know that patients usually don't cross all stages. In lower stages, they tend to climb fast. On the other hand, the

higher their stage number is the slower they evolving, in other words, they stay at higher stages for longer time period.

4.5 Cross entropy

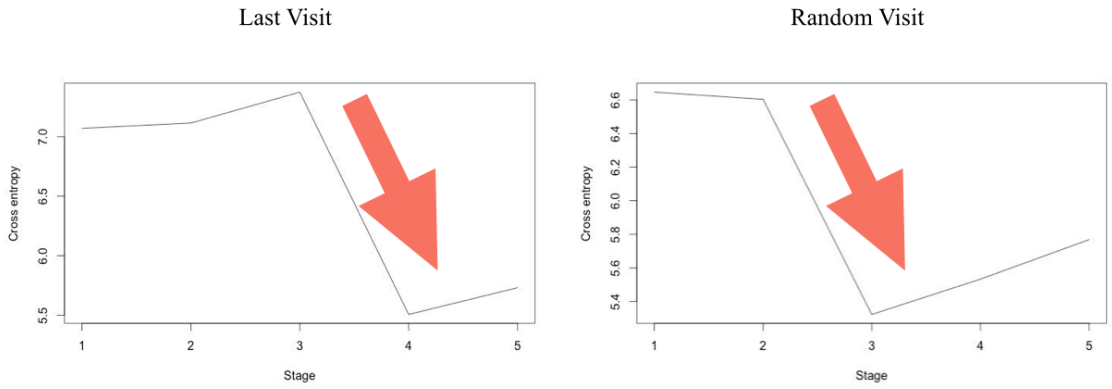
We also want to know that whether different classes with same stage have the same diseases or not as the stages went higher. To find out the answer to this question, we need to compute the cross entropy between classes. We calculate cross entropy using formula (12). (Danescu-Niculescu-Mizil et al. 2013)

$$H_s(c_1, c_2) = H'_s(c_1, c_2) + H'_s(c_2, c_1) \tag{12}$$

$$H'_s(c_1, c_2) = E_{x_{ij}|c_i=c_2, s_{ij}=s}[-logP(x_{ij}|\Theta(c_1, s))] \tag{13}$$

After cross entropy value has been computed, we got the results plotted in figure 5. We can conclude from figure 5 that as the stage go higher, the diseases in different class same stage will be more alike, since the line decrease after stage 3. This phenomenon is quit identical to chronic diseases' progression. For a chronic disease like diabetes, when the disease is getting worse, the symptoms will be more common, which match our results.

Figure 5. Cross entropy for last visit and random visit stages.



5 Conclusions

In this paper, we use a model made by previous study to model diabetic patient data. Our experiments showed that our model's performance is consistent with previous finding. In addition, our model's accuracy is better than logistic regression and is almost as good as baseline 2. As for the classes we found, the results have provided some insights for different type of diabetic patients. Moreover, by looking into each patient evolving pattern, we find that after moving to higher stage, i.e. 3 or above, the evolving rate is slower, which is a reasonable phenomenon for chronic diseases. Lastly, the cross entropy test has suggested the same trend as chronic disease's progression pattern.

However, there are still some improvements that can be done in the future, for example, allowing each patient sequence belongs to multiple classes or using different chronic disease patient data.

6 References

Y. C. Chen, W. Y. Chiou, S. K. Hung, Y. C. Su, and S. J. Hwang. 2013. Hepatitis C virus itself is a causal risk factor for chronic kidney disease beyond traditional risk factors: a 6-year nationwide cohort study across Taiwan. *BMC Nephrology*, 6;14:187.

W. C. Chiu, Y. T. Tsan, S. L. Tsai, C. J. Chang, J. D. Wang, P. C. Chen, and hDATa Research Group. 2014. Hepatitis C viral infection and the risk of dementia. *European Journal of Neurology*, 21 (8):1068-e59.

M. J. Cohen, A. D. Grossman, D. Morabito, M. M. Knudson, A. J. Butte and G. T. Manley. 2010. Identification of complex metabolic states in critically injured patients using bioinformatic cluster analysis. *Critical Care*, 14 (1):R10.

C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, C. Potts. 2013. No country for old members: user lifecycle and linguistic change in online communities. *WWW*, 307-318.

R. S. Doody, V. Pavlik, P. Massman, S. Rountree, E. Darby, and W. Chan. 2010. Predicting progression of Alzheimer's disease. *Alzheimer's Research & Therapy,* 2:2.

H. M. Fonteijn, M. Modat, M. J. Clarkson, J. Barnes, M. Lehmann, N. Z. Hobbs, R. I. Scahill, S. J. Tabrizi, S. Ourselin, N. C. Fox, and D. C. Alexander. 2012. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImange*, 60 (3):1880-1889.

Y. C. Hsu, J. T. Lin, H. J. Ho, Y. H. Kao, Y. T. Huang, N. W. Hsiao, M. S. Wu, Y. Y. Liu,

and C. Y. Wu. 2014. Antiviral treatment for hepatitis C virus infection is associated with improved renal and cardiovascular outcomes in diabetic patients. *Hepatology*, 59 (4):1293-1302.

C. C. Hsu, C. H. Lee, M. L. Wahlqvist, H. L. Huang, H. Y. Chang, L. Chen, S. F. Shih, S. J. Shin, W. C. Tsai, T. Chen, C. T. Huang, J. S. Cheng. Poverty increases type 2 diabetes incidence and inequality of care despite universal health coverage. *Diabetes Care*, 35 (11):2286-2292.

C. H. Lin, and W. H. Sheu. 2013. Hypoglycaemic episodes and risk of dementia in diabetes mellitus: 7-year follow-up study. *Journal of internal medicine*, 273 (1):102-110.

D. Mould. 2012. Models for disease progression: new approaches and uses. *Clinical Pharmacology & Therapeutics*, 92 (1):125-131.

T. M. Post, J. I. Freijer, J. DeJongh, and M. Danhof. 2005. Disease system analysis: basic disease progression models in degenerative disease. *Pharmaceutical research*, 22 (7):1038-1049.

A. Raj, A. Kuceyeski, and M. Weiner. 2012. A network diffusion model of disease progression in dementia. *Neuron*, 73 (6):1204-1215.

X. Wang. D. Sontag, and F. Wang. 2014. Unsupervised learning of disease progression models. *KDD*, 85-94.

J. Yang, J. J. McAuley, J. Leskovec, P. LePendu, and N. Shah. 2014. Finding progression stages in time-evolving event sequences. *WWW*, 783-794.

H. T. Yeh, C. F. Hsieh, Y. W. Tsai, and W. F. Huang. 2012. Effects of thiazolidinediones on cardiovascular events in patients with type 2 diabetes mellitus after drug-eluting stent implantation: a retrospective cohort study using the national health insurance database in Taiwan. *Clinical Therapy*, 34 (4):885-893.

J. Zhou, J. Liu, V. A. Narayan, and J. Ye.. 2012. Modeling disease progression via fused sparse group lasso. *KDD*, 1095-1103.

J. Zhou, L. Yuan, J. Liu, J. Ye.. 2011. A multi-task learning formulation for predicting disease progression. *KDD*, 814-822.