Disease Prediction with a Maximum Entropy Method

Michael Shub,[*] Qing Xu,[†] Xiaohua(Michael) Xuan[‡§]

Abstract

In this paper, we propose a maximum entropy method for predicting disease risks. It is based on a patient's medical history with diseases coded in ICD-10 which can be used in various cases. The complete algorithm with strict mathematical derivation is given. We also present experimental results on a medical dataset, demonstrating that our method performs well in predicting future disease risks and achieves an accuracy rate twice that of the traditional method. We also perform a comorbidity analysis to reveal the intrinsic relation of diseases.

1 Introduction

Disease prediction is an effective way to assess a person's health status. Studies have shown that in many cases, there are identifiable indicators or preventable risk factors before the onset of the patient's disease. These early warnings can effectively reduce the individual's risk of disease. Theoretically, this can reduce the number of treatments needed and increase the necessary effective interventions. However, the combination of problem factors caused by different diseases and the patient's past medical history are so complicated that no doctor can fully understand all of this. Currently, doctors can use family and health history and physical examinations to estimate the patients risk and guide laboratory tests to further evaluate the patient's health. However, these sporadic and qualitative "risk assessments" are usually only for a few diseases, depending on the experience, memory and time of the particular doctor. Therefore, the current medical care is after the fact. Once the symptoms of the disease appear, it is involved, rather than actively treating or eliminating the disease as soon as possible.

Today the prevailing model of prospective heath care is firmly based on the genome revolution. Indeed, technologies ranging from linkage equilibrium and candidate gene association studies to genome wide associations have provided an extensive list of disease-gene associations, offering us detailed information on mutations, SNPs, and the associated likelihood of developing specific disease phenotypes.

The basic assumption behind the research is that once we have classified all disease-related mutations, we can use various molecular biomarkers to predict each individual's susceptibility to future diseases, thus bringing us into a predictive medicine era. However, these rapid advances have also revealed the limitations of genome-based methods. Considering that the signals provided by most disease-related SNPs or mutations are very weak, it is becoming increasingly clear that the prospect of genome-based methods may not be realized soon.

Does this mean that prospective disease prediction methods must wait until genomics methods are sufficiently mature? Our purpose is to prove that the method based on medical history provides hope for the prospective prediction of disease.

In this paper, we mainly study the disease prediction and comorbidity of diseases. Our approach is distinctly different in that we are trying to build a general predictive system which can utilize a less constrained feature space, i.e. taking into account all available demographics and previous medical history. Moreover, we rely primarily on ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification) codes (see Section 2) for making predictions to account for the previous medical history, rather than specialized test results.

---

[*]Math Department, City College and the Graduate Center of CUNY.

[†]UniDT, Shanghai, China.

[‡]UniDT, Shanghai, China.

[§]This work was partially supported by the Smale Institute.

2 Data

2.1 Source Data and Population

Our database comprises the medical records of 354,552 patients in China with a total of 2,904,257 hospital visits. The data was originally compiled from Insurance claims during 2007 to 2017. Such medical records are highly complete and accurate, and they are frequently used for epidemiological and demographic research.

The input for our methods consists of each patient's personal information, such as gender, birthday, treatment-date, and diagnosis history, provided per patient's visit. Each data record consists of a hospital visit, represented by a patient ID and a diagnosis code per visit, as defined by the International Classification of Diseases, Tenth Revision,Clinical Modification(ICD-10-CM). The International Statistical Classification of Diseases and Related Health Problems (ICD) provides codes to classify diseases and a wide variety of signs, symptoms, abnormal findings, social circumstances, and external causes of injury or disease. It is published by the World Health Organization.

Each disease or health condition is given a unique code, and can be up to 6 characters long, such as A01.001. The first character is a letter while the others are digits. ICD-10 codes are hierarchical in nature, so the 6 characters codes can be collapsed to fewer characters identifying a small family of related medical conditions. For instance, code A01.001 is a specific code for typhoid fever. This code can be collapsed to A01.

Moreover, we classify diseases of the same category into one class. For example, A90 is the code for Dengue fever (classical dengue) and A91 is the code for Dengue hemorrhagic fever. We classify them into the same class named F_A90. Thus, the 20 thousand origin ICD-10 codes are classified into 429 classes.

A sample patient medical history is shown in Table 1. Each line represents one hospital visit. Demographic data are also available.

Table 1: Medical History Sample

patient_id	gender	treatment_date	code
14532	F	2011-10-15	F_M47
14532	F	2011-11-19	F_N91
14532	F	2012-10-09	F_L20
14532	F	2012-10-19	F_N60
14532	F	2013-05-08	F_B37
14532	F	2013-06-04	F_H10
14532	F	2013-06-15	F_K04
14532	F	2013-08-23	F_L20

In our medical database, the number of visits per patient ranges from 1 to 491, with a median of 4. Also, the average is 8.19. Table 2 shows the 20 most prevalent diseases in our database.

Table 2: 20 Most Prevalent Diseases

Disease	Prevalence
Acute upper respiratory infection	20.88%
Hypertension and its complications	7.35%
Dermatitis and pruritus	3.75%
Gastritis and duodenitis	3.49%
Chronic bronchitis	3.28%
Pulp, gum, and alveolar ridge diseases	3.15%
Hard tissue disease of teeth	2.73%
Abnormal uterine and vaginal bleeding	2.42%
Chronic rhinitis, nasopharyngitis and pharyngitis	1.99%
Non-infectious gastroenteritis and colitis	1.97%
Chronic ischemic heart disease	1.93%
Inflammation of the vagina and vulva	1.72%
Pneumonia	1.63%
Abnormal thyroid (parathyroid) function	1.62%
Other diabetes	1.56%
Backache	1.54%
Acute lower respiratory infection	1.51%
Cervical disc disease	1.44%
Type II diabetes	1.37%
Female pelvic inflammatory disease	1.18%

2.2 Quantifying the Strength of Comorbidity Relationships

In order to measure the correlation from disease comorbidity, we need to quantify the intensity of disease comorbidity by introducing the concept of distance between the two diseases. One difficulty of this method is that there are biases in different statistical measures, which overestimate or underestimate the relationship between rare or epidemic diseases. Given that the number of diagnoses (prevalence) for a particular disease follows a long tail distribution, these biases are important, which means that although most diseases are rarely diagnosed, a small number of diseases have been diagnosed in a large part of the population.

Therefore, quantifying comorbidity usually requires us to compare diseases that affect dozens of patients with diseases that affect millions of patients.

We will use two comorbidity measures to quantify the distance between two diseases: The Absolute Logarithmic Relative Risk (ALRR) and ϕ-correlation(ϕ).

The Emperical Relative Risk of observing a pair of diseases i and j affecting the same patient is given by

$$ERR_{ij} = \frac{C_{ij}N}{P_i P_j},$$

where C_{ij} is the number of patients affected by both diseases, N is the total number of patients in the population and P_i and P_j are the prevalences of diseases i and j.

$$ERR_{ij} = \frac{\frac{C_{ij}}{N}}{\frac{P_i}{N} \cdot \frac{P_j}{N}}.$$

Thus, the Relative Risk is defined as

$$RR_{ij} = \frac{p_{ij}}{p_i p_j},$$

Here, p_{ij} is the transition probability from disease i to disease j and p_i is the incidence probability of disease i. The Absolute Logarithmic Relative Risk is defined as

$$M_{ij} = \left| \log \left(\frac{p_{ij}}{p_i p_j} \right) \right|.$$

The emperical ϕ-correlation, which is Pearsons correlation for binary variables, can be expressed mathematically as

$$\frac{C_{ij}N - P_iP_j}{\sqrt{P_iP_j(N-P_i)(N-P_j)}} = \frac{\frac{C_{ij}}{N} - \frac{P_i}{N} \cdot \frac{P_j}{N}}{\sqrt{\frac{P_i}{N} \cdot \frac{P_j}{N}(1 - \frac{P_i}{N})(1 - \frac{P_j}{N})}}.$$

Therefore, the ϕ-correlation is defined as

$$\phi_{ij} = \frac{p_{ij} - p_ip_j}{\sqrt{p_ip_j(1-p_i)(1-p_j)}}.$$

These two comorbidity measures are not completely independent of each other, as they both increase with the number of patients affected by both diseases, yet both measures have their intrinsic biases. For example, RR overestimates relationships involving rare diseases and underestimates the comorbidity between highly prevalent illnesses, whereas ϕ accurately discriminates comorbidities between pairs of diseases of similar prevalence but underestimates the comorbidity between rare and common diseases.

3 Methodology

In this section. We will formulate the maximum entropy method we used to predict disease risk.

3.1 Some notations

Suppose there are n diseases and N records. Let us use d_i to denote disease i. A record is a pair of diseases (d_i, d_j) which means that there is a patient with a diagnosis of disease d_j simultaneously or after disease d_i. Let us use S_k to denote record k $(1 \le i \le n, 1 \le k \le N)$.

Assume that $S_k = (f(k), g(k))$. Here, f and g are maps.

$$f, g : \{1, 2, \cdots, N\} \to \{1, 2, \cdots, n\}.$$

$f(k)$ is called the first disease and $g(k)$ is called the second disease in record k.

In this paper, we assume that f and g are surjective. If f is not surjective, we can remove the diseases with indexes in $\{1, 2, \cdots, n\} \backslash \mathrm{Range}(f)$ from the medical history. Then the surjective assumption can be satisfied for f. The same can be done for g.

Assume that

$$X_{ik} = \begin{cases} 1, & f(k) = i \\ 0, & f(k) \ne i \end{cases}, \quad Y_{kj} = \begin{cases} 1, & g(k) = j \\ 0, & g(k) \ne j \end{cases}.$$

Denote by

$$\chi_C = \begin{cases} 1, & x \in C \\ 0, & x \notin C \end{cases}.$$

Let $A = XY, B = YX$. Then

$$A_{ij} = \sum_{k=1}^{N} X_{ik}Y_{kj} = \sum_{k=1}^{N} \chi_{\{f(k)=i, g(k)=j\}}$$

is the number of patients who suffer from disease i before disease j.

$$B_{km} = \sum_{j=1}^{n} Y_{kj}X_{jm} = \sum_{j=1}^{n} \chi_{\{f(m)=j, g(k)=j\}}.$$

B is a matrix with entries ranged in $\{0, 1\}$. $B_{km} = 1$ if and only if there is a disease j such that the patient in record k suffer a second disease j and the patient in record m suffer a first disease j. Our task is to evaluate the transition probability from record k to record m.

3.2 Entropy for Markov Chains

Suppose M is a non-negative matrix. If for each $k, m \in \{1, 2, \cdots, N\}$, there exists $l \geq 1$ such that $(M^l)_{k,m} > 0$, then M is said to be irreducible.

Now we define the entropy for Markov chains. A matrix B is called a skeleton matrix if its entries are either 0 or 1. A non-negative matrix P as called a Markov transition matrix if

$$\sum_{m=1}^{N} P_{k,m} = 1, \quad \forall \ k = 1, 2, \cdots, N.$$

Moreover, if $P_{k,m} > 0 \Leftrightarrow B_{k,m} = 1$, then P is called the Markov transition matrix associated with the skeleton matrix B.

For a non-negative vector $p = (p_1, \cdots, p_N)$, if

$$\sum_{k=1}^{N} p_k = 1, \quad pP = p,$$

then p is called a stationary distribution of P.

For a non-negative matrix $W = (w_{k,m})$, if

$$\sum_{k,m=1}^{N} w_{k,m} = 1,$$

and for all $1 \leq k \leq N$,

$$\sum_{m=1}^{N} w_{m,k} = \sum_{m=1}^{N} w_{k,m}.$$

Then W is called a Markov weight matrix.

Here are some connections between Markov transition matrix and Markov weight matrix.

For a Markov transition matrix P with stationary distribution p. Define

$$w_{k,m} = p_k P_{k,m}.$$

Then it is easy to verify that $W = (w_{k,m})$ is a Markov weight matrix.

On the contrary, given a Markov weight matrix $W = (w_{k,m})$, set

$$p_k = \sum_{m=1}^{N} w_{k,m}, \quad P_{k,m} = \frac{w_{k,m}}{p_k}.$$

Then W is the Markov weight matrix associated with P.

Now we define the entropy for a Markov transitin matrix. First, let us consider the chain with length l.

$$H_l(P) = -\sum_{i_0=1}^{N} \sum_{i_1=1}^{N} \cdots \sum_{i_l=1}^{N} p_{i_0} P_{i_0,i_1} \cdots P_{i_{l-1},i_l} \log(p_{i_0} P_{i_0,i_1} \cdots P_{i_{l-1},i_l})$$

$$= -\sum_{i_0=1}^{N} p_{i_0} \log(p_{i_0}) - l \sum_{k=1}^{N} \sum_{m=1}^{N} p_k P_{k,m} \log(P_{k,m}).$$

So the entropy for a chain with unit length is defined as

$$H(P) = \lim_{l \to \infty} \frac{H_l(P)}{l} = -\sum_{k=1}^{N} \sum_{m=1}^{N} p_k P_{k,m} \log(P_{k,m}).$$

3.3 Maximum Entropy Theorem

The principle of maximum entropy is a basic principle in information theory(see e.g. [6]). It states that the probability distribution which best represents the current state of knowledge is the one with largest entropy. Since the distribution with the maximum entropy is the one that makes the fewest assumptions about the true distribution of data, the principle of maximum entropy can be seen as an application of Occam's razor(see e.g. [7]).

Theorem 3.1. *Suppose B is irreducible. λ is the maximum eigenvalue of B, and $l = (l_1, \cdots, l_N), r = (r_1, \cdots, r_N)^T$ are the corresponding left and right eigenvectors with*

$$\sum_{k=1}^{N} l_k r_k = 1.$$

Then the entropy of the Markov chain associated with the skeleton matrix B attains the maximum $\log \lambda$ when

$$w_{k,m} = \frac{1}{\lambda} B_{k,m} l_k r_m, \quad 1 \leq k, m \leq N.$$

Here, $W = (w_{k,m})$ is the weight matrix for B.

Proof. see Appendix.

$\square$

Theorem 3.2. *Suppose A is irreducible. λ is the maximum eigenvalue of A, and $L = (L_1, \cdots, L_n), R = (R_1, \cdots, R_n)^T$ are the corresponding left and right eigenvectors with*

$$\sum_{j=1}^{n} L_j R_j = 1.$$

Then the entropy of the Markov chain associated with the skeleton matrix B attains the maximum $\log \lambda$ when

$$v_{ij} = \frac{1}{\lambda} A_{ij} L_i R_j, \quad 1 \leq i, j \leq n.$$

Here, $V = (v_{ij})$ is the weight matrix for A.

Proof. see Appendix.

$\square$

Remark 3.1. *Recall that we have assumed that f and g are surjective. Therefore, X and Y have rank n. Since $A = XY, B = YX$, we have that*

$$\det(\lambda I_N - YX) = \lambda^{N-n} \det(\lambda I_n - XY).$$

Therefore, the eigenvalues of A and B are the same except for the zeros. In particular, the largest eigenvalue of A and B are the same.

Moreover, $YA = BY$. Suppose a is an eigenvector of A with eigenvalue μ, then

$$\mu Y a = Y A a = B Y a.$$

Since $a \neq 0$ and Y is injective as a map from $\mathbb{R}^n$ to $\mathbb{R}^N$, $Ya \neq 0$. Hence, Ya is the eigenvector of B with eigenvalue μ.

3.4 Algorithm for Probability Estimation

Following is the algorithm for estimating the related probability.

Algorithm 1: Probability Estimation

Step 1. Compute the matrix A.

Step 2. Use power method to compute the maximum eigenvalue λ of A with the corresponding left and right eigenvectors L_0 and R_0. Let

$$L = \frac{L_0}{\sqrt{L_0 \cdot R_0}}, \quad R = \frac{R_0}{\sqrt{L_0 \cdot R_0}}.$$

Step 3. Compute the weight matrix V as follows.

$$v_{ij} = \frac{1}{\lambda} A_{ij} L_i R_j, \quad 1 \leq i, j \leq n.$$

Step 4. Compute the transition probability as follows.

$$p_{ij} = \frac{v_{ij}}{\sum\limits_{l=1}^{n} v_{il}}.$$

Step 5. Compute the stationary distribution as follows.

$$p_i = L_i R_i.$$

3.5 Method for Disease Prediction

The prediction task is to predict the diseases that a person is most likely to have if we know that he has already suffered from diseases $(d_{i_1}, d_{i_2}, \cdots, d_{i_T})$ which are ordered by occurance.

We first calculate the probability by Algorithm 1. Then we construct the following quantity

$$r_j = \frac{1}{T} \sum_{l=1}^{T} p_{i_l, j}. \quad (1 \leq j \leq n)$$

Then we sort the r_j and choose the top 5 disease as the predicted diseases for a person.

We also make an additional assumption, that is, the latest disease take a highest weight. Thus, some modifications are made. We first construct a decreasing sequence $\{a_n\}_{n \geq 1}$ such that $a_n > 0$. (For example, $a_n = 1/n^2$). Then we modify r_j as

$$r_j = \frac{\sum\limits_{l=1}^{T} a_{T+1-l} \cdot p_{i_l, j}}{\sum\limits_{l=1}^{T} a_l}.$$

And we use the modified r_j to choose the top 5 disease as the predicted diseases for a person.

4 Experiments

4.1 Data Cleaning

The diseases are classified by F-code as described in section 2, and there are 429 F-coded diseases in total. If someone suffered from the same disease for many times, then we keep the earliest record and remove the others. For example, for the patient with patientid=123770, she suffered from mucopurulent conjunctivitis on 2015-12-09 and 2016-03-08, then the record with 2016-03-08 is removed from the history.

We clean the data and collect the records of the same patient together into one record. The history column recorded the patient's disease history and the diseases are sorted by time and separated by a comma.

The following table is a sample of the cleaned data.

Table 3: Sample of Cleaned Data

patient_id	history
123770	F_H10,F_H00
135086	F_M65,F_J00,F_K29,F_K01,F_K04
400195	F_J00,F_K29
3218331	F_J00,F_J40
119151	F_J00,F_N60
102519	F_J00,F_L50,F_E34,F_E01
1503387	F_K29,F_K01,F_I83
7044682	F_J00,F_J20,F_J40
182660	F_E01,F_J00,F_J20
1888934	F_K31,F_K29,F_K22,F_K50,F_J00,F_J40,F_J20,F_M70,F_J30,F_J34

4.2   Calculate the Probability

We construct the matrix A as follows.

First we initialize a 429×429 matrix with entries equal to 0. We also contruct a map $\varphi :$ F-codes $\rightarrow \{0, 1, 2, \cdots, 428\}$ to index the diseases. Next, for history $(h_1, h_2, \cdots, h_T)$, we set

$$A_{\varphi(h_i),\varphi(h_j)} \leftarrow A_{\varphi(h_i),\varphi(h_j)} + 1, \quad 1 \le i \le j \le T.$$

Thus, we establish the matrix A.

Next, we use the power method to calculate the maximum eigenvalue λ of A and the corresponding left and right eigenvectors L and R.

After that, we derive the Markov weight matrix and the transition probability as described in Algorithm 1.

Finally, we can calculate r_j as described in subsection 3.5 and derive the related disease prediction.

4.3   Results of Accuracy

To compare the result, We use a method used previously by the insurance company as the benchmark. This method is called the emperical methods, that is, to calculate the incidence rate of diseases and use the top 5 prevalent diseases as predicition for each person.

We use 300,000 people's records to calculate p_{ij} and also the top 5 diseases. For another 10,000 people, we use their records from 2007-2014 to calculate the diseases with the highest r_j which is described in the previous section and choose the 5 diseases with highest r_j-score as prediction, which is known as the maximum entropy method.

Then we examine the diseases they suffer from during 2015-2017 to see how many diseases is accurately predicted by these two methods.

The measurement we use is called the hit rate. It is defined as follows.

$$H = \frac{|A \cap B|}{|B|},$$

where A is the disease set predicted by the model and B is the disease set that a person suffer from during 2015-2017.

If we predict 5 diseases using the maximum entropy method, the hit rate is 31.89%. As a contrast, the hit rate is 16.55% for the empirical method.

We also compare the hit rate with 1/2/3/4 predictions for the two methods. The following table summarize the result.

Table 4: Comparisons of Hit Rate

number of predictions	maximum entropy method	empirical method
1	15.01%	7.54%
2	20.50%	10.67%
3	25.21%	13.55%
4	29.01%	14.92%
5	31.89%	16.55%

We can see from the table that the hit rate of the maximum entropy method is approximately twice that of the empirical method.

4.4   Comorbidity Analysis

We first study the ALRR. Recall that M is calculated as follows.

$$M_{ij} = \left| \log \left(\frac{p_{ij}}{p_i p_j} \right) \right|.$$

If disease i and disease j are independent, then M_{ij} is close to 0. So if M_{ij} is large, then disease i and disease j are highly correlated.

If disease i is high blood pressure and disease j is type II diabetes. Then

$$M_{ij} = 2.17, \quad M_{ji} = 2.39.$$

Here is list of diseases with high LRR.

Table 5: ALRR of maximum entropy method

disease i	disease j	M_{ij}
Type II diabetes	Type I diabetes	3.40
Pulmonary heart disease	Acute ischemic heart disease	3.35
Type II diabetes	atherosclerosis	3.18
Diseases of lip, tongue and oral mucosa	Malignant tumors of the lip, mouth and pharynx	2.98
heart failure	Arrhythmia	2.96
Metabolic disorders	renal failure	2.84
emphysema	asthma	2.80
pneumonia	Bronchiectasis	2.67
high blood pressure	Type II diabetes	2.47
high blood pressure	renal failure	2.46
alopecia	Seborrheic keratosis	2.41
heart failure	Anal and rectal disorders	2.41
Type II diabetes	high blood pressure	2.39
heart failure	Peptic ulcer	2.38
high blood pressure	atherosclerosis	2.36
Alzheimer disease	Sleep disorders	2.33
high blood pressure	Cerebral hemorrhage or infarction and its sequelae	2.33
Over nutrition	Other diabetes	2.27
Pulmonary heart disease	Arrhythmia	2.22
high blood pressure	heart failure	2.21
Pituitary hyperfunction	Joint disorder	2.12
Pulmonary heart disease	arthritis	2.08

The following table is a list of diseases such that M_{ij} differs from M_{ji} .

Table 6: Asymmetric ALRR of maximum entropy method

disease i	disease j	M_{ij}	M_{ji}
Female pelvic inflammatory disease	Trichomoniasis	1.13	3.40
nephritic nephrotic syndrome	heart failure	1.51	3.35
Metabolic disorders	Malignant tumor of skin	1.37	3.19
Esophageal diseases	Splenic diseases	3.40	1.58
anemia	hypotension	2.89	1.30
Other diabetes	Central nervous system diseases	1.26	2.84
Benign tumor of uterus	Tumors with undetermined or unknown endocrine gland dynamics	2.89	1.31
Arrhythmia	Mental and behavioral disorders	2.58	1.03
Arthrosis	epilepsy	0.71	1.89
Arrhythmia	Diseases of autonomic nervous system	2.63	1.47
Headache syndrome	Other diseases of arteries and arterioles	1.58	2.74
Acute pancreatitis and other diseases of pancreas	Type II diabetes	2.78	1.70
asthma	emphysema	1.73	2.80
Ankylosis and other spondylosis	Hypopituitarism	1.87	0.80
Arthrosis	Myasthenia and primary muscle diseases	2.71	1.65
Malignant tumors of digestive organs	Hemangioma and lymphangioma	3.40	2.33
Refractive and accommodative disorders	glaucoma	2.41	1.35
Other diabetes	Optic neuropathy	1.71	2.74
Chronic ischemic heart disease	Pericardial disease	1.26	2.28
Type II diabetes	Over nutrition	0.82	1.83

Next, we consider the ϕ-correlation. Recall that

$$\phi_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i p_j (1 - p_i)(1 - p_j)}}.$$

Next table display 20 disease pairs with high ϕ-correlation.

Table 7: ϕ-correlation of maximum entropy method

disease i	disease j	ϕ
Mania, bipolar, depression, and anxiety disorders	sleep disorder	68.23
Type II diabetes	Hypertension and its complications	67.75
Headache syndrome	Pulp, gums and edentulous alveolar ridge diseases	60.72
Arrhythmia	Hypertension and its complications	58.55
Muscle disorders	Backache	57.28
Shingles	Dermatitis and pruritus	55.30
Headache syndrome	Backache	53.33
Benign uterine tumor	Abnormal uterine and vaginal bleeding	46.75
Upper respiratory tract diseases such as chronic laryngitis and laryngotracheitis	Chronic rhinitis, nasopharyngitis and pharyngitis	42.54
Other disorders of kidney and ureter	Other disorders of the urinary system	42.30
anemia	Pulp, gums and edentulous alveolar ridge diseases	40.85
cellulitis	Dermatophytes and other superficial fungal diseases	40.74
Other disorders of male reproductive organs	Prostatic hyperplasia and prostatitis	40.69
Urethral disorders	Other disorders of the urinary system	39.66
Other disorders of bone	Osteoporosis without pathological fracture	34.96
Upper respiratory tract diseases such as chronic laryngitis and laryngotracheitis	Chronic bronchitis	33.39
Other diseases of the digestive system	Gastritis and duodenitis	29.68
Type II diabetes	Metabolic disorders	27.51
Type II diabetes	Dermatitis and pruritus	27.16
Arrhythmia	sleep disorder	27.11

We can see from table 5 many disease pairs with large M_{ij}, such as type II diabetes and hypertension and its complications, which imply that such diseases have intrinsic relations. Tedesco [3] have mentioned that Hypertension is frequently associated with diabetes mellitus and its prevalence doubles in diabetics compared to the general population. This high prevalence is associated with increased stiffness of large arteries. Our result is consistent with their medical research.

5   Conclusions

In this paper, we propose a maximum entropy method for predicting disease risks. It is based on a patient's medical history with diseases coded in ICD-10 which can be used in various cases. The complete algorithm with strict mathematical derivation is given. We also present experimental results on a medical dataset, demonstrating that our method performs well in predicting future disease risks and achieves an accuracy rate twice that of the traditional method. We also perform a comorbidity analysis to reveal the intrinsic relation of diseases.

Acknowledgement

Appendix

Proof of Theorem 3.1. Suppose $W = (w_{k,m})$ is the weight matrix. Then the entropy of the Markov chain can be rewritten as

$$H(w) = -\sum_{k=1}^{N} \sum_{m=1}^{N} w_{k,m} \left(\log(w_{k,m}) - \log \left(\sum_{m=1}^{N} w_{k,m} \right) \right)$$

$$= -\sum_{k=1}^{N} \sum_{m=1}^{N} w_{k,m} \log(w_{k,m}) + \sum_{k=1}^{N} \left(\sum_{m=1}^{N} w_{k,m} \right) \log \left(\sum_{m=1}^{N} w_{k,m} \right).$$

Let us construct the Lagrangian

$$L = -\sum_{k=1}^{N} \sum_{m=1}^{N} w_{k,m} \log(w_{k,m}) + \sum_{k=1}^{N} \left(\sum_{m=1}^{N} w_{k,m} \right) \log \left(\sum_{m=1}^{N} w_{k,m} \right)$$

$$+ \sum_{k=1}^{N} h_k \left(\sum_{m=1}^{N} w_{m,k} - \sum_{m=1}^{N} w_{k,m} \right) + \mu \left(1 - \sum_{k=1}^{N} \sum_{m=1}^{N} w_{k,m} \right).$$

If $w_{k,m} \neq 0$, we have that

$$\frac{\partial L}{\partial w_{k,m}} = -1 - \log(w_{k,m}) + \log \left(\sum_{m=1}^{N} w_{k,m} \right) + 1 - h_k + h_m - \mu = 0.$$

$$\frac{w_{k,m}}{\sum_{m=1}^{N} w_{k,m}} = e^{-\mu} \frac{e^{h_m}}{e^{h_k}}.$$

If $w_{k,m} = 0$, then $B_{k,m} = 0$. Therefore,

$$\frac{w_{k,m}}{\sum_{m=1}^{N} w_{k,m}} = e^{-\mu} \frac{B_{k,m} e^{h_m}}{e^{h_k}}.$$

Set $\lambda = e^{\mu}$, then

$$\sum_{m=1}^{N} B_{k,m} e^{h_m} = \lambda e^{h_k}.$$

By the Perron-Frobenius theorem(see e.g. [11]), There are no nonnegative eigenvectors for B other than the Perron vector r and its positive multiples. Hence,

$$\frac{e^{h_m}}{e^{h_k}} = \frac{r_m}{r_k}.$$

And

$$P_{k,m} = \frac{w_{k,m}}{\sum_{m=1}^{N} w_{k,m}} = \frac{1}{\lambda} B_{k,m} \frac{r_m}{r_k}.$$

On the other hand,

$$\lambda \frac{w_{k,m}}{r_m} = B_{k,m} \frac{p_k}{r_k} \Rightarrow \lambda \frac{p_m}{r_m} = \sum_{k=1}^{N} B_{k,m} \frac{p_k}{r_k}.$$

Therefore, there exists $t > 0$ such that

$$\frac{p_k}{r_k} = t l_k.$$

$$\sum_{k=1}^{N} l_k r_k = 1 \Rightarrow t = 1.$$

Hence,

$$p_k = l_k r_k, \quad w_{k,m} = \frac{1}{\lambda} l_k B_{k,m} r_m.$$

Recall that

$$H = -\sum_{k=1}^{N} \sum_{m=1}^{N} p_k P_{k,m} \log(P_{k,m}),$$

and $0 \cdot \log 0 = \lim_{x \to 0+} x \log x = 0$. It follows that

$$H = -\sum_{k=1}^{N} \sum_{m=1}^{N} l_k r_k \frac{1}{\lambda} B_{k,m} \frac{r_m}{r_k} \log\left(\frac{1}{\lambda} B_{k,m} \frac{r_m}{r_k}\right)$$

$$= -\sum_{k=1}^{N} \sum_{m=1}^{N} l_k \frac{1}{\lambda} B_{k,m} r_m (\log B_{k,m} + \log r_m - \log r_k - \log \lambda)$$

Since $B_{k,m} \in \{0, 1\}$, $B_{k,m} \log B_{k,m} = 0$,

$$H = -\sum_{k=1}^{N} \sum_{m=1}^{N} l_k \frac{1}{\lambda} B_{k,m} r_m (\log r_m - \log r_k) + \log \lambda$$

$$= -\sum_{m=1}^{N} l_m r_m \log r_m + \sum_{k=1}^{N} l_k r_k \log r_k + \log \lambda$$

$$= \log \lambda.$$

$\square$

Next, we will prove Theorem 3.2. We first prove an auxiliary lemma.

Lemma 1. *Suppose λ is the maximum eigenvalue of B and X, Y are matrices in section 3.1. $l = (l_1, \cdots, l_N), r = (r_1, \cdots, r_N)^T$ are the corresponding left and right eigenvectors with*

$$\sum_{k=1}^{N} l_k r_k = 1.$$

Then

$$\begin{pmatrix} (lY)_1 & & \\ & \ddots & \\ & & (lY)_n \end{pmatrix} A \begin{pmatrix} (Xr)_1 & & \\ & \ddots & \\ & & (Xr)_n \end{pmatrix} = \lambda^2 X \begin{pmatrix} l_1 r_1 & & \\ & \ddots & \\ & & l_N r_N \end{pmatrix} Y. \quad (1)$$

Proof. Recall that

$$f, g : \{1, 2, \cdots, N\} \to \{1, 2, \cdots, n\},$$

and

$$X_{i,k} = \chi_{\{f(k)=i\}}, \quad Y_{k,j} = \chi_{\{g(k)=j\}}.$$

$$A_{i,j} = \sum_{k=1}^{N} X_{i,k} Y_{k,j} = \sum_{k: f(k)=i, g(k)=j} 1 = |f^{-1}(i) \cap g^{-1}(j)|.$$

Here, $|S|$ denote the number of elements in S.

$$B_{k,m} = \sum_{j=1}^{n} Y_{k,j} X_{j,m} = \sum_{j=1}^{n} \chi_{\{g(k)=j, f(m)=j\}} = \chi_{\{g(k)=f(m)\}}.$$

13

Since $Br = \lambda r$, we have that

$$\sum_{m=1}^{N} B_{k,m} r_m = \lambda r_k \Rightarrow \sum_{m=1}^{N} \chi_{\{g(k)=f(m)\}} r_m = \lambda r_k.$$

Since $lB = \lambda l$, we have that

$$\sum_{k=1}^{N} B_{k,m} l_k = \lambda l_m \Rightarrow \sum_{k=1}^{N} \chi_{\{g(k)=f(m)\}} l_k = \lambda l_m.$$

The (i,j) element of the left hand side in (1) is

$$(lY)_i A_{i,j} (Xr)_j = |f^{-1}(i) \cap g^{-1}(j)| \left(\sum_{k=1}^{N} l_k Y_{k,i} \right) \left(\sum_{m=1}^{N} r_m X_{j,m} \right)$$

$$= |f^{-1}(i) \cap g^{-1}(j)| \left(\sum_{k=1}^{N} l_k \chi_{\{g(k)=i\}} \right) \left(\sum_{m=1}^{N} r_m \chi_{\{f(m)=j\}} \right)$$

Assume that
$$f^{-1}(i) \cap g^{-1}(j) = \{u_1, u_2, \cdots, u_q\}.$$

Then
$$f(u_1) = f(u_2) = \cdots = f(u_q) = i, \quad g(u_1) = g(u_2) = \cdots = g(u_q) = j.$$

$$(lY)_i A_{i,j} (Xr)_j = q \left(\sum_{k=1}^{N} l_k \chi_{\{g(k)=i\}} \right) \left(\sum_{m=1}^{N} r_m \chi_{\{f(m)=j\}} \right)$$

$$= \sum_{t=1}^{q} \left(\sum_{k=1}^{N} l_k \chi_{\{g(k)=f(u_t)\}} \right) \left(\sum_{m=1}^{N} r_m \chi_{\{f(m)=g(u_t)\}} \right)$$

$$= \lambda^2 \sum_{t=1}^{q} l_{u_t} r_{u_t}.$$

The (i,j) element of the right hand side in (1) is

$$\lambda^2 \sum_{k=1}^{N} X_{i,k} l_k r_k Y_{k,j} = \lambda^2 \sum_{k=1}^{N} l_k r_k \chi_{\{f(k)=i, g(k)=j\}} = \lambda^2 \sum_{k \in f^{-1}(i) \cap g^{-1}(j)} l_k r_k$$

$$= \lambda^2 \sum_{t=1}^{q} l_{u_t} r_{u_t}.$$

Thus, we complete the proof. $\qquad\square$

Proof of Theorem 3.2. Suppose V is the weight matrix of A, then

$$V = X \begin{pmatrix} l_1 r_1 & & \\ & \ddots & \\ & & l_N r_N \end{pmatrix} Y.$$

By the above lemma, we have that

$$V = \frac{1}{\lambda^2} \begin{pmatrix} (lY)_1 & & \\ & \ddots & \\ & & (lY)_n \end{pmatrix} A \begin{pmatrix} (Xr)_1 & & \\ & \ddots & \\ & & (Xr)_n \end{pmatrix}$$

Set
$$L = \frac{lY}{\sqrt{\lambda}}, \quad R = \frac{Xr}{\sqrt{\lambda}}.$$

Then L, R are the left and right eigenvectors of A corresponding to the eigenvalue λ. And

$$\sum_{j=1}^{n} L_j R_j = LR = \frac{1}{\lambda} lYXr = lr = 1.$$

$$v_{i,j} = \frac{1}{\lambda} L_i A_{i,j} R_j.$$

$\square$

References

[1] Glasgow R.E. et al, Does the chronic care model serve also as a template for improving prevention, *Milbank Q*, 2001, 79(4):579-612.

[2] Cesar A. Hidalgo, Nicholas Blumm, Albert-Laszlo Barabasi, Nicholas A. Christakis, A Dynamic Network Approach for the Study of Human Phenotypes, *PLoS Computational Biology*, 2009, 5, e1000353.

[3] M. A. Tedesco, F. Natale, G. Di. Salvo, S. Caputo, M. Capasso & R. Calabro, Effects of coexisting hypertension and type II diabetes mellitus on arterial stiffness, *Journal of Human Hypertension*, 2004, 18:469-473.

[4] Roque, F. S. et al., Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Computational Biology*, 2011, 7, e1002141.

[5] Melton, G. B. et al., Interpatient distance metrics using SNOMED CT defining relationships, *J. Biomed. Inform.*, 2006, 39:697-705.

[6] Shannon, C.E., A Mathematical Theory of Communication, *Bell System Technical Journal*, 1948, 27:379-423.

[7] Maurer, A., Ockham's Razor and Chatton's Anti-Razor, *Mediaeval Studies*, 1948, 46:463-475.

[8] Perlis, R. H. et al. Using electronic medical records to enable largescale studies in psychiatry: treatment resistant depression as a model *Psychol. Med.*, 2012, 42:41-50.

[9] Thomas E. Booth, Power Iteration Method for the Several Largest Eigenvalues and Eigenfunctions, *Nuclear Science and Engineering*, 2006, 154(1):48-62.

[10] Bruce Kitchens, *Symbolic dynamics, one-sided, two-sided, and countable state Markov shifts*, 1998, Springer.

[11] Carl D. Meyer, *Matrix analysis and applied linear algebra. With solutions to problems*, 2001, SIAM.