

RESEARCH ARTICLE

# eDRAM: Effective early disease risk assessment with matrix factorization on a large-scale medical database: A case study on rheumatoid arthritis

Chu-Yu Chin<sup>1</sup>, Sun-Yuan Hsieh<sup>1</sup>, Vincent S. Tseng<sup>2\*</sup>

**1** Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, **2** Computer Science and Information Engineering, National Chiao Tung University, Hsinchu, Taiwan

\* [vtseng@cs.nctu.edu.tw](mailto:vtseng@cs.nctu.edu.tw)



**OPEN ACCESS**

**Citation:** Chin C-Y, Hsieh S-Y, Tseng VS (2018) eDRAM: Effective early disease risk assessment with matrix factorization on a large-scale medical database: A case study on rheumatoid arthritis. PLoS ONE 13(11): e0207579. <https://doi.org/10.1371/journal.pone.0207579>

**Editor:** Lars Kaderali, Universitätsmedizin Greifswald, GERMANY

**Received:** January 8, 2018

**Accepted:** November 2, 2018

**Published:** November 26, 2018

**Copyright:** © 2018 Chin et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** This study used the Longitudinal Health Insurance Database 2000 dataset (LHID2000). The ownership of the data underlying this study belongs to the National Health Insurance Research Database (NHIRD) of Taiwan and cannot be made publicly available due to legal restrictions on data confidentiality. However, the data are available for research proposed through formal application to the Health and Welfare Data Science Center at Ministry of Health and Welfare, Taiwan (<https://dep.mohw.gov.tw/DOS/np-2500-113.html>). The detailed acts and

## Abstract

Recently, a number of analytical approaches for probing medical databases have been developed to assist in disease risk assessment and to determine the association of a clinical condition with others, so that better and intelligent healthcare can be provided. The early assessment of disease risk is an emerging topic in medical informatics. If diseases are detected at an early stage, prognosis can be improved and medical resources can be used more efficiently. For example, if rheumatoid arthritis (RA) is detected at an early stage, appropriate medications can be used to prevent bone deterioration. In early disease risk assessment, finding important risk factors from large-scale medical databases and performing individual disease risk assessment have been challenging tasks. A number of recent studies have considered risk factor analysis approaches, such as association rule mining, sequential rule mining, regression, and expert advice. In this study, to improve disease risk assessment, machine learning and matrix factorization techniques were integrated to discover important and implicit risk factors. A novel framework is proposed that can effectively assess early disease risks, and RA is used as a case study. This framework comprises three main stages: *data preprocessing*, *risk factor optimization*, and *early disease risk assessment*. This is the first study integrating matrix factorization and machine learning for disease risk assessment that is applied to a nation-wide and longitudinal medical diagnostic database. In the experimental evaluations, a cohort established from a large-scale medical database was used that included 1007 RA-diagnosed patients and 921,192 control patients examined over a nine-year follow-up period (2000–2008). The evaluation results demonstrate that the proposed approach is more efficient and stable for disease risk assessment than state-of-the-art methods.

norms can be found in the following official websites: <http://law.moj.gov.tw/Eng/LawClass/LawAll.aspx?PCode=I0050021>; [http://nhird.nhri.org.tw/en/Data\\_Protection.html](http://nhird.nhri.org.tw/en/Data_Protection.html).

**Funding:** Telecommunication Laboratories, Chunghwa Telecom Co., Ltd provided support in the form of salaries for the author CY Chin, but it did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section. Ministry of Science and Technology, Taiwan grant 105-2218-E-009-031 to VST. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** There are no conflicts of interest to declare, although one of the authors is affiliated with a commercial company [Telecommunication Laboratories, Chunghwa Telecom Co., Ltd]. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

## Introduction

Rheumatoid arthritis (RA), a systemic autoimmune rheumatism disease (SARD), is rare and causes chronic bone damage and deterioration. RA is difficult to diagnose at an early stage, and with disease progression, RA may lead to bone deformation, swelling, pain, and permanent disability [1, 2]. Unfortunately, this disease is not easily cured and requires long-term follow-up, controller medications, and regular healthcare visits. Although RA does not directly cause death, it can clearly reduce the patient's ability to work or live independently, as it affects a wide range of activities, such as walking, eating, personal hygiene, and even mental health [1, 3, 4]. This significantly increases long-term domestic expenditure and affects national productivity and medical resource allocation [5, 6]. Accordingly, early detection of RA has been extensively studied [7–16] over the past few years, as it allows effective symptom management and prevents joint deterioration by appropriate medication therapy. Therefore, early diagnosis of this serious disease is fundamental in a successful treatment strategy [1, 12–14, 17]. Thus, disease prediction for RA is an important issue in medical informatics.

Recent advances in electronic medical record (EMR) standardization and medical information exchange systems have substantially enlarged EMR data sets. Efficient and effective analytical techniques are important for discovering new medical knowledge from big EMR data. The discovered rules can be used to improve disease prediction and prevention, assess patient prognosis, and increase diagnostic precision.

There are two issues in EMR analysis. First, a small number of diagnostic records are inadequate to represent the complete picture of a patient's health status. For instance, symptoms of several serious diseases, such as cancer, are obscure during early disease development stages. Therefore, the lack of patient medical records may lead to misdiagnosis, resulting in delayed medical treatment and proper care. Secondly, personal EMRs are scattered in a number of hospitals because patients are not confined to one hospital for treatment. Thus, it is difficult to combine personal EMR data for analysis, and the possibility of misdiagnosis increases.

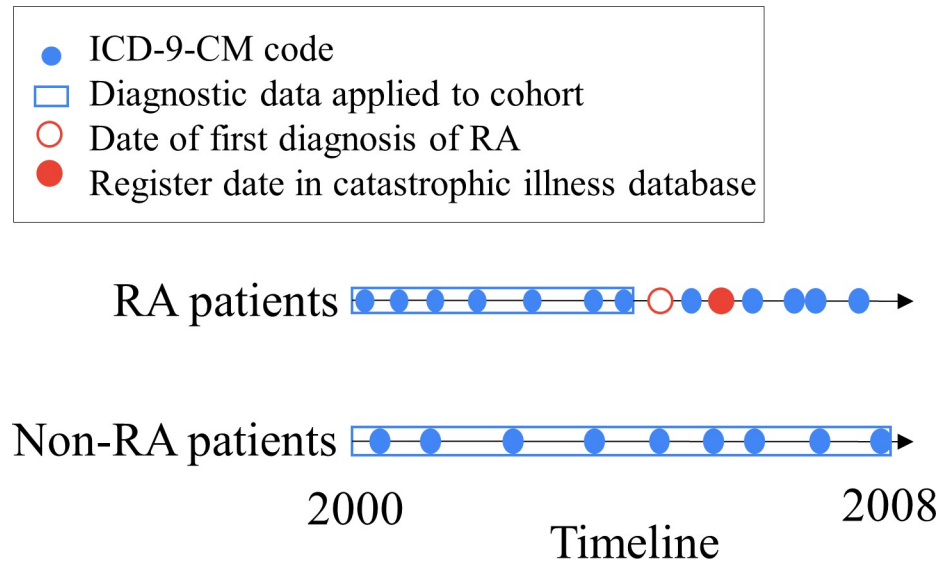
To address these issues, a universal National Health Insurance (NHI) program was conducted in Taiwan to generate a database called *NHIRD (National Health Insurance Research Database)* containing physician diagnostic records and prescriptions. Large-scale medical information is recorded by physician visits; this information is diverse and has been collected from all hospitals in Taiwan. Moreover, it is suitable for investigating personal health trends. *NHIRD* has great potential for discovering novel information, such as hidden disease risk factors, the causal relationship between diseases and symptoms, disease development, and a disease risk assessment model to promote treatment.

Although a number of past studies have considered this issue [8, 11, 18–25], it is difficult to design a disease risk assessment system that can accurately reflect the health status of a patient. This has recently attracted attention owing to the need for improving the accuracy of disease prediction, based on information in large-scale EMR databases. Thus, analytical techniques have been proposed. Liao, *et al.* (2011) developed classification algorithms by using penalized logistic regression. To validate the proposed algorithm, it was applied to two external hospitals using different electronic health record (EHR) systems [8, 11]. Carroll, *et al.* (2011) applied support vector machines (SVMs) to identify RA cases using EHR (ICD-9 codes, medications, and natural language processing-derived clinical notes) [9]. Kuo, *et al.* (2013) utilized SVMs to predict the onset of bullous pemphigoid, and the risk factors were determined by logistic regression [18]. Rau, *et al.* (2015) used artificial neural networks and logistic regression to construct a prediction model for liver cancer development in patients with type II diabetes mellitus. Furthermore, a user interface was designed to compute the probability of liver cancer occurrence using physician-proposed risk factors [19]. Chin, *et al.* (2015) proposed a

framework based on associative classification for mining risk patterns to assess the onset of early RA, and the mined classifiable patterns exhibited highly significant associations with disease risk. To estimate the novelty of risk patterns, a method for calculating the number of related studies integrated in the PubMed search engine was included in the analysis stage of the framework [26]. These associative classification techniques are based on frequent and high confidence association rules to classify objects. Classification based on multiple association rules (CMAR) [27] and classification based on associations [28] are effective associative classification techniques. Cheng, *et al.* (2017) proposed a framework integrating the “classify-by-sequence” (CBS) method [29] to mine for sequential risk patterns from time-series information in diagnostic records for early assessment of chronic diseases [30]. CBS and BayesFM [31] are sequential classification techniques that primarily combine the algorithms of sequential pattern mining, rule selection, and classification. In the above studies, the classifiable sequential patterns and classifiable patterns are considered disease risk factors for evaluating disease progression. Patient phenotyping is used to identify patients who match criteria from a large-scale population. The features of EHR are utilized to identify the cohort by machine learning and statistical methods [25]. In this framework, early disease risk assessment is aimed at discovering hidden factors and establishing assessment models based on diagnostic data that are collected before formal diagnosis of the target disease, such as RA (Fig 1). By using the model, the target disease can be assessed before its onset. This is generally called early disease risk assessment [30].

Non-negative matrix factorization (NMF) is an unsupervised analytical technique for part-based representation that achieves significant reduction in the dimensions of objective data and discovers latent factors [32–40]. It has been successfully applied to image recognition and text mining and has effectively improved accuracy and efficiency [33]. Recently, various analytical methods have been developed for different types of medical data by using the NMF algorithm. For example, Yang, *et al.* (2016) proposed unsupervised clustering to analyze gene expression data [36], Paine, *et al.* (2016) proposed unsupervised analysis using desorption electrospray ionization datasets [39], and Ozaki, *et al.* (2016) proposed analysis of complex actions in sports from electromyography data [41]. However, the above studies neglected the investigation of diagnostic data. Moreover, they have several limitations: 1) Identifying risk factors requires expert advice [42, 43]. For a large amount of medical data, the trend is to identify risk factors without human supervision. 2) A large data size or number of risk factors requires longer execution time and results in lower assessment accuracy. However, in medical decision-making, both efficiency and assessment accuracy should be considered. Thus, the ability of NMF to significantly reduce dimensionality and maintain data quality is important. 3) Recent studies have shown that SVM is useful for identifying phenotyping [9, 11, 18, 44]; however, an overly large number of EMR features may adversely affect performance and accuracy. To analyze EMR data with a large number of features and improve the assessment effectiveness, NMF was utilized to significantly reduce data dimension, discover latent factors, and improve data quality. Few studies have considered the application of NMF integrated with SVM [37] in patient phenotyping analysis. In the present study, a method integrating NMF with SVM is proposed to analyze diagnostic data for disease risk assessment.

To overcome the aforementioned limitations, an innovative approach is proposed for high precision RA prediction by using NMF. The main contributions of this study can be summarized as follows: 1) A novel framework called *eDRAM* (early disease risk assessment) is proposed for assessing disease risk in early development stages. In contrast with traditional practice, in the proposed framework, disease risk factors are approximately reconstructed by matrix factorization. 2) To the best of the authors' knowledge, this is the first study on matrix factorization techniques integrated with machine learning for disease risk assessment based on a



**Fig 1. Timeline for data collection and definition of RA patients.**

<https://doi.org/10.1371/journal.pone.0207579.g001>

nationwide medical diagnostic database. For a large number of diagnostic attributes, the proposed method can effectively approximate an optimal dimensionality. This improves both performance and data quality. 3) In the experiments, comprehensive evaluations were performed by comparing the proposed method with CBS, BayesFM, and CMAR for disease prediction. The results demonstrate that *eDRAM* is more effective than the other methods in terms of disease risk assessment metrics. To make the experiments more robust, wide-coverage data were used, and a sufficient number of evaluations were performed.

## Methods

### Overview of the proposed framework

[Fig 2](#) shows the framework of the proposed *eDRAM* approach. It comprises three main stages: data preprocessing, risk factor optimization, and early disease risk assessment. The preprocessing stage comprises noise reduction, cohort selection, and matrix transformation. To discover the optimized risk factors, the NMF algorithm with parameter optimization was used for constructing the NMF-based matrix. In the assessment model learning and early disease risk assessment stages, the machine learning classifier SVM was used for disease modeling with the NMF-based matrix, yielding the final disease risk assessment, which serves as an excellent reference for physicians and patients. The instructions for executing the experimental protocols is available at [dx.doi.org/10.17504/protocols.io.rv2d68e](https://dx.doi.org/10.17504/protocols.io.rv2d68e).

### Data preprocessing

**Noise reduction.** The EMR database contains noise, which may lead to biased disease risk assessment. Three types of data noise should be eliminated from the study cohort: 1) Incorrect data formats, such as inconsistent ICD-9-CM encoding rules or patient identification numbers with erroneous lengths. To determine this, the ICD-9-CM codes were formatted to five-digit codes. For instance, the formatting code 714.0, which represents RA, was formatted to 71400. 2) Missing, incomplete, or unreasonable data. 3) Meaningless or garbled data. For noise reduction, 795 records were removed.

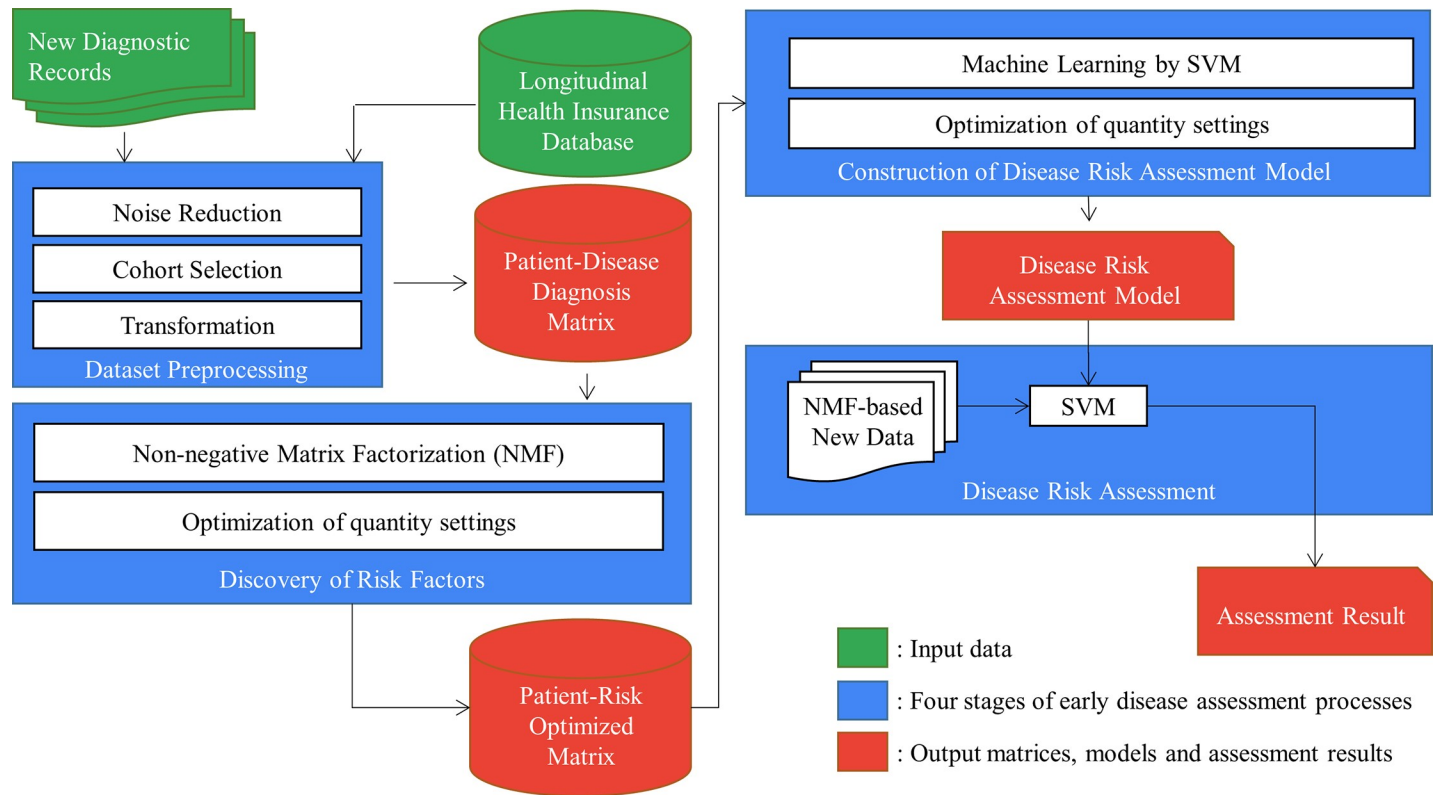


Fig 2. Framework of proposed approach.

<https://doi.org/10.1371/journal.pone.0207579.g002>

**Cohort selection.** The patient data that were collected from the EMR database satisfied selection criteria related to the following: the time-period of the study, the ICD-9-CM codes of the studied disease, and the search strategy in the two subject databases.

The original NHIRD contains information from 1996 to 2008; ICD-9-CM codes were adopted on January 1, 2000. To ensure consistent and standard codes, the cohort with outpatient diagnoses was used from 2000 to 2008. All RA cases met the criteria of the ICD-9-CM code 714.0 and were confirmed by using the registry of patients with catastrophic illnesses and the ambulatory care databases. RA cases were excluded in the controls. The cohort selection procedure flowchart is shown in Fig 3.

**Data transformation.** The outpatient clinical data analyzed here include patient ID, visiting date, and diagnostic disease codes generated at each clinic visit. An example is shown in Fig 4(A).

To analyze the relationship between the patient and the disease, the patient–disease diagnosis matrix is proposed by adopting a novel matrix-based analysis approach involving disease alignment for each patient. Given an  $N \times M$  matrix, each row represents the medical history of a diagnosed patient across all diseases or symptoms ( $DS$ ). Each column indicates the diagnostic record status of all patients for a single  $DS$ . The patient–disease diagnosis matrix is defined as follows.

**Definition 1** (*patient–disease diagnosis matrix*). Given a set of unique patients  $P = \{p_1, p_2, \dots, p_m, \dots, p_{|P|}\}$  (the total number of patients is  $|P|$ ) and a set of unique diagnostic codes  $D = \{d_1, d_2, \dots, d_n, \dots, d_{|D|}\}$  (the total number of diagnostic codes in the EMR cohort is  $|D|$ ), then the patient–disease diagnosis matrix is defined as  $PDP \rightarrow D [v_{m,n}]$ , where  $D$  is the

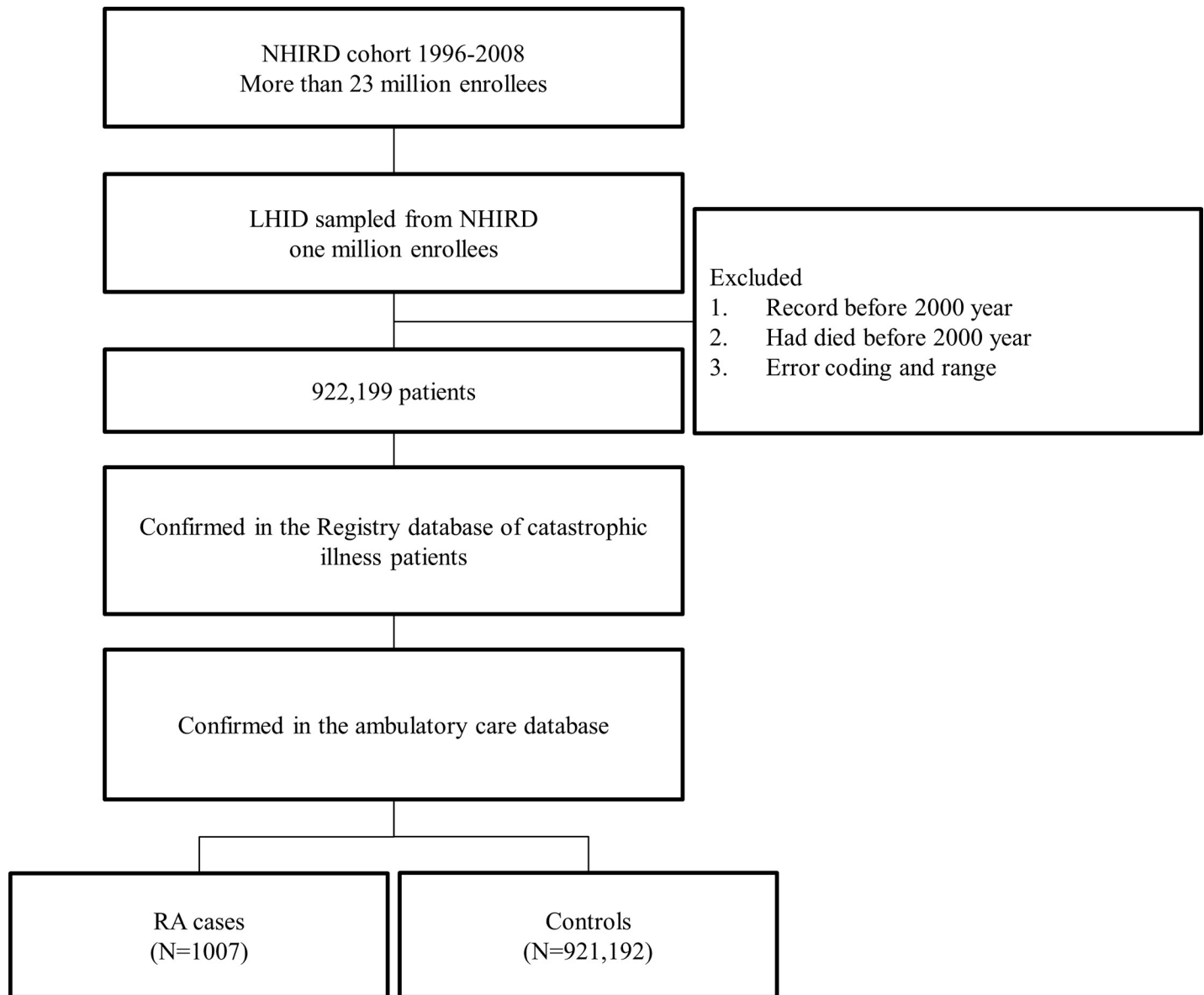


Fig 3. Flow chart of study cohort enrollment.

<https://doi.org/10.1371/journal.pone.0207579.g003>

diagnostic code set, and  $v$  is a binary value (0 or 1), representing true or false for  $0 < n \leq |D|$ . Fig 4(B) shows an example of a patient–disease diagnosis matrix.

### Discovery of latent risk factors

After the patient–disease diagnosis matrix has been generated, the next operation is to approximate a better information matrix by executing the NMF algorithm.

NMF is a multivariate analysis algorithm for matrix factor optimization, matrix decomposition, and factor reconstruction [32–39]. It should be noted that the matrix model cannot contain negative values and is suitable for the analysis of medical diagnostic data after the transformation of the patient–disease diagnosis matrix. In this operation, the aim is to reduce the matrix dimension and to discover the latent risk factors. The new risk factors are



(a) EMR data

Patient ID	Date	Diagnostic record 1	Diagnostic record 2	Diagnostic record 3
1	2000/01/01	521.00	719.40	403.00
1	2000/01/08	264.60	523.50	/
2	2000/01/06	478.21	719.49	/

(b) Patient-Disease Diagnosis matrix

\* DS = Disease or Symptom

Patient	$d_1$	$d_2$	$d_3$	$d_4$	$d_n$
$p_1$	0	1	0	0	1
$p_2$	1	1	0	1	0
$p_3$	1	0	0	0	0
$p_m$	0	0	1	1	1

(c) Dimension variations when using NMF to decompose Patient-Disease Diagnosis matrix

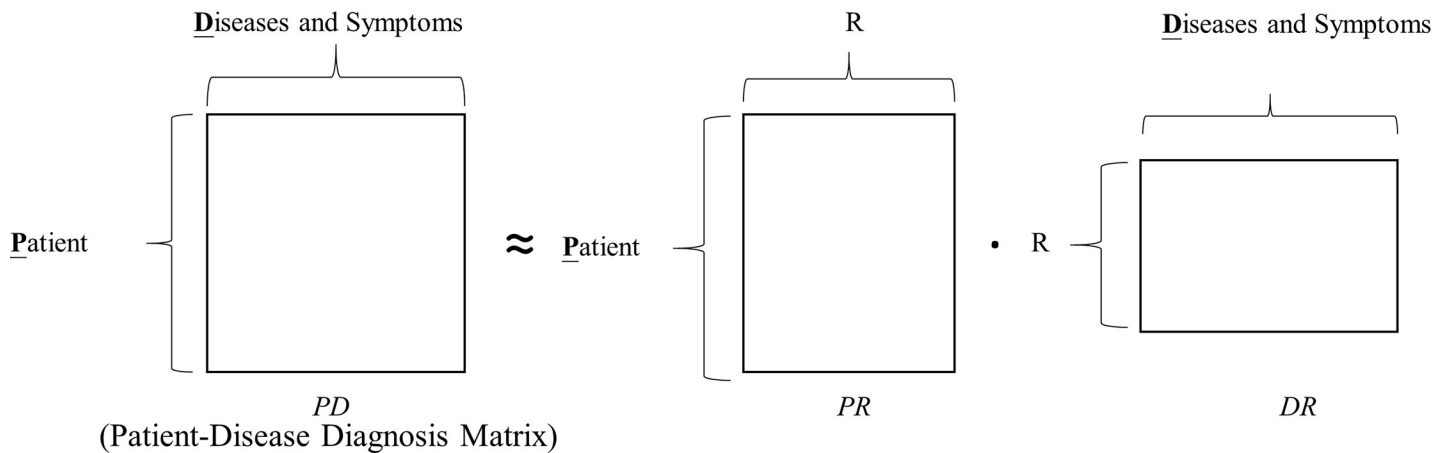


Fig 4. Example and concept of transformed patient-disease diagnosis matrix.

<https://doi.org/10.1371/journal.pone.0207579.g004>

multiplicative factors that are hidden among the original factor relationships, and their discovery allows more effective and efficient disease risk assessment. By Definition 1, the patient-disease diagnosis matrix is approximated by the two matrices in Eq (1). An example is shown in Fig 4(C).

$$PD_{P \rightarrow D}[v_{n,m}] \approx PR_{P \rightarrow R}[pv_{n,r}] \cdot DR_{D \rightarrow R}[dv_{m,r}]^T \tag{1}$$

PR and DR represent the factor matrices, each patient  $p$  and disease  $d$  are modeled by a factor vector set  $R$ ,  $0 < r \leq |R|$ , and the elements in the two matrices are nonnegative.

The cost function, which quantifies the approximation, is defined as follows:

$$\|PD - PR \cdot DR^T\|^2 = \sum_{n,m} (v - \sum_{r=1}^{|R|} pv_{n,r} - dv_{m,r})^2. \tag{2}$$

Eq (2) is minimized by multiplicative algorithms using Eq (3), which iteratively updates and improves the latent risk factors [32, 33, 35, 38, 45].

$$pv_{n,r} \leftarrow pv_{n,r} \frac{(PD \cdot DR^T)_{n,r}}{(PR \cdot DR^T \cdot DR)_{n,r}} \text{ and } v_{m,r} \leftarrow dv_{m,r} \frac{(PR^T \cdot PD)_{m,r}}{(DR \cdot PR^T \cdot PR)_{m,r}}. \tag{3}$$

According to the NMF algorithm,  $PD$  is decomposed into two risk factor matrices, namely,  $PR$  and  $DR$ .  $PR$  is called NMF-based matrix and contains the novel disease risk factors applied for disease risk assessment. For the example in Fig 4(B), the results of NMF are shown in Fig 5.

### Construction of disease risk assessment model

For constructing the disease risk model, SVM is applied to the stage for learning NMF-based matrix. This learned model can be a support to the disease risk assessment phase.

### Disease risk assessment

In this stage, the goal is to identify RA patients with disease risk from a cohort. Based on the disease risk assessment model, diagnostic records of unknown patient can be predicted using the SVM classifier [46]. Because SVM is a well-known classifier widely used by recent researches in the field of machine learning [9, 11, 18, 37, 44], it will not be described here further.

### Parameters

**Parameters of non-negative matrix factorization.** After the patient–disease diagnosis matrix transformation, NMF [33] was utilized for discovering risk factors by decomposing the patient–disease diagnosis matrix, yielding a new risk matrix of size  $N \times R$ , where  $R$  must be less than  $M$  to reduce the factor dimensions and thus achieve data compression. As adjustments to  $R$  can affect the effectiveness of disease risk assessment, an optimal  $R$  value must be experimentally determined for each individual database.

**SVM parameters with RBF kernel function.** The SVM parameters are  $C$  and  $\gamma$ .  $C$  indicates the extent to which misclassification should be avoided, and thus higher values represent higher sensitivity.  $\gamma$  defines how far the influence of a training example reaches. Larger  $\gamma$  values form a small support vector, resulting in overfitting. The best combination of the two parameters can be obtained using the grid search method [46]. In this experiment,  $C$  was set to 2 and  $\gamma$  was set to 0.03125.


### Materials

In this study, a large-scale nationwide medical outpatient dataset, namely, Longitudinal Health Insurance Database 2000 (LHID2000) sampled from Taiwan’s NHIRD was used. NHIRD covers more than 99.6% of the general population of Taiwan, with approximately 23 million people [8], and is thus highly representative. The data is from the period 1996–2008.

To ensure that in the proposed approach, the specified disease model is appropriately learned, the database was divided into two datasets: RA cases and controls. The definitions of the two datasets are as follows: 1) RA cases included patients diagnosed more than twice with the RA diagnostic code and who were simultaneously enrolled in the registry database of patients with catastrophic illnesses. The RA patient data were collected from 2000 until the



Patient No.	DS 1	DS 2	DS 3	DS 4	DS m
1	0	1	0	0	1
2	1	1	0	1	0
3	1	0	0	0	0
n	0	0	1	1	1

 **Transforming**

Patient No.	New Risk Factor 1	New Risk Factor 2	New Risk Factor 3	r
1	0	0.99477	0.82013	0.0075758
2	1.0374	0.73335	0	0.96785
3	0.97234	0.98089	0	0
n	0.68671	0	1.0202	0.12876

\* DS = Disease or Symptom

**Fig 5. Using NMF to decompose the NMF-based matrix from the patient–disease diagnosis matrix.**

<https://doi.org/10.1371/journal.pone.0207579.g005>

patients were diagnosed with RA for the first time (Fig 1). The hypothesis is that the disease patterns/models were hidden in the diagnostic records as early features/symptoms/relationships before the formal diagnosis of RA. The data after the patients were diagnosed with RA were not included in this dataset. The proposed method assesses whether patients would develop RA based on diagnostic data that were recorded before RA had been formally diagnosed. 2) Patients who did not meet the criteria that define RA and had medical diagnostic records from 2000 to 2008 were classified as controls.

In the cohort, there were a large number of outpatient diagnostic records of approximately 163 million individuals, containing 13,392 ICD-9-CM codes that also represented a number of diseases/symptoms. Each code represented a specific disease or symptom. With regard to gender, a statistically significant difference was observed, namely, the proportion of women in the dataset consisting of RA cases and controls was 76.3% and 48.9%, respectively, suggesting that the dataset consisting of RA cases had a larger number of women. The means of diagnostic records, diagnostic codes, and clinical meetings per year exhibited statistically significant differences in the comparison, indicating that RA cases involved more frequent meetings with physicians as well as more types of diseases. The frequency and distribution of the continuous variables for all patients were compared using Student’s *t*-test and Pearson’s chi-squared test. The prevalence of RA in the cohort is 0.1%, which is approximately equal to that reported in a previous study in Taiwan (97.5 cases in a population of 100,000) [47]. More details are shown in Table 1.

### Experiments

In this section, the details of the experiments are presented, namely, experimental dataset, experimental environment, experimental measures, experimental settings for parameter R, effectiveness evaluation, efficiency evaluation, and discussion.

**Table 1. Baseline characteristics of RA patients in the cohort (2000–2008).**

	RA cases (N = 1007)	Controls (N = 921,192)	p-value
Mean age (SD), years	57.76 ± 15.41	41.78 ± 20.32	< 0.0001
Female, %	76.3	48.9	< 0.0001
Clinic visits			
No. of all visits	88,713	109,777,857	
Mean no. per year	21.59 ± 15.72	14.10 ± 12.6	< 0.0001
Median no. per year	18	10.57	
Diagnostic records			
No. of all diagnostic records	141,394	163,043,706	
Mean no. per year	34.51 ± 29.84	21.12 ± 22.80	< 0.0001
Median no. per year	26.31	13.87	
Diagnostic codes (ICD-9-CM)			
No. of diagnostic codes	2328	13,392	
Mean no. per year	7.98 ± 4.93	4.19 ± 2.56	< 0.0001
Median no. per year	6.96	3.7	

<https://doi.org/10.1371/journal.pone.0207579.t001>

### Experimental dataset

The experimental data were randomly sampled from the cohort (Table 1). They contained three sets, namely, Datasets 1, 2, and 3. Dataset 1 was used for the experimental setting of parameter R. Dataset 2 was used to evaluate performance. Dataset 3 was used to evaluate efficiency. As shown in Table 2, Dataset 1 consisted of 500 RA patients and 500 non-RA patients, Dataset 2 consisted of 500 RA patients and 25000 non-RA patients, and Dataset 3 consisted of 25000 RA patients and 25000 non-RA patients. The patients in Dataset 1 were different from those in Dataset 2. In Dataset 2, the non-RA patients were divided into 50 groups of controls, with 500 patients in each group. Each group had the same RA patients and different non-RA patients. Thus, Dataset 2 was divided into 50 new datasets. In Dataset 3, the number of RA patients was replicated from 1000 to 25000 and 25000 non-RA patients.

The dimension of each dataset was reduced by performing NMF separately. In the validation step, the stratified 10-fold cross-validation strategy [46, 48] was performed according to the proportion of the categories (RA and non-RA), with each fold having an equal proportion of RA patients and non-RA patients. Each iteration comprised nine folds as training data for construction of the disease risk model and one fold as testing data for performance evaluation. Ten iterations were performed in sequence. The results were averaged calculated (Fig 6).

### Experimental environment

The experiments were implemented on a server with two Intel CPU E5-2630 v4 2.2GHz and 32GB RAM, running Windows 7 Enterprise. All classification algorithms were implemented in Java. The NMF library of Matlab 2016a and libSVM [46] were used in the study.

### Experimental measures

To evaluate the proposed method, the following metrics were employed: accuracy, sensitivity, specificity, and standard deviation. They are described as follows, and the corresponding

**Table 2. Description of experimental data.**

Description	Dataset 1	Dataset 2	Dataset 3
RA Patient No.	500	500	25000
non-RA Patient No.	500	25000	25000

<https://doi.org/10.1371/journal.pone.0207579.t002>

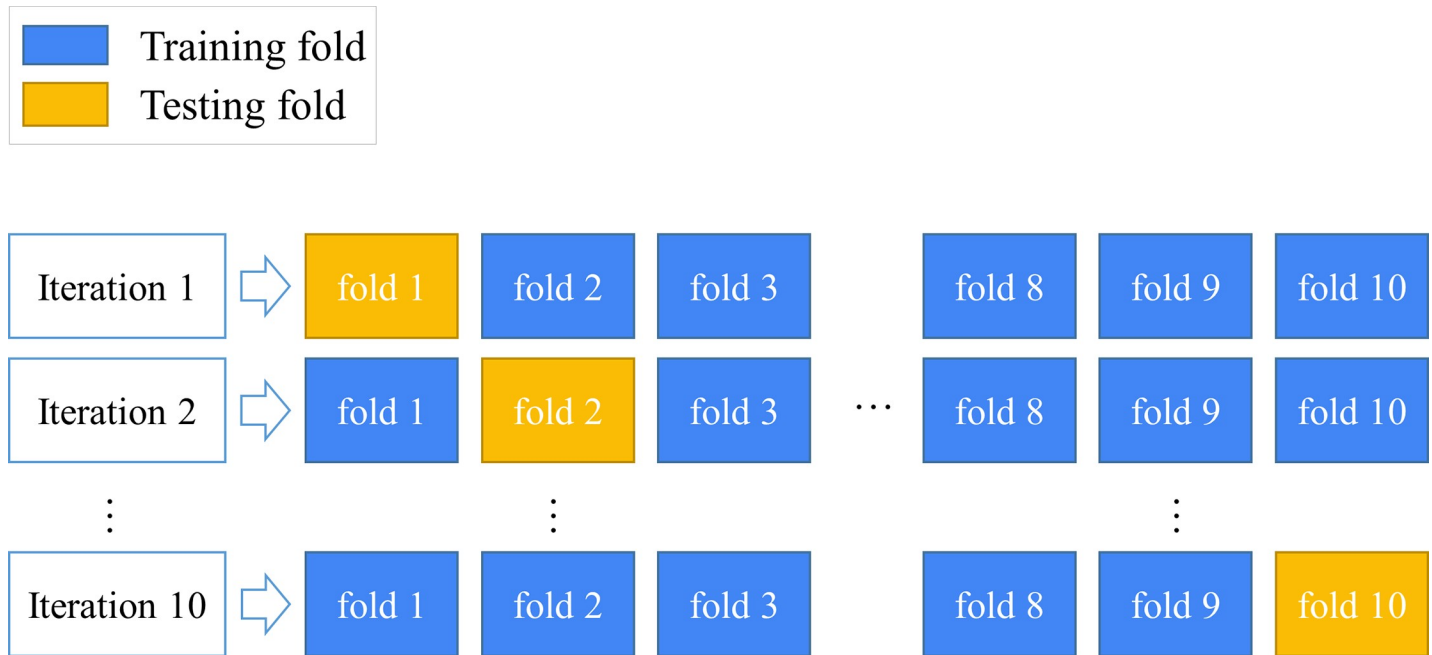


Fig 6. 10-fold cross-validation model.

<https://doi.org/10.1371/journal.pone.0207579.g006>

formulas are given in Eqs 4–7. *True positive* is the number of RA cases correctly assessed. *True negative* is the number of control cases correctly assessed. *Condition positive* is the number of all RA cases. *Condition negative* is the number of all control cases. *Sensitivity* (equivalent to recall) represents the ability to correctly assess patients with RA. *Specificity* indicates the ability to correctly assess patients without RA. *Accuracy* indicates the ability to correctly assess the cases of RA and controls. In the experiment setting phase, the purpose of adjusting the parameters is to balance sensitivity and specificity with the highest performance. Accuracy can be regarded as the average of sensitivity and specificity, as the proportion of patients has been adjusted. Standard deviation (SD) is used to measure the amount of variation of a set of measurements (sensitivity, specificity, and accuracy). This represents a measure of stability of a classifier. If  $\{x_1, x_2, \dots, x_n\}$  are the observed values,  $\mu$  is their mean value, and  $n$  is their number, then

$$Accuracy = \frac{\sum True\ positive + \sum True\ negative}{\sum Condition\ positive + \sum Condition\ negative}, \tag{4}$$

$$Sensitivity = \frac{\sum True\ positive}{\sum Condition\ positive}, \tag{5}$$

$$Specificity = \frac{\sum True\ negative}{\sum Condition\ negative}, \tag{6}$$

$$Standard\ Deviation = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}. \tag{7}$$

These metrics were used to compare the proposed *eDRAM* method with three representative approaches, namely, *CBS* [29], *CMAR* [27], and *BayesFM* [31]. To obtain a highly effective

disease risk assessment, the parameters in this experiment should be adjusted using all approaches. The details are described below.

## Experimental results

In the experiments, the evaluations comprised: 1) Selection of the  $R$  value based on effectiveness. 2) Effectiveness comparisons between the proposed method and existing well-known disease risk assessment systems in terms of sensitivity, specificity, accuracy, and standard deviation of specificity. 3) Efficiency evaluation of all methods.

**Experimental settings for parameter  $R$ .** For NMF, the diagnostic data should be transformed to the patient-disease diagnosis matrix. The number of diseases/symptoms used in the patient-disease diagnosis matrix was 13392. To optimize the NMF-based risk factors, the  $R$  value should be determined. It represents the number of risk factors refined from the  $M$  columns of the original disease diagnostics matrix, where  $R < M$ . Particularly, overly large or small values of  $R$  would decrease the effectiveness of disease risk assessment. To determine the optimal  $R$  value, experiments were conducted by using Dataset 1 and varying  $R$  from 100 to 900 with an interval of 100.

[Fig 7](#) shows the effectiveness of RA assessment for various risk factor numbers. The following should be noted. First, the trend of the curve shows that high  $R$  values result in relatively low accuracy. When the number of risk factors reaches 800, an unstable assessment is obtained, that is, the difference in sensitivity and specificity is greater than 10%. This implies that excessive risk factors will reduce accuracy. Secondly, as the  $R$  value decreases, the measured value tends to stabilize and converge. However, lower  $R$  values will reduce the overall measurement values resulting in observations varying from 200 to 100. Thirdly, the measure values has a cyclic relationship with the trend of the  $R$  value. For example, in [Fig 7](#), when the  $R$  value is in the range 400–700, the continuous value of the sensitivity forms a peak. Fourthly, accuracy can be considered a combination of sensitivity and specificity. Fifthly, when  $R$  is 200, the accuracy and its standard deviation achieved the best result. Based on the highest accuracy (the lowest standard deviation and the acceptable distance between sensitivity and specificity was smaller than 5%),  $R$  was set to 200 for the following experiments.

**Effectiveness evaluation.** In this experiment, the performance of the proposed framework *eDRAM* was evaluated against CBS [29], CMAR [27], and BayesFM [31] on Dataset 2. The  $\min\_sup$  values of CBS, CMAR, and BayesFM were set to 0.063, 0.005, and 0.006, respectively. The comparison shows that *eDRAM* achieved a better assessment rate than CBS, CMAR, and BayesFM in terms accuracy, sensitivity, and specificity on the cohort ([Fig 8](#)). Moreover, *eDRAM* maintained sensitivity and specificity closer to each other and better balanced than the other approaches ([Fig 8](#)); thus, *eDRAM* is more practical. Furthermore, *eDRAM* proved highly efficient, as it used fewer risk factors and still achieved better efficacy. Indeed, the number of risk factors in *eDRAM* was reduced to 200 ([Fig 7](#)), whereas the other approaches had 13,392 risk factors ([Table 1](#)).

The stability of the proposed disease risk assessment approaches is now evaluated. The experiment was based on the results of the previous sub-section for the standard deviation of sensitivity, specificity, and accuracy. The standard deviation was used (Eq 10) to evaluate the performance stability of the approaches. When the standard deviation is smaller, stability is higher. [Fig 8](#) shows the results of comparing *eDRAM*, CBS, CMAR, and BayesFM in terms of the standard deviation of sensitivity, specificity, and accuracy. It can be seen that *eDRAM* is the most stable method because it converts visible variables to latent factors [32, 33] that are non-redundant and more concise so as to achieve high-quality disease risk assessment.

**Efficiency evaluation.** In this experiment, the average runtime in the assessment phase was calculated by using Dataset 3. The experimental results in [Table 3](#) show that the proposed

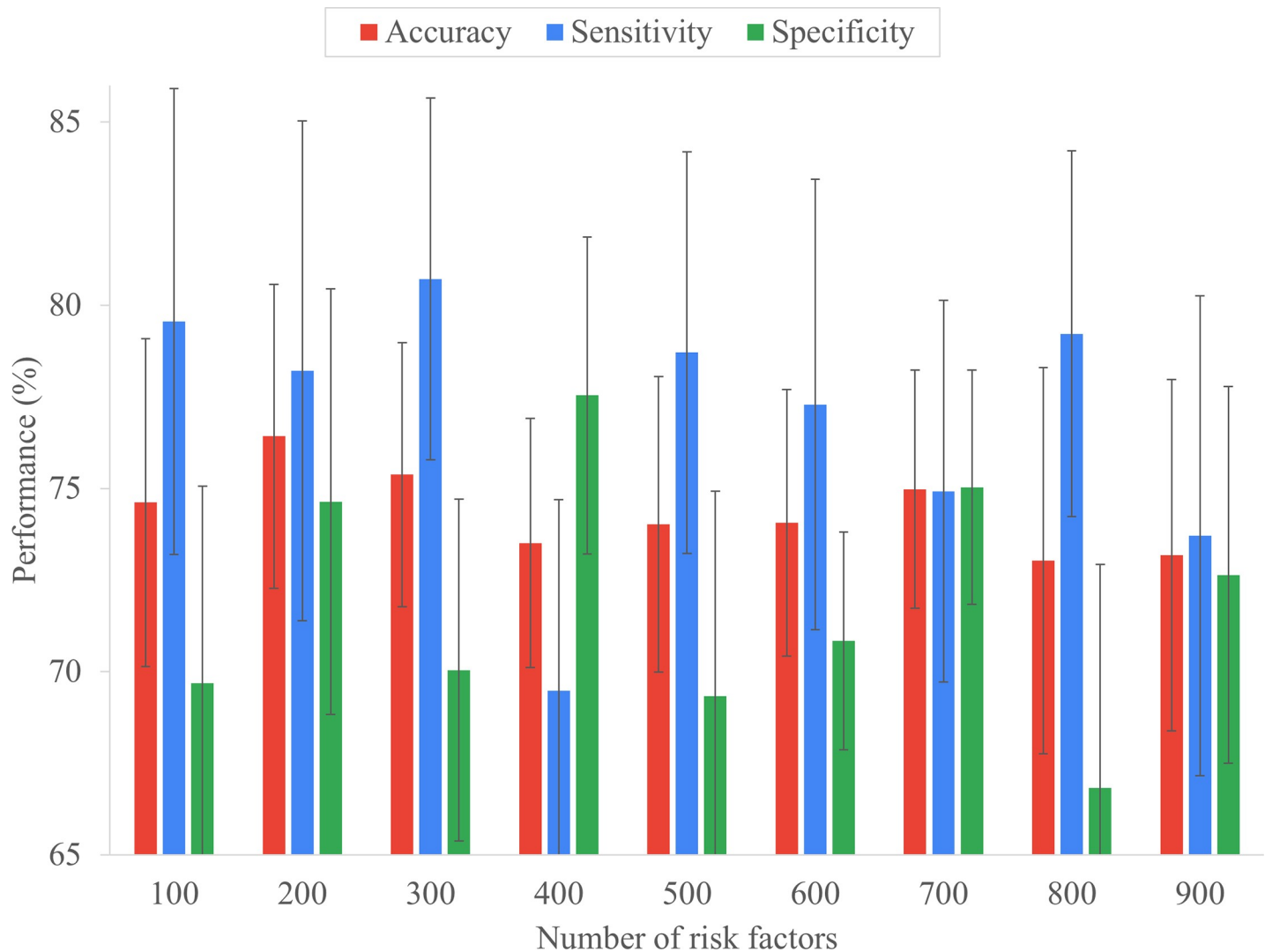


Fig 7. Effectiveness of RA risk assessment under different number of risk factors.

<https://doi.org/10.1371/journal.pone.0207579.g007>

method had the best performance in terms of assessment time. In the assessment phase, the proposed method was 2.5 times as efficient as CMAR and five times as efficient as CBS. Owing to dimensionality reduction, the proposed method can reduce loading and execution time. Regarding the other methods, the diagnostic datasets have several features that result in an increase in the number of disease patterns and hence an increase in the search time during the disease risk assessment phase.

### Discussion

Based on the performance and the stability measures obtained from the experimental evaluation, the following can be concluded:

1. The experimental evaluation demonstrates that *eDRAM* is superior in terms of accuracy, sensitivity, and specificity (Fig 8), as it establishes a matrix-based diagnostic data analysis model to decompose the NMF-based matrix for identifying important disease risk factors.

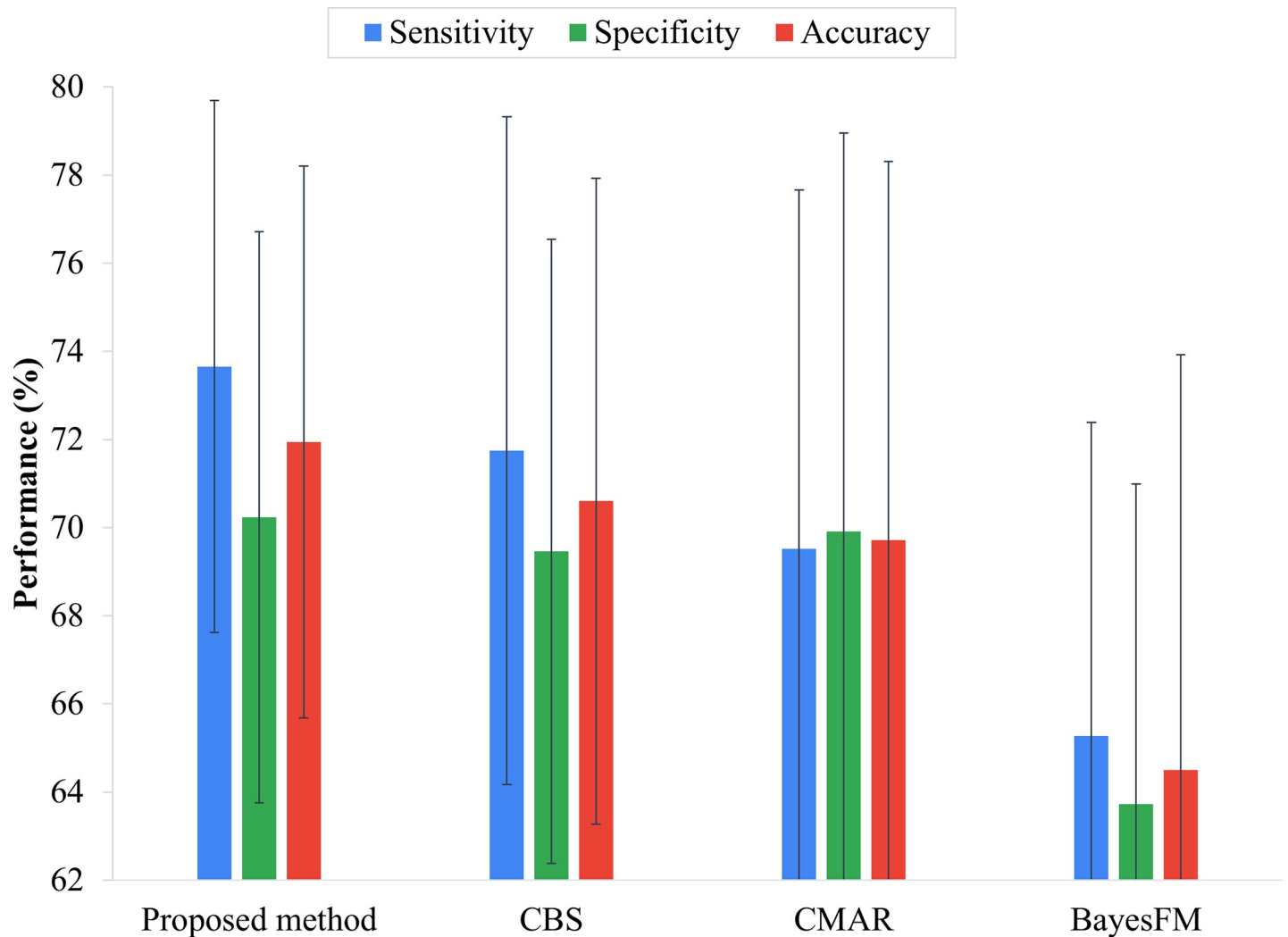


Fig 8. Comparison of the performance of the *proposed method*, CBS, CMAR, and BayesFM approaches.

<https://doi.org/10.1371/journal.pone.0207579.g008>

This indicates that the proposed approach has the advantage of finding more associated factors hidden in the diagnostic data than the other approaches.

2. Fig 7 shows the trade-off between sensitivity and specificity, that is, it is not easy to perform well in terms of both sensitivity and specificity. For example, Fig 8 shows CMAR has better specificity than CBS but is poor in terms of sensitivity. By contrast, with *eDRAM*, both sensitivity and specificity are robust for early disease risk assessment.
3. In disease risk assessment, stability and efficacy are equally important. Therefore, an experiment involving standard deviation was conducted to evaluate stability. It was demonstrated that high performance can be achieved, but its stability is not necessarily optimal. For example, Fig 8 shows that CBS ranks second in performance, but third in stability. The experimental evaluation shows that *eDRAM* is reliable in terms of both stability and performance for early disease risk assessment.
4. In the experiment for selecting  $R$ , the dimension was reduced from the original 13392 to 200, the ratio being approximately 66. Even though the amount of data was significantly



Table 3. Efficiency comparison on Dataset 3.

	Proposed method	CMAR	CBS	BayesFM
Assessment time (sec.)	0.12	0.30	0.61	41.90

<https://doi.org/10.1371/journal.pone.0207579.t003>

reduced, the experimental results on performance demonstrated that the proposed method achieved better results compared to the other methods.

- In the experiment, BayesFM discovered an excessively large number of sequential rules (features). Although the pruning algorithm was employed, there was still a large number of rules employed for assessment (> 25000), which is 50 times more than in CBS. This indicates that an excessively large number of features will result in low efficiency (Table 3), lower effectiveness (Fig 8), and render the assessment results unstable (higher standard deviations, as shown in Fig 8).
- The patient-disease diagnosis matrix is transformed into a NMF-based matrix with significantly reduced dimensions, instead of selecting specific factors. The NMF-based matrix is still associated with the original matrix and can approximate it. For machine learning, an overly large number of attributes and less data correlation may lead to misjudgment and reduce efficiency. Thus, NMF is suitable for extensive analysis of big data [34, 49, 50].

## Conclusions

Several serious diseases are not apparent during the early stages of their development. Hence, they are difficult to diagnose. This delayed detection results in missing the optimal time for treatment initiation that may be critical for controlling the disease. To address this, a novel method called *eDRAM* was proposed for early disease risk assessment with high efficacy, efficiency, and stability. *eDRAM* discovered novel risk factors from a large-scale nationwide outpatient diagnostic database using matrix factorization. Based on the optimal risk factors discovered, a disease risk assessment model was established using machine learning techniques. Thereupon, the model successfully assessed the disease risk. In summary, the contributions of this study are as follows. First, to the best of the authors' knowledge, this is the first study to apply the NMF algorithm. The main advantages of the proposed method using NMF lie in that the optimal factors hidden in the data can be approximated to achieve high assessment accuracy, and the traditional problems of big data can be resolved by significant dimensionality reduction. Secondly, a diagnostic data model called patient-disease diagnosis matrix was proposed for mapping the medical diagnostic dataset. It facilitates further data analysis and factor discovery by using matrix factorization and classification techniques, as was demonstrated in this study. Moreover, it provides a new perspective for the problem of disease risk assessment. Thirdly, modeling of disease risk assessment based on the longitudinal nationwide EMR is effective, reliable, and robust. The experimental results demonstrated that the proposed approach is superior to the three modern classification approaches used for disease risk assessment.

For future work, several research directions could be further explored. First, medications play an important role in disease treatment. Hence, the prescription database and associations between prescriptions and diseases can be considered important risk factors. Secondly, environmental conditions associated with diseases, such as place of residence, season, and occupation, are potential risk factors that should be taken into consideration to enhance the effectiveness of disease risk assessment. Thirdly, as temporal information is an important potential factor, a temporal information model could be advantageously incorporated. Finally,

using the proposed matrix-based analytic approach in combination with novel effective classifiers can aid in the discovery of deeper risk factors and the early detection of several serious diseases.

## Author Contributions

**Conceptualization:** Chu-Yu Chin, Sun-Yuan Hsieh.

**Data curation:** Chu-Yu Chin.

**Formal analysis:** Chu-Yu Chin, Sun-Yuan Hsieh.

**Funding acquisition:** Vincent S. Tseng.

**Investigation:** Chu-Yu Chin.

**Methodology:** Chu-Yu Chin, Vincent S. Tseng.

**Project administration:** Vincent S. Tseng.

**Resources:** Sun-Yuan Hsieh, Vincent S. Tseng.

**Software:** Chu-Yu Chin.

**Supervision:** Sun-Yuan Hsieh, Vincent S. Tseng.

**Validation:** Chu-Yu Chin, Vincent S. Tseng.

**Visualization:** Chu-Yu Chin.

**Writing – original draft:** Chu-Yu Chin.

**Writing – review & editing:** Chu-Yu Chin, Vincent S. Tseng.

## References

1. Smolen JS, Landewé R, Bijlsma J, Burmester G, Chatzidionysiou K, Dougados M, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. *Annals of the Rheumatic Diseases*. 2017. <https://doi.org/10.1136/annrheumdis-2016-210715> PMID: 28264816
2. Bergström M, Ahlstrand I, Thyberg I, Falkmer T, Börsbo B, Björk M. 'Like the worst toothache you've had'—How people with rheumatoid arthritis describe and manage pain. *Scandinavian Journal of Occupational Therapy*. 2017; 24(6):468–76. <https://doi.org/10.1080/11038128.2016.1272632> PMID: 28052711
3. Foti Daniela P, Greco M, Paella E, Gulletta E. New laboratory markers for the management of rheumatoid arthritis patients. *Clinical Chemistry and Laboratory Medicine (CCLM)*2014. p. 1729.
4. D. DK, I. ODC, Wolfgang H, S. MD, A. LA, A. DL, et al. The number of elevated cytokines and chemokines in preclinical seropositive rheumatoid arthritis predicts time to diagnosis in an age-dependent manner. *Arthritis & Rheumatism*. 2010; 62(11):3161–72. <https://doi.org/10.1002/art.27638> PMID: 20597112
5. Wang BCM, Hsu P-N, Furnback W, Ney J, Yang Y-W, Fang C-H, et al. Estimating the Economic Burden of Rheumatoid Arthritis in Taiwan Using the National Health Insurance Database. *Drugs—Real World Outcomes*. 2016; 3(1):107–14. <https://doi.org/10.1007/s40801-016-0063-8> PMC4819475. PMID: 27747810
6. Mora C, Díaz J, Quintana G. Costos directos de la artritis reumatoide temprana en el primer año de atención: simulación de tres situaciones clínicas en un hospital universitario de tercer nivel en Colombia. *Biomédica*. 2009; 29:43–50.
7. Jansen H, Willenborg C, Lieb W, Zeng L, Ferrario PG, Loley C, et al. Rheumatoid Arthritis and Coronary Artery Disease: Genetic Analyses Do Not Support a Causal Relation. *The Journal of Rheumatology*. 2017; 44(1):4–10. <https://doi.org/10.3899/jrheum.151444> PMID: 27744395
8. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-Treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research*. 2010; 62(8):1120–7. <https://doi.org/10.1002/acr.20184> PMC3121049. PMID: 20235204

9. Carroll RJ, Eyer AE, Denny JC. Naïve Electronic Health Record Phenotype Identification for Rheumatoid Arthritis. *AMIA Annual Symposium Proceedings*. 2011; 2011:189–96. PMC3243261. PMID: [22195070](https://pubmed.ncbi.nlm.nih.gov/22195070/)
10. Scott DL. Early rheumatoid arthritis. *British Medical Bulletin*. 2007; 81-82(1):97–114. <https://doi.org/10.1093/bmb/ldm011> PMID: [17540693](https://pubmed.ncbi.nlm.nih.gov/17540693/)
11. Carroll RJ, Thompson WK, Eyer AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*. 2012; 19(e1):e162–e9. <https://doi.org/10.1136/amiajnl-2011-000583> PMID: [22374935](https://pubmed.ncbi.nlm.nih.gov/22374935/)
12. Cader MZ, Filer A, Hazlehurst J, de Pablo P, Buckley CD, Raza K. Performance of the 2010 ACR/EULAR criteria for rheumatoid arthritis: comparison with 1987 ACR criteria in a very early synovitis cohort. *Annals of the Rheumatic Diseases*. 2011; 70(6):949–55. <https://doi.org/10.1136/ard.2010.143560> PMID: [21285117](https://pubmed.ncbi.nlm.nih.gov/21285117/)
13. Gibofsky A. Overview of epidemiology, pathophysiology, and diagnosis of rheumatoid arthritis. *The American Journal of Managed Care*. 2012; 18(13 Suppl):S295–302. PMID: [23327517](https://pubmed.ncbi.nlm.nih.gov/23327517/)
14. Goekoop-Ruiterman YPM, De Vries-Bouwstra JK, Allaart CF, Van Zeben D, Kerstens PJSM, Hazes JMW, et al. Clinical and radiographic outcomes of four different treatment strategies in patients with early rheumatoid arthritis (the BeSt study): A randomized, controlled trial. *Arthritis & Rheumatism*. 2005; 52(11):3381–90. <https://doi.org/10.1002/art.21405> PMID: [16258899](https://pubmed.ncbi.nlm.nih.gov/16258899/)
15. SJ A., HA R., Siamak N. Accuracy of veterans administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Care & Research*. 2004; 51(6):952–7. <https://doi.org/10.1002/art.20827> PMID: [15593102](https://pubmed.ncbi.nlm.nih.gov/15593102/)
16. Turk SA, van Schaardenburg D, Boers M, de Boer S, Fokker C, Lems WF, et al. An unfavorable body composition is common in early arthritis patients: A case control study. *PLOS ONE*. 2018; 13(3): e0193377. <https://doi.org/10.1371/journal.pone.0193377> PMID: [29565986](https://pubmed.ncbi.nlm.nih.gov/29565986/)
17. Schneider M, Krüger K. Rheumatoid Arthritis—Early Diagnosis and Disease Management. *Deutsches Ärzteblatt International*. 2013; 110(27–28):477–84. <https://doi.org/10.3238/arztebl.2013.0477> PMC3722643. PMID: [23964304](https://pubmed.ncbi.nlm.nih.gov/23964304/)
18. Kuo CC, Yang FC, Yang MH, Lee DD, editors. Predicting the onset of bullous pemphigoid with co-morbidities: A survey based on a nationwide medical database. 2013 IEEE International Conference on Bioinformatics and Biomedicine; 2013 18–21 Dec. 2013.
19. Rau H-H, Hsu C-Y, Lin Y-A, Atique S, Fuad A, Wei L-M, et al. Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. *Computer Methods and Programs in Biomedicine*. 2016; 125:58–65. <https://doi.org/10.1016/j.cmpb.2015.11.009>. PMID: [26701199](https://pubmed.ncbi.nlm.nih.gov/26701199/)
20. Lam C, Kuan C-F, Miser J, Hsieh K-Y, Fang Y-A, Li Y-C, et al. Emergency department utilization can indicate early diagnosis of digestive tract cancers: A population-based study in Taiwan. *Computer Methods and Programs in Biomedicine*. 2014; 115(3):103–9. <https://doi.org/10.1016/j.cmpb.2014.04.002>. PMID: [24835615](https://pubmed.ncbi.nlm.nih.gov/24835615/)
21. Liao J-N, Chao T-F, Liu C-J, Wang K-L, Chen S-J, Tuan T-C, et al. Risk and prediction of dementia in patients with atrial fibrillation — A nationwide population-based cohort study. *International Journal of Cardiology*. 2015; 199:25–30. <https://doi.org/10.1016/j.ijcard.2015.06.170> PMID: [26173170](https://pubmed.ncbi.nlm.nih.gov/26173170/)
22. Chao T-F, Liu C-J, Tuan T-C, Chen S-J, Chen T-J, Lip GYH, et al. Risk and Prediction of Sudden Cardiac Death and Ventricular Arrhythmias for Patients with Atrial Fibrillation—A Nationwide Cohort Study. *Scientific Reports*. 2017; 7:46445. <https://doi.org/10.1038/srep46445> <https://www.nature.com/articles/srep46445#supplementary-information>. PMID: [28422144](https://pubmed.ncbi.nlm.nih.gov/28422144/)
23. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*. 2016; 6:26094. <https://doi.org/10.1038/srep26094> PMC4869115. PMID: [27185194](https://pubmed.ncbi.nlm.nih.gov/27185194/)
24. Yang H, Chen Y-H, Hsieh T-F, Chuang S-Y, Wu M-J. Prediction of Mortality in Incident Hemodialysis Patients: A Validation and Comparison of CHADS2, CHA2DS2, and CCI Scores. *PLOS ONE*. 2016; 11(5):e0154627. <https://doi.org/10.1371/journal.pone.0154627> PMID: [27148867](https://pubmed.ncbi.nlm.nih.gov/27148867/)
25. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association: JAMIA*. 2014; 21(2):221–30. <https://doi.org/10.1136/amiajnl-2013-001935> PMC3932460. PMID: [24201027](https://pubmed.ncbi.nlm.nih.gov/24201027/)
26. Chin CY, Weng MY, Lin TC, Cheng SY, Yang YHK, Tseng VS. Mining Disease Risk Patterns from Nationwide Clinical Databases for the Assessment of Early Rheumatoid Arthritis Risk. *PLOS ONE*. 2015; 10(4):e0122508. <https://doi.org/10.1371/journal.pone.0122508> PMID: [25875441](https://pubmed.ncbi.nlm.nih.gov/25875441/)
27. Wenmin L, Jiawei H, Jian P, editors. CMAR: accurate and efficient classification based on multiple class-association rules. *Proceedings 2001 IEEE International Conference on Data Mining*; 2001 2001.

28. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*; New York, NY. 3000305: AAAI Press; 1998. p. 80–6.
29. Tseng VS, Lee C-H. Effective temporal data classification by integrating sequential pattern mining and probabilistic induction. *Expert Systems with Applications*. 2009; 36(5):9524–32. <http://dx.doi.org/10.1016/j.eswa.2008.10.077>.
30. Cheng YT, Lin YF, Chiang KH, Tseng VS. Mining Sequential Risk Patterns From Large-Scale Clinical Databases for Early Assessment of Chronic Diseases: A Case Study on Chronic Obstructive Pulmonary Disease. *IEEE Journal of Biomedical and Health Informatics*. 2017; 21(2):303–11. <https://doi.org/10.1109/JBHI.2017.2657802> PMID: 28129195
31. Lesh N, Zaki MJ, Oglhara M. Scalable feature mining for sequential data. *IEEE Intelligent Systems and their Applications*. 2000; 15(2):48–56. <https://doi.org/10.1109/5254.850827>
32. Lee DD, Seung HS, editors. *Algorithms for non-negative matrix factorization*. *Advances in neural information processing systems*; 2001.
33. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999; 401(6755):788–91. <https://doi.org/10.1038/44565> PMID: 10548103
34. Liao R, Zhang Y, Guan J, Zhou S. CloudNMF: A MapReduce Implementation of Nonnegative Matrix Factorization for Large-scale Biological Datasets. *Genomics, Proteomics & Bioinformatics*. 2014; 12(1):48–51. <https://doi.org/10.1016/j.gpb.2013.06.001>.
35. Cai D, He X, Han J, Huang TS. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2011; 33(8):1548–60. <https://doi.org/10.1109/TPAMI.2010.231> PMID: 21173440
36. Yang H, Seoighe C. Impact of the Choice of Normalization Method on Molecular Cancer Class Discovery Using Nonnegative Matrix Factorization. *PLOS ONE*. 2016; 11(10):e0164880. <https://doi.org/10.1371/journal.pone.0164880> PMID: 27741311
37. Padilla P, Lopez M, Gorris JM, Ramirez J, Salas-Gonzalez D, Alvarez I. NMF-SVM Based CAD Tool Applied to Functional Brain Images for the Diagnosis of Alzheimer's Disease. *IEEE Transactions on Medical Imaging*. 2012; 31(2):207–16. <https://doi.org/10.1109/TMI.2011.2167628> PMID: 21914569
38. Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*. 1994; 5(2):111–26. <https://doi.org/10.1002/env.3170050203>
39. Paine MRL, Kim J, Bennett RV, Parry RM, Gaul DA, Wang MD, et al. Whole Reproductive System Non-Negative Matrix Factorization Mass Spectrometry Imaging of an Early-Stage Ovarian Cancer Mouse Model. *PLOS ONE*. 2016; 11(5):e0154837. <https://doi.org/10.1371/journal.pone.0154837> PMID: 27159635
40. Cao B, Shen D, Sun J-T, Wang X, Yang Q, Chen Z, editors. *Detect and Track Latent Factors with Online Nonnegative Matrix Factorization*. *IJCAI*; 2007.
41. Ozaki Y, Aoki R, Kimura T, Takashima Y, Yamada T, editors. *Characterizing muscular activities using non-negative matrix factorization from EMG channels for driver swings in golf*. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016 16–20 Aug. 2016.
42. Ho JC, Ghosh J, Sun J. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*; New York, New York, USA. 2623658: ACM; 2014. p. 115–24.
43. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association: JAMIA*. 2013; 20(1):117–21. <https://doi.org/10.1136/amiajnl-2012-001145> PMC3555337. PMID: 22955496
44. Wei W-Q, Tao C, Jiang G, Chute CG. A High Throughput Semantic Concept Frequency Based Approach for Patient Identification: A Case Study Using Type 2 Diabetes Mellitus Clinical Notes. *AMIA Annual Symposium Proceedings*. 2010; 2010:857–61. PMC3041302. PMID: 21347100
45. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*. 2007; 52(1):155–73. <https://doi.org/10.1016/j.csda.2006.11.006>.
46. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol*. 2011; 2(3):1–27. <https://doi.org/10.1145/1961189.1961199>
47. Kuo CF, Luo SF, See LC, Chou IJ, Chang HC, Yu KH. Rheumatoid arthritis prevalence, incidence, and mortality rates: a nationwide population study in Taiwan. *Rheumatology International*. 2013; 33(2):355–60. <https://doi.org/10.1007/s00296-012-2411-7> PMID: 22451027

48. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence—Volume 2; Montreal, Quebec, Canada. 1643047: Morgan Kaufmann Publishers Inc.; 1995. p. 1137–43.
49. Mejía-Roa E, Tabas-Madrid D, Setoain J, García C, Tirado F, Pascual-Montano A. NMF-mGPU: non-negative matrix factorization on multi-GPU systems. BMC Bioinformatics. 2015; 16(1):43. <https://doi.org/10.1186/s12859-015-0485-4> PMID: 25887585
50. Erichson NB, Mendible A, Wihlborn S, Kutz JN. Randomized nonnegative matrix factorization. Pattern Recognition Letters. 2018; 104:1–7. <https://doi.org/10.1016/j.patrec.2018.01.007>.