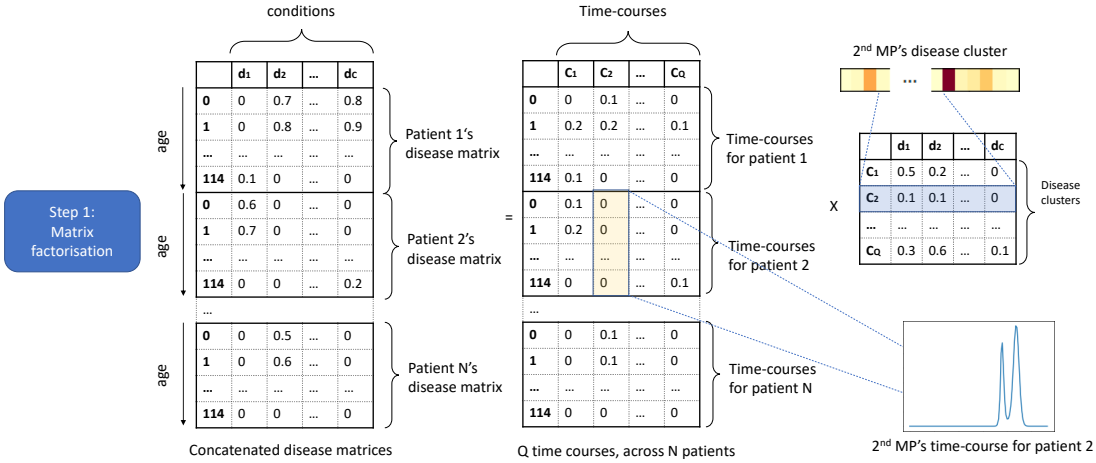
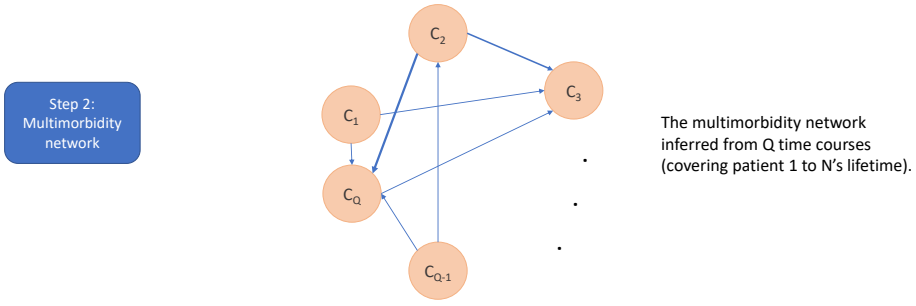
Graphical Abstract

Learning Multimorbidity Patterns from Electronic Health Records Using Non-negative Matrix Factorisation

Abdelaali Hassaine, Dexter Canoy, Jose Roberto Ayala Solares, Yajie Zhu, Shishir Rao, Yikuan Li, Mariagrazia Zottoli, Kazem Rahimi, Gholamreza Salimi-Khorshidi





Highlights

Learning Multimorbidity Patterns from Electronic Health Records Using Non-negative Matrix Factorisation

Abdelaali Hassaine, Dexter Canoy, Jose Roberto Ayala Solares, Yajie Zhu, Shishir Rao, Yikuan Li, Mariagrazia Zottoli, Kazem Rahimi, Gholamreza Salimi-Khorshidi

- Despite the increasing prevalence of multimorbidity (i.e., the presence of several medical conditions in the same individual) in the population, its patterns and trajectories remain poorly understood.
- In this paper, we introduce a new approach for temporal multimorbidity phenotyping – using non-negative matrix factorisation (NMF) – and provide a solution for benchmarking/evaluating the resulting disease clusters and the multimorbidity networks (that are derived from these disease clusters' time-courses).
- Our approach, which is validated on an EHR dataset of 7 million patients, can provide the field of multimorbidity research with new solutions to investigate the patterns by which diseases occur over time in the population and in one's life.

Learning Multimorbidity Patterns from Electronic Health Records Using Non-negative Matrix Factorisation

Abdelaali Hassaine[a,b,c], Dexter Canoy[a,b,c], Jose Roberto Ayala Solares[a,b,c], Yajie Zhu[a,b], Shishir Rao[a,b], Yikuan Li[a,b], Mariagrazia Zottoli[a,b,c], Kazem Rahimi[a,b,c,*], Gholamreza Salimi-Khorshidi[a,b]

[a] *Deep Medicine, Oxford Martin School, University of Oxford, Oxford, United Kingdom*
[b] *The George Institute for Global Health (UK), University of Oxford, Oxford, United Kingdom*
[c] *NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom*

Abstract

Multimorbidity, or the presence of several medical conditions in the same individual, has been increasing in the population – both in absolute and relative terms. Nevertheless, multimorbidity remains poorly understood, and the evidence from existing research to describe its burden, determinants and consequences has been limited. Previous studies attempting to understand multimorbidity patterns are often cross-sectional and do not explicitly account for multimorbidity patterns' evolution over time; some of them are based on small datasets and/or use arbitrary and narrow age ranges; and those that employed advanced models, usually lack appropriate benchmarking and validations. In this study, we (1) introduce a novel approach for using Non-negative Matrix Factorisation (NMF) for temporal phenotyping (i.e., simultaneously mining disease clusters and their trajectories); (2) provide quantitative metrics for the evaluation of these clusters and trajectories; and (3) demonstrate how the temporal characteristics of the disease clusters that result from our model can help mine multimorbidity networks and generate new hypotheses for the emergence of various multimorbidity patterns over time. We trained and evaluated our models on one of the world's largest electronic health records (EHR) datasets, containing

*Corresponding author: kazem.rahimi@georgeinstitute.ox.ac.uk

more than 7 million patients, from which over 2 million where relevant to, and hence included in this study.

Keywords: Non-negative Matrix Factorisation, Multimorbidity, Temporal phenotyping, Disease Trajectories, Electronic Health Records.

1. Introduction

Multimorbidity is generally defined as the presence of two or more chronic conditions in an individual [1]. There is growing evidence that the number of people with multimorbidity has been increasing in many populations, both in relative and absolute terms. This increasing burden has been attributed to a number of factors, including the trend towards an ageing population, as well as factors relating to changes in lifestyle, and health-seeking behaviour and the environment [2]. Research in this area has been growing, but most investigations have focused on predicting, preventing and managing disorders in isolation from one another. Therefore, more research is needed for a better understanding of this emerging burden and its underlying patterns and mechanisms, in order to anticipate its consequences for the health services and the provision of appropriate care [3].

Different types of studies in the past have tried several methods to investigate multimorbidity. We refer to the first group of such studies as "pairwise methods"; in these methods, disease pairs that show co-occurrence frequencies that are different from what their individual frequencies in the population would predict, are considered to be "connected" [4, 5, 6, 7]. Treating diseases as nodes and connectedness as edges, these studies formed networks; the properties of such networks were then used to characterise various multimorbidity pathways. Liu et al. [8] took the idea further and formed a network where nodes were medications, tests and diagnoses. While pairwise methods are valuable in generating comorbidity hypotheses for disease pairs, their inability to address conditional independence (i.e., where the correspondence between disease c_i and c_j is due to a disease c_k they both are linked to, i.e., $P(c_i|c_j, c_k) = P(c_i|c_k)$, or $c_i \perp\!\!\!\perp c_j \mid c_k$ [9]) can make the multi-disease networks resulting from them misleading.

This has led to the rise of alternative methods for disease phenotyping and other approaches of the study of multimorbidity that can deal with multiple diseases simultaneously. Except a limited number of studies that used "probabilistic methods" such as latent class growth analysis [7] and Hidden Markov Model [10], majority of recent studies in the field have relied on "factorisation methods". The earlier factorisation methods started by forming a matrix, where the entry i, j denotes a metric related to disease i (or other concepts in EHR, such as medication or clinical measurements) in patient j; factor methods decompose such a matrix into Q multimorbidity patterns (MPs), each consisting of a disease cluster (DC) and the expression of that DC across patients. While relatively effective, such matrix factorisation approaches did not take into account the temporal aspects of MPs, and have been mostly limited to mining static disease (and sometimes medication) clusters [11, 12, 13, 14, 15]. As an exception, Zhou et al. [16] introduced a matrix factorisation method that considers the temporal patterns in EHR data; while their method can be potentially employed for multimorbidity analyses, the primary focus was on the prediction of future diseases (as opposed to multimorbidity).

On the other hand, solutions for joint phenotyping (based on diseases plus other concepts, for instance) have been achieved through the use of tensor factorisation. For instance, both [17] and [18], used non-negative tensor factorisation for joint phenotyping of diagnoses and medications, through adding medications as a new dimension to the input matrix (and hence the tensor). Following a similar approach Perros et al. [19], introduced a time dimension and used tensor factorisation for joint phenotyping of diseases and time (i.e., disease-based temporal phenotyping); instead of explicitly dealing with time, however, their model deals with the chronological order/index of encounters. In another recent study of temporal phenotyping, Zhao et al. [20] used tensor factorisation and showed the effectiveness of their resulting phenotypes for stratifying the risk of cardiovascular diseases. In summary, previous temporal phenotyping studies relied on relatively complex tensor-based factorisation approaches, which makes the use of simpler matrix factorisation techniques for this goal an under-explored topic.

In this study, we introduce a novel design that can enable matrix factorisation

techniques, such as NMF [21, 22], for disease-based temporal phenotyping (i.e., to mine the DCs and their expression over time in patients). This was facilitated by an assumption that is key to our study: the underlying DCs are the same for all patients, but their expression pattern (i.e., the strength of a DC's expression in a given year of one's life) varies from one person-age to the other. We trained our models on one of the world's largest EHR datasets – consisting of more than 7 million individuals (2 million were appropriate for this study) – and evaluated its results for the study of multimorbidity. Given that the past studies of multimorbidity were often cross-sectional, used arbitrary narrow age ranges, and lacked appropriate benchmarking and validations (of their methodology and results), this paper can introduce a new methodology pipeline to the field of multimorbidity research that can employ the temporal information – hidden in longitudinal datasets such as EHR – to generate and evaluate new hypotheses on how diseases occur over time.

2. Materials and methods

In this section, we explain the source of our EHR data, our approach in using NMF, and the network analyses that we carried out to show the power of temporal disease-based phenotyping for the study of MPs. Note that, for the rest of the paper, matrices will be denoted by upper case bold fonts (e.g., $\mathbf{A}$), vectors will be denoted by lower case bold fonts (e.g., $\mathbf{a}$), and everything else (scalar and indices) will be denoted with no bolding of the fonts.

2.1. EHR Data

An important development that can help address the aforementioned issues in the current state of multimorbidity studies, is the rapid growth in the adoption of healthcare information systems, and the growing interest in utilising EHR. Particularly, the longitudinal nature of EHR and their richness (e.g., containing diagnoses, medications, and tests/measurement) can provide a unique opportunity to study temporal multi-modal phenotyping. In this study, we used the Clinical Practice Research Datalink (CPRD) [23]; a longitudinal primary care data from a network of 674 general practices (GP) in the UK, linked to secondary care (i.e., Hospital

Episode Statistics, HES) and other health and administrative databases (e.g., Office for National Statistics' death registration). The data encompass 45 million patients, including 13 million currently registered patients. Our data cut is based on an older version with 7 million patients. CPRD is broadly representative of the population by age, sex, and ethnicity [24]. It has been extensively validated and is considered as the most comprehensive longitudinal primary care database [25], with several large-scale epidemiological reports [26, 27, 28] adding to its credibility.

HES, on the other hand, contains data on hospitalisations, outpatient visits, accident and emergency for all admissions to National Health Service (NHS) hospitals in England [29]. Approximately 75% of the CPRD GPs in England (58% of all UK CPRD GPs) participate in patient-level record linkage with HES, which is performed by the Health and Social Care Information Centre [30]. In this study, we only considered the data from GPs that consented to (and hence have) record linkage with HES. The importance of primary care at the central point of the national health system in the UK, the additional linkages, and all the aforementioned properties, make CPRD one of the most suitable EHR datasets in the world for data-driven clinical/medical discovery and machine learning.

2.2. Study Population

In this study, we only included patients whose primary and secondary care data are linked; for the included patients, we only considered disease-related events in both GP and HES data (after removing the duplicates events). In order to have a comprehensive coverage of patients' health journey, we only considered patients with at least 5 years of follow-up; this resulted in a total number of 2,204,178 patients (among whom 994,563 were men and 1,209,615 were women) with a total number of 25,791,493 clinical encounters. Note that, in this study, we are interested in "incident cases" and not in the "prevalent cases" of diseases, i.e. new occurrences of diseases rather than diseases carried over. Therefore, we only considered the first occurrence of each disease happening after 1 year of any patient's registration date with the general practice clinic (as the first year after registration is likely to contain diseases carried over rather than new occurrences of diseases). This process is summarised in

Figure 1.

Another important step in processing CPRD was to create consistent disease classifications between GP and HES data and choosing the appropriate level of granularity in diseases' hierarchy . In HES, diseases are coded using ICD-10 (International Classification of Diseases [31]), whereas in the GP records diseases are coded using Read Code [32]. ICD-10 offers a hierarchical form that makes its use for data mining and machine learning much more convenient. Therefore, we decided to map the diseases to the ICD-10 domain; that is, we mapped the Read Codes to ICD-10 codes using the mapping provided by NHS Digital [33]. When no direct mapping was available, Read Codes were first mapped to SNOMED-CT codes [34] (also provided by NHS Digital [35]), and the latter then mapped to ICD-10-CM codes using the mapping provided by the US National Library of Medicine [36]. Note that ICD-10-CM is the US version of the ICD-10 codes and is generally more granular than the official WHO ICD-10 code; when working at the block level of the hierarchy, however, the two coding schemes are equivalent.

As mentioned earlier, ICD-10 codes are organised in a tree-like hierarchy. Working at the highest level of the ICD-10 hierarchy will result in only a few diseases and hence is not likely to help unravel the complex underlying dependencies among diseases. Conversely, working at the lowest level of the hierarchy will generate thousands of diseases, each with small number of occurrences (and even a smaller number of co-occurrences) and hence leads to an overall difficulty of mining useful inter-disease patterns. In this study, similarly to the work in [15], we chose to work at the ICD-10 block level, which provides a good trade off between granularity and co-occurrence, and can lead to medically interpretable results. Furthermore, similar to [6], we eliminated all the diagnoses relating to pregnancy, general symptoms, external causes and administration (i.e., ICD-10 chapters XV, XVI, XVIII, XIX, XX and XXI); . The resulting 142 ICD-10 blocks are what we will refer to as diseases from here on.

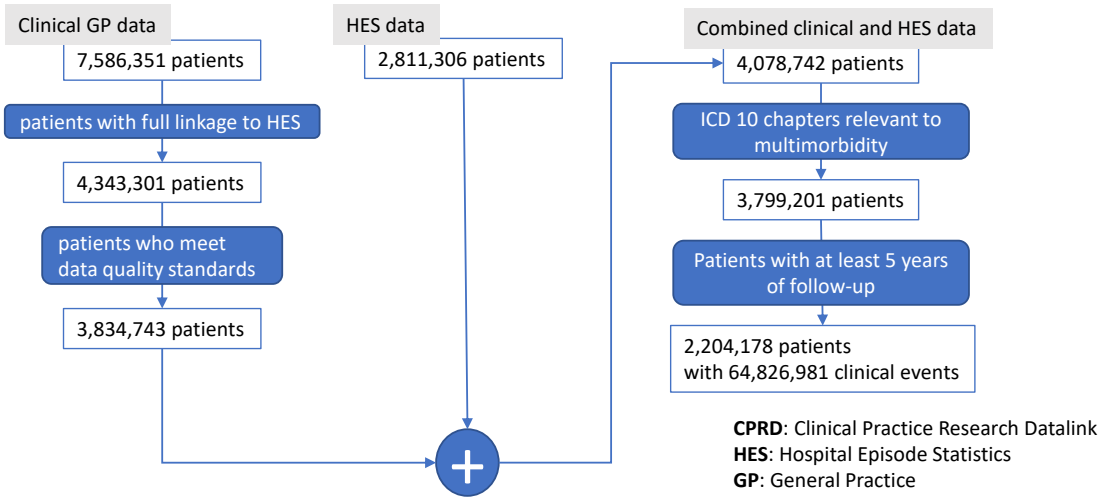
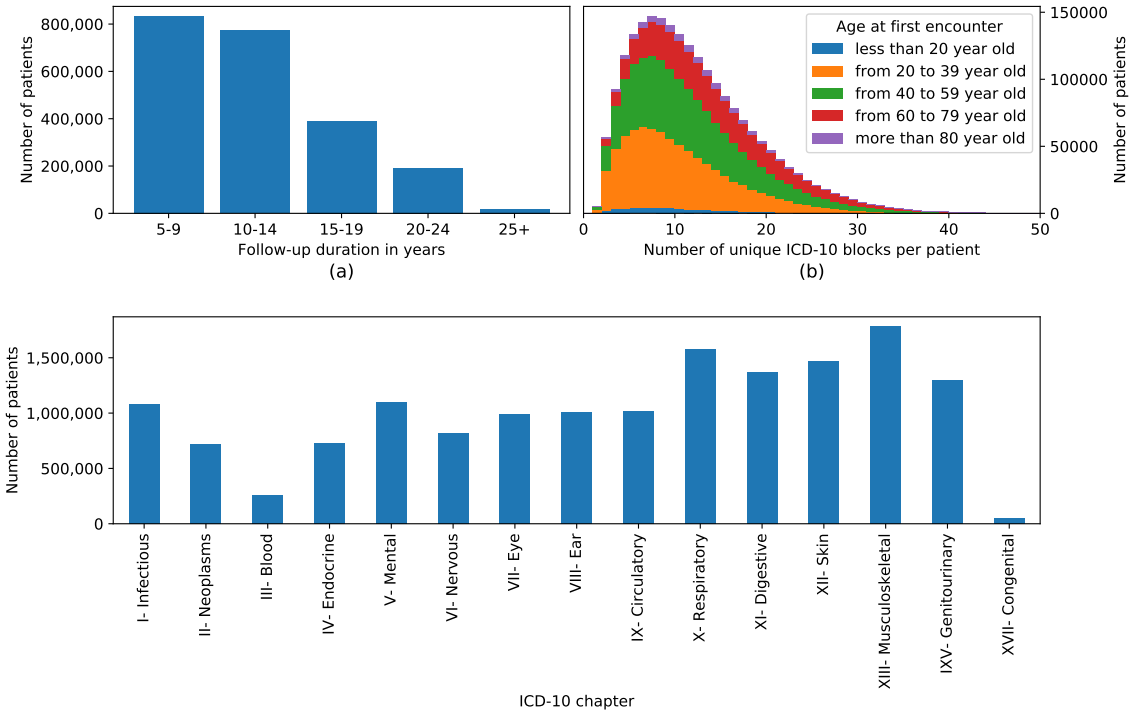


Figure 1: Inclusion/exclusion criteria for the study. In order to make sure that the data is appropriate for the study's objectives (described at the end of Section 1), we only included the patients that meet our criteria, e.g., follow up duration, record quality (as indicated by CPRD data manual), linkage between primary and secondary care, and having sufficient follow-up information. Note that the number of patients after combining GP and HES data does not shrink as there are patients with GP events without any event in HES and vice-versa.

At the end of this process, the subset of CPRD that we will use for this study includes nearly 2.2M adult patients (aged 16 years and over) and 65M events ranging from calendar years 1985 to 2014; from here on, we will refer to this subset as data. Figure 2 shows some of the key characteristics of the data, including the distributions of year of birth for both male and female patients, distributions of ages and follow-up durations, number of patients per disease count, as well as the number of patients for each ICD-10 block and chapter. When compared to the past studies, in terms of the number of patients, length of follow-up and being representative of the population, the data in this study are substantially better than those of the past studies.



Age (year) at first encounter	Male patients	Female patients
0-19	24,955	31,375
20-29	125,524	186,888
30-39	207,915	258,420
40-49	205,799	223,315
50-59	184,335	189,636
60-69	142,339	154,549
70-79	82,868	116,721
80+	20,828	48,711
Total	**994,563**	**1,209,615**
All patients	**2,204,178**	

(d)

Figure 2: An exploratory analysis of the data, showing the key characteristics of the population we studied: (a) Number of patients per follow-up duration, (b) number of patients per counts of unique diseases, for each age ranges, (d) number of unique patients per ICD-10 chapter, and (d) age range at the first event for male and female patients, per age band.

2.3. Non-negative Matrix Factorisation

Non-negative matrix factorisation (NMF, or NNMF) refers to a group of algorithms that decompose a matrix $\mathbf{D}$ into (usually) two matrices $\mathbf{A}$ and $\mathbf{B}$, with the property that all three matrices have no negative elements, i.e.,

$$\mathbf{D} \approx \mathbf{A} \times \mathbf{B} \mid \mathbf{A} \geq 0, \mathbf{B} \geq 0. \tag{1}$$

This non-negativity makes the resulting matrices easier to inspect and interpret. Also, in applications such as processing of count data, which is the starting point of our analysis, non-negativity is inherent to the data being considered. Since the NMF problem does not have an exact analytical solution in general, there has been a range of numerical approximations for it [37, 38, 39, 40, 41, 42, 43, 44, 45, 46]. In this paper, we use the Kullback-Leibler divergence and simple multiplicative updates [37, 39], enhanced to avoid numerical underflow [41] as implemented by Nimfa package in Python [44].

2.4. Modelling Pipeline

Our NMF modelling pipeline is a five-step process. The first step of the pipeline starts by forming a "disease matrix" $\mathbf{D}_p$ for each patient p, where $\mathbf{D}_p(i, j) = 1$ if patient p had the first incidence of disease i at age (in years) j ($\mathbf{D}_p(i, j) = 0$ otherwise). This makes $\mathbf{D}_p$ a $T \times C$ matrix, where C is the number of unique conditions/diseases in the study (i.e., 142 in this study) and T is the maximum age a patient is tracked for (i.e., 114 years in this study). Denoting the total number of patients by N, this process will result in N such $\mathbf{D}_p$ matrices.

Given the variability in disease prevalences (i.e., some diseases are more common than others), the counts corresponding to rare diseases (such as tuberculosis) are expected to be much lower than the counts corresponding to more common diseases (such as respiratory infections). Therefore, when carrying out the NMF analysis, the results can be biased towards explaining the more frequent diseases (i.e., the higher counts). In order to correct for this, in the second step of the pipeline, we used an adjustment inspired by TF–IDF (term frequency–inverse document frequency),

which is commonly used in natural language processing and information retrieval [47]. More specifically, we introduce DF–IPF as the product of two statistics: disease frequency (DF) and inverse patient frequency (IPF), where $DF = 1$ (as we only considered the first occurrence of each disease), and $IPF(i) = \log(N/N_i)$ for disease i (with N_i denoting the number of patients who had disease i and N being the total number of patients in our data cut). The DF–IPF adjustment for $\mathbf{D}_p$ will simply result from the multiplication of its entries with the appropriate inverse patient frequency, i.e., $\mathbf{D}_p(i,j) \leftarrow \mathbf{D}_p(i,j) * DF * IPF(i)$.

As each patient will only have a relatively small number of diseases, $\mathbf{D}_p$ is expected to be sparse. On the other hand, NMF does not explicitly model age as a temporal concept (i.e., no explicit model for the relationship among rows in $\mathbf{D}$). In medicine, however, one will not see a meaningful difference between a disease happening at age a or $a\pm$ 1 or 2 years; in factorisation of a matrix like $\mathbf{D}$ using standard NMF, this property will not exist and the two scenarios will not be seen as similar. Furthermore, we know from the medical practice that the date at which certain chronic disease gets recorded is a noisy concept. For instance, one's diabetes diagnosis can be delayed by months or years due to not noticing or ignoring the symptoms, and/or delaying a doctor visit; or, diseases such dementia are known to have long preclinical periods, where patients who visit their doctors less regularly are, on average, more likely to have their diagnosis delayed. Therefore, in the third step of our modelling pipeline, we smooth each column of $\mathbf{D}_p$, using a Gaussian kernel with a standard deviation σ – the optimal value of this parameter will be determined empirically. We expect this step to address NMF's lack of temporal regularisation, and help it take into account the noise and/or uncertainty we have around the age that a particular disease has actually occurred.

As mentioned earlier, to the best of authors' knowledge, the only MP analyses that took time into account used *tensor* factorisation methods. Therefore, one of the key contributions of this paper is to enable *matrix* factorisation techniques to result in temporal phenotyping. Assuming that DCs are the same across all patients – and what varies from patient to patient is the strength of these DCs expressing themselves at different patient-ages – enables us to use NMF in a unique way for

temporal phenotyping. In the fourth step of our modelling pipeline, we concatenate the $\mathbf{D}_p$ matrices along the age axis and form $\mathbf{D}$ (i.e., $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_N]$) – a $(T * N) \times C$ matrix. The fifth and final step of our modelling pipeline is the NMF decomposition of $\mathbf{D}$: For a given number of factors, Q, this will result in two matrices $\mathbf{A}$ and $\mathbf{B}$ (a la Eq. 1), where $\mathbf{A}$ is $(T * N) \times Q$ matrix and $\mathbf{B}$ a $Q \times C$ matrix. A row in $\mathbf{B}$ is a DC (i.e., what we assumed are the same for all patients), and a column in $\mathbf{A}$ is the time-course associated with a DC across patients (i.e., how strongly a DC is expressed in each and every patient-age). In summary, we factorise $\mathbf{D}$ (which can be viewed as the population/group disease matrix) using NMF; our assumption that DCs are the same across patients (and what's variable is their expression in a patient-age), enabled us to interpret $\mathbf{A}$ and $\mathbf{B}$ that result from this factorisation as matrices containing DCs and their time-courses, respectively. See Figure 3 for an illustration of our NMF analysis, and Table 1 for a glossary of some of the terms introduced in this paper and frequently used in for describing the methodology.

Note that, in this study we focus on disease-based temporal phenotyping, but the approach we introduce can easily incorporate additional clinical features (e.g., medications) to the input matrix and result in temporal phenotypes that are not disease-only.

2.5. Ascendency Analysis

While mining disease clusters is the key objective of multimorbidity research, providing a view into how such clusters evolve over time is another important goal of multimorbidity research that is less studied. Given that our approach results in a time-course for each cluster, we aim to treat these clusters as nodes and use their time-courses to define their connectivity. More specifically, we employed a simple network modelling technique, which was originally introduced by Patel et al [48] for the study of the networks in the brain; it was shown to outperform many network modelling techniques in both finding edges and their directions in a range of settings where ground truth for network edges are available (i.e., a simulation study) [49].

Imagine we have two DCs, with binary time-courses $\mathbf{v}$ and $\mathbf{w}$ (each of length T*N); in our method, if $\mathbf{v}$ and $\mathbf{w}$ are active together and inactive together, we con-

Table 1: Glossary of the terms frequently used in this paper, when describing the methodology.

Term	**Definition**
Disease cluster (DC), or multimorbidity clusters (MC)	Each DC (or MC) is a vector; the entries are weights that are assigned to each and every disease to denote their extent of belonging to the cluster. Note that, as implied, each disease can belong (i.e., have nonzero values) to more than one DC.
DC time-course	The strength by which a DC is expressed in a given patient over time. For instance, if the time-course shows a larger value at age 50 than 40, it means that its corresponding DC has a stronger expression/presence at age 50 than 40.
Multimorbidity pattern (MP) and temporal phenotype	In our study, each MP or temporal phenotype consists of a DC and its corresponding time-course. For instance, our NMF approach, decomposes $\mathbf{D}$ to a matrix of DCs (i.e., $\mathbf{B}$) plus a matrix of their time-courses (i.e., $\mathbf{A}$). An MP might also be referred to as component (e.g., in PCA and ICA) or factor (e.g., in NMF)

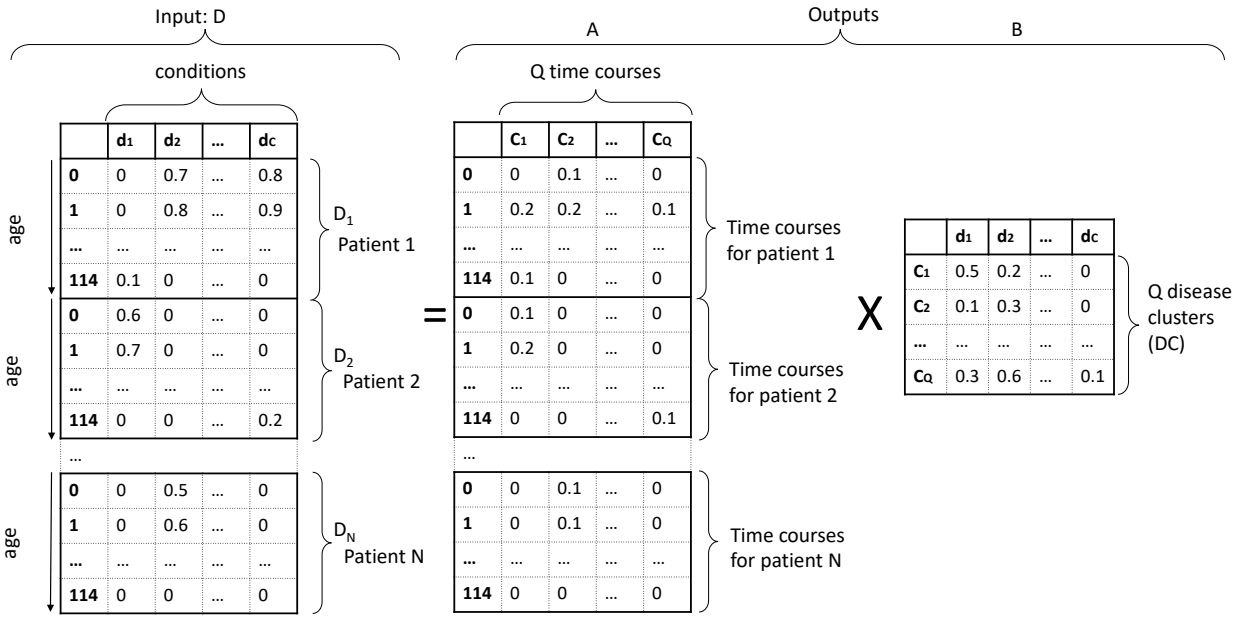


Figure 3: An illustration of formation and decomposition of disease matrix $\mathbf{D}$. As the figure shows, $\mathbf{D}$ results from vertical concatenation of $\mathbf{D}_p$ matrices. The matrix $\mathbf{B}$ resulting from NMF decomposition has Q rows, each denoting a disease cluster (i.e., a vector of C weights, denoting each disease's belonging to that cluster). The strength of the q'th cluster's expression in each patient-age can be found in the q'th column of $\mathbf{A}$, at the entry that corresponds to that patient-age.

sider them associated/connected. This concept is measured by $\kappa_{\mathbf{v},\mathbf{w}} \in [-1, 1]$, and can be seen as an undirected edge in a network; it monotonically increases with joint activation of $\mathbf{v}$ and $\mathbf{w}$, as their individual activations remain fixed; it monotonically decreases with $\mathbf{v}$'s activation as $\mathbf{w}$'s activation and the joint activation remain fixed, or conversely, it monotonically decreases with $\mathbf{w}$'s activation as $\mathbf{v}$'s activation and the joint activation remain fixed; it will be 0 when $\mathbf{v}$ and $\mathbf{w}$ are statistically independent. More formally, in this approach a bivariate Bernoulli Bayesian model is constructed for the joint activation of each pair of time-courses using a multinomial likelihood with a Dirichlet prior. The data to model the joint activation/inactivation

probabilities for the two time-courses be:

$$\begin{aligned} z_1 &= \textstyle\sum_{T*N} I(\mathbf{v}=1,\mathbf{w}=1) \\ z_2 &= \textstyle\sum_{T*N} I(\mathbf{v}=1,\mathbf{w}=0) \\ z_3 &= \textstyle\sum_{T*N} I(\mathbf{v}=0,\mathbf{w}=1) \\ z_4 &= \textstyle\sum_{T*N} I(\mathbf{v}=0,\mathbf{w}=0), \end{aligned} \tag{2}$$

The multinomial likelihood of our data takes the form:

$$p(z|\theta) \propto \prod_{i=1}^{4} \theta_i^{z_i}, \tag{3}$$

where each parameter θ is defined as

$$\begin{aligned} \theta_1 &= P(\mathbf{v}=1,\mathbf{w}=1) \\ \theta_2 &= P(\mathbf{v}=1,\mathbf{w}=0) \\ \theta_3 &= P(\mathbf{v}=0,\mathbf{w}=1) \\ \theta_4 &= P(\mathbf{v}=0,\mathbf{w}=0). \end{aligned} \tag{4}$$

κ is defined as a ratio, with its numerator measuring the difference between the joint activation probability and the expected joint activation probability under independence, and its denominator forcing it to range from -1 to 1, i.e.,

$$\kappa = \frac{\theta_1 - E}{D(max(\theta_1) - E) + (1 - D)(E - min(\theta_1)))} \tag{5}$$

where $E = (\theta_1 + \theta_2)(\theta_1 + \theta_3)$, $max(\theta_1) = min(\theta_1 + \theta_2, \theta_1 + \theta_3)$, $min(\theta_1) = max(0, 2\theta_1 + \theta_2 + \theta_3 - 1)$ and

$$D = \begin{cases} \frac{\theta_1 - E}{2(max(\theta_1) - E)} + 0.5, \text{ if } \theta_1 \geq E \\ 0.5 - \frac{\theta_1 - E}{2(E - min(\theta_1))}, \text{ otherwise.} \end{cases} \tag{6}$$

Given the connectivity between two nodes with $\mathbf{v}$ and $\mathbf{w}$ time-courses, if $\mathbf{v}$ exhibits an elevated expression for a subset of the period in which $\mathbf{w}$ exhibits an elevated expression, we consider $\mathbf{w}$ to be ascendant to $\mathbf{v}$ and vice versa. This is measured by $\tau_{\mathbf{v},\mathbf{w}} \in [-1, 1]$, and $\tau_{\mathbf{v},\mathbf{w}} > 0$ means that $\mathbf{v}$ is ascendant to $\mathbf{w}$, whereas $\tau_{\mathbf{v},\mathbf{w}} < 0$ means that $\mathbf{w}$ is ascendant to $\mathbf{v}$. More formally:

$$\tau_{\mathbf{v},\mathbf{w}} = \begin{cases} 1 - \frac{\theta_1+\theta_3}{\theta_1+\theta_2)}, \text{ if } \theta_2 \geq \theta_3 \\ \frac{\theta_1+\theta_2}{\theta_1+\theta_3} - 1, \text{ otherwise} \end{cases} \tag{7}$$

This idea is illustrated in Figure 4.

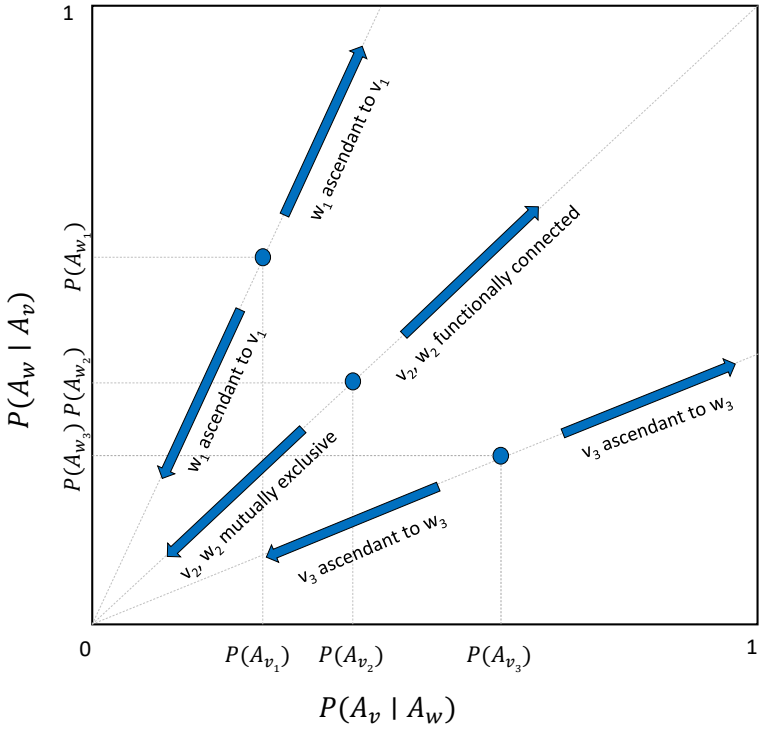


Figure 4: Three MP pairs (v_1,w_1), (v_2,w_2), and (v_3,w_3) are illustrated, each with a different hierarchical relationship. A_v denotes an indicator of elevated expression of node v. As the slope of the line from the origin to $(P(A_v), P(A_w))$ gets further from 1, the degree of ascendancy between the MP pair increases. The ascendancy metric of two nodes is determined by the ratio of their respective expression probabilities. Specifically, for two functionally connected time-courses **v** and **w**, we say that **v** is ascendant to **w** if the probability of an elevated expression of **v** is larger than that of **w**. Credits to Patel et al. [48]

2.6. Benchmarks and Evaluations

It is always challenging to assess the quality of unsupervised/exploratory models, particularly when there is limited access to relevant objective data on their goodness.

Therefore, in this study, we aim to provide the field with an objective approach to evaluate the goodness of the DCs that result from multimorbidity research. We used two different sources of objective medical knowledge, each providing a list of "comorbid" disease pairs: (1) Jensen et al. [6] concluded a list of 1,556 comorbid disease pairs based on the analysis of a large national health dataset, followed by thorough medical due diligence; and (2) Beam et al. [50] studied a large body of medical literature to conclude their list of 113 comorbid condition pairs. We denote a comorbid disease pair as $p_{ij} = (c_i, c_j)$, which means according to the medical literature, diseases c_i and c_j are comorbid. Furthermore, through the study of large body of medical literature Beam et al. [50] also concluded a list of 156 causal pairs, where one disease causes the other. We denote a causal disease pair as $p_{i \to j} = (c_i, c_j)$, which means disease c_i causes disease c_j. Note that, the number of unique pairs mentioned are the counts after converting the disease notations in the source to ICD-10 blocks (i.e., the level at which our analysis has been carried out).

In order to assess the correspondence between the comorbid disease pairs in the medical literature (i.e., ground truth) and the DC's obtained from our analysis, we defined a score C_L, to measure the proportion of comorbid disease pairs, whose constituent diseases fall in the same DC. For instance, in order to evaluate whether or not our DCs capture the relationship between diabetes and hypertension, we compute the proportion of DCs which have both diabetes and hypertension over the number of DCs which have either diabetes or hypertension. More formally, the k'th DC can be represented as $\mathbf{q}_k = \{w_{ki}\}_{i=1}^{C}$, where w_{ki} is the weight for disease c_i in the k'th DC; following this representation, $\mathbf{l}_k$ denotes the list of diseases that correspond to top L weights in $\mathbf{q}_k$. We defined $C_L = a/b$, where b is the number of p_{ij}'s such that either c_i or c_j belong to the top L diseases in any cluster (i.e., belong to $\cup_{k=1}^{Q} \mathbf{l}_k$), and a is the number of p_{ij}'s such that both c_i and c_j belong to the top L diseases in the same cluster (i.e., there is at least one k, such that $\mathbf{l}_k$ contains both c_i and c_j). Under this definition, $C_L = 0$ indicates that there is no disease pair that has both its constituent diseases in the same $\mathbf{l}_k$; on the other hand, $C_L = 1$ indicates that for every disease pair that has a constituent disease in $\cup_{k=1}^{Q} \mathbf{l}_k$, there is at least one $\mathbf{l}_k$ that has both its constituent diseases.

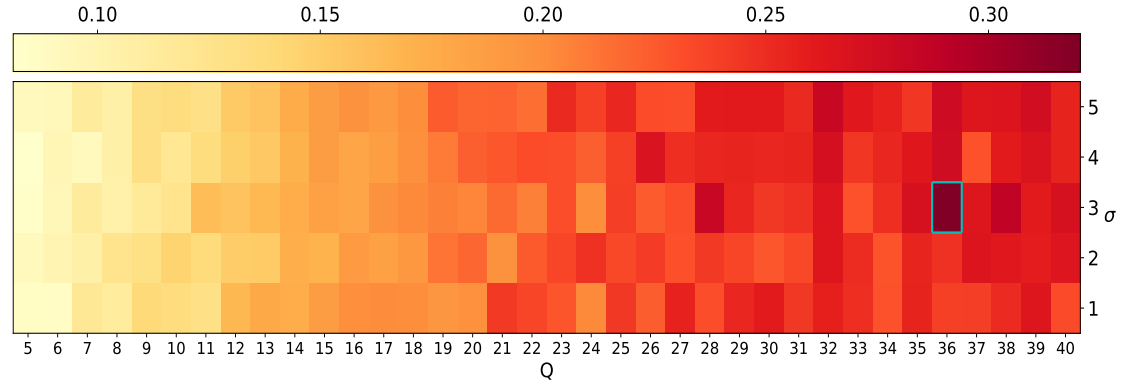
We defined a similar score based on causal pairs, to assess the goodness of our ascendancy analysis. An ascendancy edge from our analyses can be denoted as $o_{m\to n}$ (meaning an arrow from cluster m to n). We defined $A_L = a/b$, where b is the number of $p_{i\to j}$'s that either i or j belong to the top L diseases in either $\mathbf{q}_m$ or $\mathbf{q}_n$, and a is the number of $p_{i\to j}$'s such that diseases i and j belong to the top L diseases in clusters $\mathbf{q}_m$ and $\mathbf{q}_n$, respectively. Less formally, A_L measures the proportion of causal pairs from the medical literature that traverse our ascendancy graph when considering the top L diseases in each DC.

Given that the optimal values of our modelling pipeline are expected to lead to better C_L, it can be used to guide the grid search for σ and Q. Furthermore, In order to check if the C_L and A_L values we observe are higher than what would be expected by chance alone, we sampled from their distribution under the null hypothesis ($\mathcal{H}_0$) that diseases co-occur at random. Under this $\mathcal{H}_0$, we can permute the disease labels in our NMF results; in each permutation, we shuffle the labels of matrix $\mathbf{B}$'s columns, and calculate the corresponding C_L and A_L by comparing them to the ground truth pairs. We carried out this label permutation for 1,000 times; the resulting null distributions for C_L and A_L have been used to assess the statistical significance of the observed A_L and C_L values.

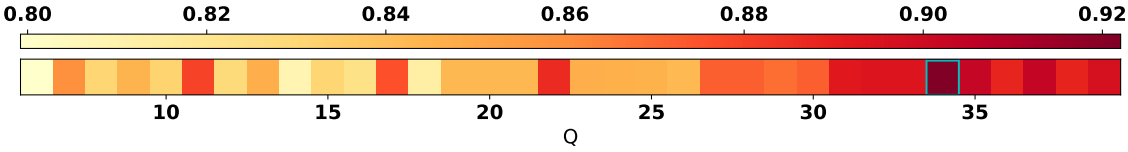
3. Results

Our NMF analysis relies on some key free parameters: Q (i.e., the number of DCs) and the standard deviation of the Gaussian filter, σ. We computed C_L for a range of of Q and σ values on Jensen disease pairs, which consists of a more comprehensive list of disease pairs than Beam, and hence more suitable for the overall parameter tuning. The resulting C_L scores are shown in Figure 5.a; using these results, we see a better performance at $\sigma = 3$ and $Q \approx 36$. In order to find the best value for Q, we calculated the cophenetic correlation coefficient, which measures the stability of clustering derived from NMF results at a given rank [41] for a range of Q values. According to the results shown in Figure 5.b, we see that the factorisation at $Q = 34$ results in the highest score. Thus, the rest of our analyses and results will be based

on 34 DCs. Furthermore, to illustrate an example P-value estimation, Figure 6 shows the value of C_5 when using 1,000 random permutations on Jensen pairs.



(a)



(b)

Figure 5: (a) Evaluation of DCs, using C_L, over a 2D grid corresponding to different values of σ and Q. The highest values (highlighted in cyan border) are observed at $\sigma = 3$ and at $Q \approx 36$. All P-values are ≤ 0.00001. (b) The change in cophenetic coefficient as a function of number of factors. As can be seen, 34 seems to be the optimal number of factors for the factorisation of $\mathbf{D}$ using NMF.

Another free parameter of our analysis is the binarisation threshold that we used to prepare the time-courses for ascendancy analysis. In order to set this parameter, we calculated A_L for a range of L values at different binarisation thresholds and saw the best results to be at 75%. Therefore, for the rest of our analysis we thresholded/binarised each time-course $\mathbf{w}$ at $0.75 * max(\mathbf{w})$.

Given that our approach to NMF does not include additional variables such as gender, we decided to carry out the analyses separately for men and women. Tables 2 and 3 shows the top 3 diseases in each DC along with their corresponding weights for male for female patients respectively (readers are referred to an alternative

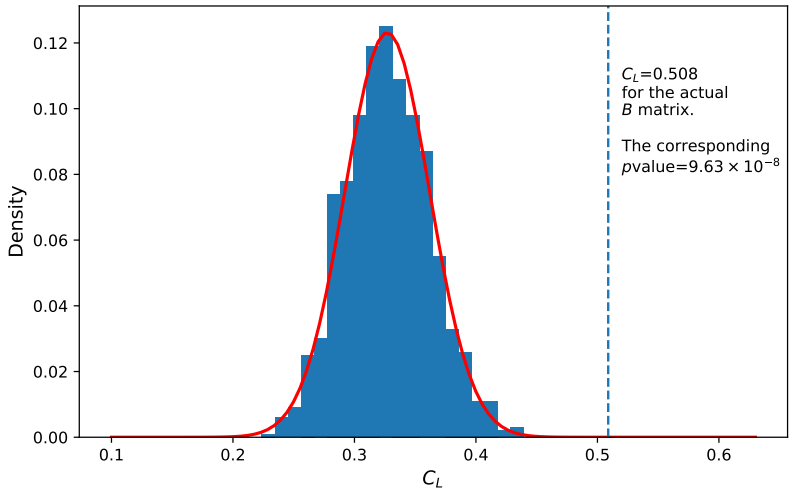


Figure 6: Histogram of C_5 on pairs by Jensen et al. for 1,000 random permutations of the disease labels in the B matrix (corresponding to the male patients analysis).

representation using a heatmap in Figure A.2 in the appendix).

Note that, corresponding to each one of these clusters we have a time-course which shows how that DC is expressed for a given patient at any given age. Interestingly, many DCs have diseases that are known comorbidities. For instance, DC^m_{22} shows that Papulosquamous disorder (L40-L45) co-occurs with Obesity (E65-E68). Obesity is also associated with Hypertension (I10-I15) in DC^m_{26}, and with Diabetes mellitus (E08-E14) in DC^m_{20}. These links are also confirmed for female patients in DC^f_{14}. Renal failure (N17-N19) is also associated with circulatory diseases as highlighted in DC^m_9 for male patients or DC^f_{28} for female patients. Also, note that some disease clusters may highlight associations between seemingly unrelated conditions, this is often triggered by a common factor between these conditions such as psychoactive substance use (F10-F19) as in DC^m_{29}.

Table 2: Disease clusters for male patients. Top 3 diseases are shown for each cluster. The weights denote each disease's belonging to the cluster, where 1 denotes a strong presence whereas 0 denotes an absence.

DC_1^m	J09-J18: Influenza and pneumonia (1.00) J90-J94: Other diseases of pleura (0.20) J20-J22: Other acute lower respiratory infections (0.18)	DC_{18}^m	I60-I69: Cerebrovascular diseases (1.00) G40-G47: Episodic and paroxysmal disorders (0.43) G80-G83: Cerebral palsy and other paralytic syndromes (0.37)
DC_2^m	N40-N53: Diseases of male genital organs (1.00) N30-N39: Other diseases of urinary system (0.17) A50-A64: Infections with a predominantly sexual mode of transmission (0.10)	DC_{19}^m	H30-H36: Disorders of choroid and retina (1.00) H43-H45: Disorders of vitreous body and globe (0.49) H40-H42: Glaucoma (0.15)
DC_3^m	H65-H75: Diseases of middle ear and mastoid (1.00) H90-H95: Other disorders of ear (0.50) J00-J06: Acute upper respiratory infections (0.14)	DC_{20}^m	E08-E14: Diabetes mellitus (1.00) E15-E16: Other disorders of glucose regulation and pancreatic internal secretion (0.30) E65-E68: Obesity and other hyperalimentation (0.27)
DC_4^m	A30-A49: Other bacterial diseases (1.00) J90-J94: Other diseases of pleura (0.21) E65-E68: Obesity and other hyperalimentation (0.16)	DC_{21}^m	N20-N23: Urolithiasis (1.00) N10-N16: Renal tubulo-interstitial diseases (0.28) N30-N39: Other diseases of urinary system (0.15)
DC_5^m	C00-C97: Malignant neoplasms (1.00) D37-D49: Neoplasms of uncertain or unknown behaviour (0.46) D00-D09: In situ neoplasms (0.31)	DC_{22}^m	L40-L45: Papulosquamous disorders (1.00) E65-E68: Obesity and other hyperalimentation (0.65) F10-F19: Mental and behavioural disorders due to psychoactive substance use (0.19)
DC_6^m	I80-I89: Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified (1.00) I26-I28: Pulmonary heart disease and diseases of pulmonary circulation (0.19) L00-L08: Infections of the skin and subcutaneous tissue (0.14)	DC_{23}^m	L55-L59: Radiation-related disorders of the skin and subcutaneous tissue (1.00) L80-L99: Other disorders of the skin and subcutaneous tissue (0.35) B00-B09: Viral infections characterized by skin and mucous membrane lesions (0.15)
DC_7^m	D60-D64: Aplastic and other anaemias (1.00) D50-D53: Nutritional anaemias (0.91) D70-D77: Other diseases of blood and blood-forming organs (0.85)	DC_{24}^m	E70-E90: Metabolic disorders (1.00) I10-I15: Hypertensive diseases (0.09) K70-K77: Diseases of liver (0.09)
DC_8^m	F50-F59: Behavioural syndromes associated with physiological disturbances and physical factors (1.00) G40-G47: Episodic and paroxysmal disorders (0.07) M60-M79: Soft tissue disorders (0.07)	DC_{25}^m	H53-H54: Visual disturbances and blindness (1.00) H15-H22: Disorders of sclera, cornea, iris and ciliary body (0.26) H55-H59: Other disorders of eye and adnexa (0.12)
DC_9^m	N17-N19: Renal failure (1.00) I30-I52: Other forms of heart disease (0.85) I95-I99: Other and unspecified disorders of the circulatory system (0.46)	DC_{26}^m	I10-I15: Hypertensive diseases (1.00) E65-E68: Obesity and other hyperalimentation (0.27) H80-H83: Diseases of inner ear (0.18)
DC_{10}^m	K50-K52: Noninfective enteritis and colitis (1.00) K55-K64: Other diseases of intestines (0.17) L60-L75: Disorders of skin appendages (0.08)	DC_{27}^m	K80-K87: Disorders of gallbladder, biliary tract and pancreas (1.00) K70-K77: Diseases of liver (0.40) K20-K31: Diseases of oesophagus, stomach and duodenum (0.17)
DC_{11}^m	F30-F39: Mood [affective] disorders (1.00) G30-G32: Other degenerative diseases of the nervous system (0.96) F40-F48: Neurotic, stress-related and somatoform disorders (0.71)	DC_{28}^m	H60-H62: Diseases of external ear (1.00) H90-H95: Other disorders of ear (0.40) G50-G59: Nerve, nerve root and plexus disorders (0.37)
DC_{12}^m	I20-I25: Ischaemic heart diseases (1.00) I30-I52: Other forms of heart disease (0.23) J20-J22: Other acute lower respiratory infections (0.08)	DC_{29}^m	F10-F19: Mental and behavioural disorders due to psychoactive substance use (1.00) K00-K14: Diseases of oral cavity, salivary glands and jaws (0.93) B95-B98: Bacterial, viral and other infectious agents (0.47)
DC_{13}^m	B85-B89: Pediculosis, acariasis and other infestations (1.00) L20-L30: Dermatitis and eczema (0.20) L00-L08: Infections of the skin and subcutaneous tissue (0.15)	DC_{30}^m	B25-B34: Other viral diseases (1.00) J00-J06: Acute upper respiratory infections (0.13) H80-H83: Diseases of inner ear (0.13)
DC_{14}^m	D10-D36: Benign neoplasms (1.00) L80-L99: Other disorders of the skin and subcutaneous tissue (0.26) L60-L75: Disorders of skin appendages (0.18)	DC_{31}^m	K90-K95: Other diseases of the digestive system (1.00) B00-B09: Viral infections characterized by skin and mucous membrane lesions (0.56) K20-K31: Diseases of oesophagus, stomach and duodenum (0.23)
DC_{15}^m	K20-K31: Diseases of oesophagus, stomach and duodenum (1.00) J40-J47: Chronic lower respiratory diseases (0.94) J30-J39: Other diseases of upper respiratory tract (0.92)	DC_{32}^m	H10-H13: Disorders of conjunctiva (1.00) H00-H06: Disorders of eyelid, lacrimal system and orbit (0.93) H55-H59: Other disorders of eye and adnexa (0.80)
DC_{16}^m	K40-K46: Hernia (1.00) K20-K31: Diseases of oesophagus, stomach and duodenum (0.16) J40-J47: Chronic lower respiratory diseases (0.09)	DC_{33}^m	H25-H28: Disorders of lens (1.00) H40-H42: Glaucoma (0.50) H15-H22: Disorders of sclera, cornea, iris and ciliary body (0.13)
DC_{17}^m	A00-A09: Intestinal infectious diseases (1.00) G40-G47: Episodic and paroxysmal disorders (0.11) K55-K64: Other diseases of intestines (0.09)	DC_{34}^m	I70-I79: Diseases of arteries, arterioles and capillaries (1.00) J40-J47: Chronic lower respiratory diseases (0.10) L80-L99: Other disorders of the skin and subcutaneous tissue (0.08)

Table 3: Disease clusters for female patients. Top 3 diseases are shown for each cluster. The weights denote each disease's belonging to the cluster, where 1 denotes a strong presence and 0 denotes an absence.

DC_1^f	I30-I52: Other forms of heart disease (1.00) I20-I25: Ischaemic heart diseases (0.54) I60-I69: Cerebrovascular diseases (0.25)	DC_{18}^f	E00-E07: Disorders of thyroid gland (1.00) I10-I15: Hypertensive diseases (0.12) E70-E90: Metabolic disorders (0.08)
DC_2^f	A00-A09: Intestinal infectious diseases (1.00) G50-G59: Nerve, nerve root and plexus disorders (0.36) K20-K31: Diseases of oesophagus, stomach and duodenum (0.09)	DC_{19}^f	G30-G32: Other degenerative diseases of the nervous system (1.00) F10-F19: Mental and behavioural disorders due to psychoactive substance use (0.24) F40-F48: Neurotic, stress-related and somatoform disorders (0.22)
DC_3^f	H00-H06: Disorders of eyelid, lacrimal system and orbit (1.00) J40-J47: Chronic lower respiratory diseases (0.52) H10-H13: Disorders of conjunctiva (0.24)	DC_{20}^f	I80-I89: Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified (1.00) L00-L08: Infections of the skin and subcutaneous tissue (0.18) L80-L99: Other disorders of the skin and subcutaneous tissue (0.17)
DC_4^f	H55-H59: Other disorders of eye and adnexa (1.00) H10-H13: Disorders of conjunctiva (0.32) H15-H22: Disorders of sclera, cornea, iris and ciliary body (0.21)	DC_{21}^f	N80-N98: Noninflammatory disorders of female genital tract (1.00) D10-D36: Benign neoplasms (0.68) G50-G59: Nerve, nerve root and plexus disorders (0.41)
DC_5^f	N60-N65: Disorders of breast (1.00) C00-C97: Malignant neoplasms (0.16) B35-B49: Mycoses (0.15)	DC_{22}^f	M80-M94: Osteopathies and chondropathies (1.00) M30-M36: Systemic connective tissue disorders (0.12) K20-K31: Diseases of oesophagus, stomach and duodenum (0.12)
DC_6^f	K00-K14: Diseases of oral cavity, salivary glands and jaws (1.00) H10-H13: Disorders of conjunctiva (0.59) H53-H54: Visual disturbances and blindness (0.24)	DC_{23}^f	E20-E35: Disorders of other endocrine glands (1.00) H15-H22: Disorders of sclera, cornea, iris and ciliary body (0.72) L40-L45: Papulosquamous disorders (0.18)
DC_7^f	J09-J18: Influenza and pneumonia (1.00) J20-J22: Other acute lower respiratory infections (0.17) G40-G47: Episodic and paroxysmal disorders (0.16)	DC_{24}^f	H65-H75: Diseases of middle ear and mastoid (1.00) H90-H95: Other disorders of ear (0.39) H60-H62: Diseases of external ear (0.31)
DC_8^f	H80-H83: Diseases of inner ear (1.00) H90-H95: Other disorders of ear (0.16) H60-H62: Diseases of external ear (0.10)	DC_{25}^f	B85-B89: Pediculosis, acariasis and other infestations (1.00) B35-B49: Mycoses (0.26) L20-L30: Dermatitis and eczema (0.22)
DC_9^f	N30-N39: Other diseases of urinary system (1.00) E70-E90: Metabolic disorders (0.47) J20-J22: Other acute lower respiratory infections (0.38)	DC_{26}^f	E50-E64: Other nutritional deficiencies (1.00) D50-D53: Nutritional anaemias (0.21) L55-L59: Radiation-related disorders of the skin and subcutaneous tissue (0.11)
DC_{10}^f	D60-D64: Aplastic and other anaemias (1.00) D50-D53: Nutritional anaemias (0.66) K90-K95: Other diseases of the digestive system (0.35)	DC_{27}^f	K80-K87: Disorders of gallbladder, biliary tract and pancreas (1.00) K20-K31: Diseases of oesophagus, stomach and duodenum (0.30) K40-K46: Hernia (0.26)
DC_{11}^f	K50-K52: Noninfective enteritis and colitis (1.00) K55-K64: Other diseases of intestines (0.16) K90-K95: Other diseases of the digestive system (0.14)	DC_{28}^f	N17-N19: Renal failure (1.00) I10-I15: Hypertensive diseases (0.39) E08-E14: Diabetes mellitus (0.37)
DC_{12}^f	H25-H28: Disorders of lens (1.00) H53-H54: Visual disturbances and blindness (0.89) H30-H36: Disorders of choroid and retina (0.69)	DC_{29}^f	I70-I79: Diseases of arteries, arterioles and capillaries (1.00) M30-M36: Systemic connective tissue disorders (0.36) L40-L45: Papulosquamous disorders (0.20)
DC_{13}^f	F30-F39: Mood [affective] disorders (1.00) F40-F48: Neurotic, stress-related and somatoform disorders (0.68) G40-G47: Episodic and paroxysmal disorders (0.45)	DC_{30}^f	D70-D77: Other diseases of blood and blood-forming organs (1.00) D50-D53: Nutritional anaemias (0.24) K20-K31: Diseases of oesophagus, stomach and duodenum (0.10)
DC_{14}^f	E65-E68: Obesity and other hyperalimentation (1.00) E08-E14: Diabetes mellitus (0.41) I10-I15: Hypertensive diseases (0.21)	DC_{31}^f	B00-B09: Viral infections characterized by skin and mucous membrane lesions (1.00) L55-L59: Radiation-related disorders of the skin and subcutaneous tissue (0.29) K20-K31: Diseases of oesophagus, stomach and duodenum (0.19)
DC_{15}^f	N10-N16: Renal tubulo-interstitial diseases (1.00) N20-N23: Urolithiasis (0.25) N30-N39: Other diseases of urinary system (0.12)	DC_{32}^f	H49-H52: Disorders of ocular muscles, binocular movement, accommodation and refraction (1.00) C00-C97: Malignant neoplasms (0.32) G20-G26: Extrapyramidal and movement disorders (0.20)
DC_{16}^f	L50-L54: Urticaria and erythema (1.00) L20-L30: Dermatitis and eczema (0.18) L60-L75: Disorders of skin appendages (0.15)	DC_{33}^f	N70-N77: Inflammatory diseases of female pelvic organs (1.00) B35-B49: Mycoses (0.17) N80-N98: Noninflammatory disorders of female genital tract (0.17)
DC_{17}^f	B95-B98: Bacterial, viral and other infectious agents (1.00) J30-J39: Other diseases of upper respiratory tract (0.23) B35-B49: Mycoses (0.21)	DC_{34}^f	B25-B34: Other viral diseases (1.00) H60-H62: Diseases of external ear (0.35) J30-J39: Other diseases of upper respiratory tract (0.33)

Lastly, to demonstrate how the time-courses that results from our model can be used for the analysis of disease trajectories, we carried out an ascendancy analysis, where each DC is treated as a node and the binarised time-courses for each disease pair are used to calculate their corresponding κ and τ. Connecting the pairs with high κ values, and defining the direction of the resulting edges based on their corresponding τ, a network map can be derived that is shown in Figures 7 and 8, for male and female patients, respectively. In both these figures, we chose a threshold for κ, so that the graph has 60 edges to simplify the visual investigation. Note that for every cluster we show the top 3 diseases (using a bar plot showing their weights). We can see for instance in the network corresponding to male patients (Figure 7 [1]) that DC_{20}^{m} in which the top disease is "Diabetes mellitus" leads to DC_{24}^{m} in which the top disease is "melabolic disorders". We can also see in the network corresponding to female patients (Figure 8 [2]) that DC_{27}^{f} in which the top disease is "Disorders of gallbladder, biliary tract and pancreas" leads to DC_{18}^{f} in which the top disease is "Disorders of thyroid gland". The association (κ) and ascendency (τ) scores between all disease clusters for male and female patients are shown in Appendix in Figure A.3.

Note that disease clusters for male and female patients are not necessarily the same as these have been computed separately for male and female patients to account for the fact that certain health and wellbeing issues are either gender-specific, or more commonly associated with one gender. Nevertheless, our results show DCs emerging in one gender that have similar counterparts in the other gender (i.e., diseases with highest weights in a male DCs are the same as or similar to those in a female DC). For instance, DC_{1}^{m} and DC_{7}^{f} correspond to clusters of lower respiratory diseases; DC_{3}^{m} and DC_{24}^{f} correspond to diseases of the middle ear and mastoid and upper respiratory diseases; DC_{6}^{m} and DC_{20}^{f} correspond to vein diseases and related disorders; DC_{11}^{m} and DC_{13}^{f} correspond to mood disorder clusters; DC_{27}^{m} and DC_{27}^{f} correspond to disorders of gallbladder, biliary tract, pancreas and related diseases. On the other hand, many DCs are specific to a single gender; for instance, DC_{2}^{m} (diseases of male genital organs and related diseases), and DC_{8}^{m} (behavioural syndromes associated with physiological disturbances and physical factors) are more specific to male patients, while DC_{5}^{f} (disorders of breast) and DC_{33}^{f} (inflammatory diseases of female pelvic organs) which are more specific to female patients.

[1] online version available here: `https://deepmedicine.github.io/JBI_NMF/circular_net_males.html`

[2] online version available at: `https://deepmedicine.github.io/JBI_NMF/circular_net_females.html`

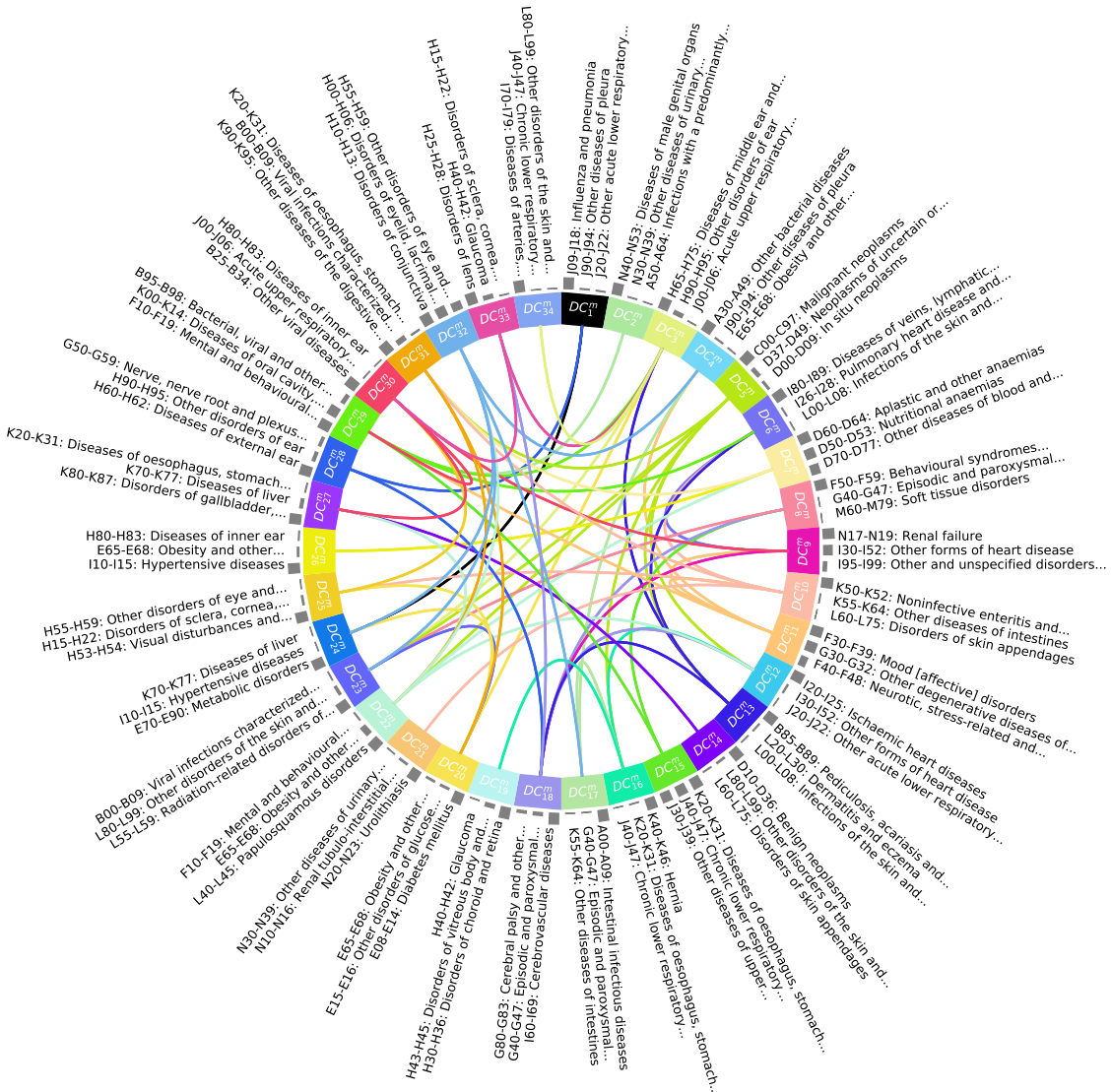


Figure 7: The ascendancy network of MPs for male patients. For visual clarity, we demonstrated the network corresponding to the top 60 edges (i.e., 0.534 threshold for κ). Note that edges are coloured with the colour of the node they originate from (i.e., the ascendant node).

Figure 8: The ascendancy network of MPs for female patients. For visual clarity, we demonstrated the network corresponding to the top 60 edges (i.e., 0.5327 threshold for κ). Note that edges are coloured with the colour of the node they originate from (i.e., the ascendant node).

4. Conclusions and Discussion

In this study, we employed a well-known matrix factorisation technique called NMF to mine the MPs using one of the largest EHR datasets in the world (i.e., CPRD). The key reason behind this research was to provide a simple and effective solution for multimorbidity research. To be more specific, our study attempted to build on the past studies' learnings, while addressing some of their limitations (e.g., using relatively small data, relying on a narrow observation windows, focusing on a small number of diseases, solely extracting the DCs (instead of temporal phenotyping), and lack of appropriate quantitative benchmarking and evaluation of the results). To the best of the authors' knowledge, our approach is the first one in the literature that has used a matrix factorisation technique for temporal phenotyping and the study of MPs' temporal patterns.

Enabling NMF to result in temporal phenotyping was achieved through a simple assumption, which is fairly common across similar studies: Every disease belongs to a number of DCs (with a degree of membership); a linear combination of these DCs for each patient at each year of his/her life explains the observed patterns of diagnoses. This is very similar to a common approach in the study of functional MRI data in neuroimaging, for instance, known as multi-session ICA through temporal concatenation [51]. In cases where one is looking for DCs that are common across population, without assuming them having a consistent expression across different people's lifetime, this approach can be even preferred over certain tensor-based factorisation approaches, which assumes the same cluster and time-course for everyone.

Given a number of DCs, the next important objective of multimorbidity analyses is to conclude a disease network, which summarises how diseases interact with one another and influence each other's occurrence. While network models have the potential to solve such a problem (when given the time-courses for diseases), the definition of nodes can be a challenge for researchers. That is, if we operate in the ICD-10 universe, we can see scenarios where the network can have anything ranging from 22 nodes (at chapter level) to 10,138 nodes (at level 4). We know from the network modelling literature that the search space for finding the best network is of a super-exponential size on the number of nodes (i.e., $O(n!2^{\binom{n}{2}})$) [52]. This makes the optimisation for learning a network of 10K nodes a huge challenge; both in terms of data (relatively small number of patients, and low prevalence and hence co-occurrence of most diseases at this level) and computing. On the other hand, using ICD-10 chapters, which will result in network with 22 nodes, is likely to lead to results that are hard to be clinically meaningful and interpretable; due to the heterogeneity of the diseases they each contain. Operating at levels such as ICD-10

blocks and 3-character ICD-10 codes, while not suffering from too many nodes, is still likely to have many highly correlated/co-occurring nodes that might make sense to be combined (particularly, given the data and computing challenges that we face) when learning large networks.

In this study, we introduced a new concept for the node (i.e., the DCs resulting from the NMF) and a new framework to mine these nodes' relationships with each other. This definition of nodes has a few advantages. Firstly, as our analysis suggests, it leads to a relatively small number of nodes, for which the corresponding network will be easier to learn. Secondly, from an empirical point of view, given that such nodes are driven by diseases that usually co-occur, splitting them into sub-nodes is not likely to be the source of any advantage (specially that such a split will make the network more complex and hence more difficult to learn). And lastly, from a clinical perspective, we are implying that diseases tend to happen in clusters and what the network will tell us is the influence of one cluster on another, given the rest of the clusters. This is in correspondence with what many in the clinical world have been arguing for that the definition of diseases today might not be the most accurate one (and hence various research on phenomapping of the diseases) [53]. Based on all these, we carried out our analysis and derived a network, using a simple and yet powerful technique known as ascendancy analysis. Of course, the data from our NMF approach is equally useful for any other network model (e.g., Bayesian networks), and hence there is need for follow-up research on the use of alternative network modelling techniques (an exhaustive list of such techniques have been used and compared in Smith et al [49]).

Given the importance of automatic evaluation of disease clusters and DC networks that result from analyses like ours, we introduced two new metrics in this study: C_L and A_L, to assess the goodness of DCs, and ascendancy scores between two DCs, respectively. While we showed some of the results on C_L for parameter tuning earlier, here we also show both A_L and C_L for all the disease pairs, using the optimal values of the free parameters. Note that, while ascendancy analysis does not necessarily mean causal relationship, it has empirically been shown to have high correspondence with it. For instance, Smith et al. [49] have shown how ascendancy metrics are better than almost all well-known causal modelling tools, when assessed on simulated data with known ground truth. Furthermore, disease trajectories and various diseases' relationship with each other are complex; there will be cases where different experts might not agree on whether a relationship is causal or comorbid. Therefore, showing A_L and C_L for all pairs will provide additional insight about the strengths and weaknesses of our approach. Figure 9 shows these results for a number of L values. For instance, at $L = 15$, which represents the top 17% of the diseases, the results

show that our analysis has successfully identified a considerable number of disease pairs in our lookups either as part of the same cluster ($> 33\%$), or as linked via ascendancy ($> 61\%$). Refer to Section 2.6 for more details on these metrics.

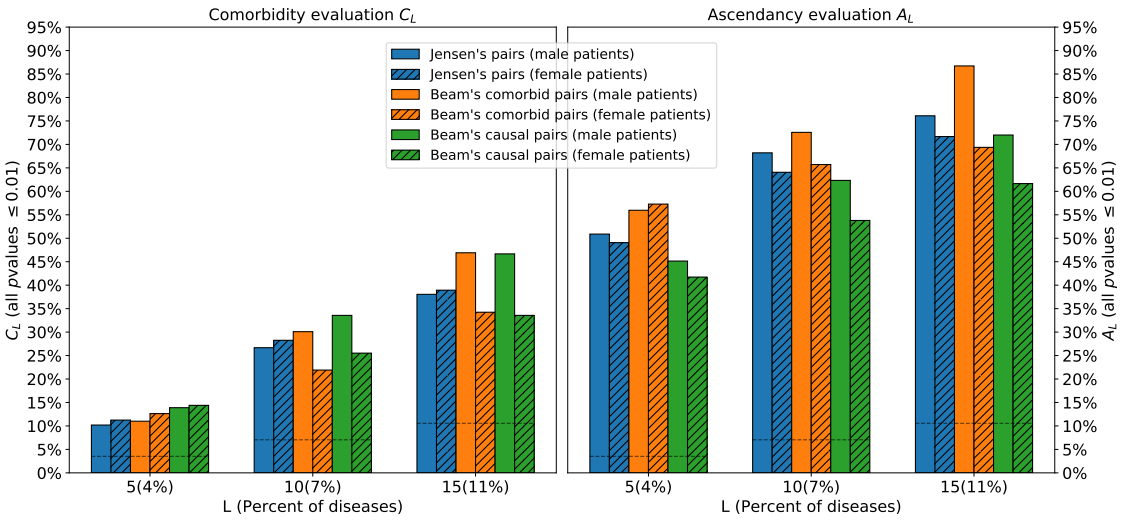


Figure 9: Evaluation of the comorbidity (left) and ascendancy (right). Note that, according to our non-parametric test of significance, all these values/bars have P-values less than 0.01. The horizontal lines correspond to the percents of diseases when considering 5, 10 and 15 top conditions. For instance, when considering the top 5 conditions in our DCs (out of 142, i.e. 3.5%), our analysis successfully identifies more than 10% of Jensen's disease pairs.

As in most such analyses, our modelling pipeline relied on a mix of choices and assumptions we made; from preprocessing to factorisation technique and beyond. For instance, NMF has a range of different implementations; a follow-up research on comparing different implementations of NMF can surely improve the approach we introduced. Furthermore, the data preprocessing we introduced, from smoothing to adjustment for count can all be done with some differences that can be subject of future research. Lastly, our approach was focused on $\mathbf{D}$ as defined in Section 2. The earlier works, however, have introduced more concepts (e.g., measurements and medications) to the starting matrix. Given that our approach can simply accommodate additional phenotypes to the input matrix, without necessarily needing to add a new dimension to it (not needing to go from matrix to tensor space) there is an opportunity that can lead to richer analyses that not only can take time/age into account, but also can benefit from additional concepts and interventions that do

influence the DCs and their trajectories. However, for the multimorbidity analysis, disease-only analysis was a necessary starting point.

It can be argued that many comorbidities happen at a more granular level than the ICD-10 block. For instance all malignant neoplasms are under the same ICD-10 block, however one would expect Hepatocellular carcinoma (C22.0) to cooccur with cirrhosis of liver (K74) or hepatitis B (B16) or hepatitis C (B17.1) but not hepatitis A (B15) nor hepatitis E (B17.2). Similarly, diabetes mellitus type 1 and type 2 have two different clinical courses and associated comorbidities. In order to have a closer look, we have identified patients with these conditions and computed the disease clusters in which they showed an expression (Table 4). Interestingly, although many of these conditions belong to the same ICD-10 block, the patients with these conditions show different expressions over different clusters. For instance, type 1 and type 2 diabetes male patients show both a high expression in DC^m_{20} (diabetes cluster) and DC^m_8 (physiological disturbances cluster), however type 1 diabetes patients are more likely to show a high expression in DC^m_{19} (choroid and retina disorders), DC^m_{30} (other viral infections) and DC^m_{34} (vascular diseases cluster), instead type 2 diabetes patients are more likely to show a high expression in DC^m_{26} (hypertension cluster), DC^m_{24} (metabolic disorders) and DC^m_{12} (Ischaemic heart diseases). Similarly, patients with the considered liver diseases show different expressions over different clusters. This confirms that our modelling is able to distinguish between granular diseases despite being performed at the ICD-10 block level. However, we believe that a more granular analysis is likely to unravel further relationships.

Note that the DCs resulting from the NMF analysis and the relationships among them are likely to have both true and false positives (as well as true and false negatives). Therefore, additional works are needed to further scrutinise such approaches' use for clinical decision making. For instance, in our approach (as well as all other factorisation-based approaches we found in the literature), disease clusters have weights for each and every disease; this means the researchers need to define a threshold to extract a subset of diseases that are truly co-morbid; the use of probabilistic NMF [54, 44] and is likely to alleviate the use of such subjective/arbitrary thresholds. Similarly, while our ascendency analysis resulted in some associations which are in correspondence with medical knowledge (based on expert evaluation, as well as the metrics we introduced in this paper), there are examples among them that are harder to explain; for instance, rare or unknown DCs such as DC^m_{25} (visual disturbance and blindness) leading to DC^m_{30} (other viral diseases), or DC^m_{28} (diseases of

Table 4: Top disease clusters for male (♂) and female (♀) patients with specific conditions. For reference, the diseases with the highest weigths in each of these clusters are appended.

	E10: Type 1 diabetes mellitus				**E11: Type 2 diabetes mellitus**			
	♂:5174		♀:4110		♂:76913		♀:65532	
Top 5 clusters	DC_{20}^{m}	74%	DC_{14}^{f}	21%	DC_{20}^{m}	71%	DC_{28}^{f}	23%
	DC_{8}^{m}	20%	DC_{28}^{f}	17%	DC_{8}^{m}	21%	DC_{14}^{f}	20%
	DC_{19}^{m}	14%	DC_{7}^{f}	12%	DC_{26}^{m}	12%	DC_{7}^{f}	12%
	DC_{10}^{m}	10%	DC_{18}^{f}	11%	DC_{24}^{m}	12%	DC_{18}^{f}	12%
	DC_{34}^{m}	10%	DC_{29}^{f}	10%	DC_{12}^{m}	11%	DC_{30}^{f}	10%

	K74: Cirrhosis of liver				**C22.0: Liver cell carcinoma**			
	♂:2166		♀:2755		♂:616		♀:210	
Top 5 clusters	DC_{27}^{m}	20%	DC_{27}^{f}	19%	DC_{20}^{m}	23%	DC_{27}^{f}	22%
	DC_{31}^{m}	18%	DC_{20}^{f}	15%	DC_{31}^{m}	21%	DC_{20}^{f}	16%
	DC_{20}^{m}	14%	DC_{7}^{f}	14%	DC_{5}^{m}	19%	DC_{30}^{f}	14%
	DC_{8}^{m}	12%	DC_{30}^{f}	13%	DC_{8}^{m}	16%	DC_{7}^{f}	14%
	DC_{10}^{m}	12%	DC_{22}^{f}	13%	DC_{27}^{m}	12%	DC_{10}^{f}	14%

	B15: Acute hepatitis A				**B16: Acute hepatitis B**			
	♂:550		♀:508		♂:101		♀:52	
Top 5 clusters	DC_{8}^{m}	13%	DC_{7}^{f}	18%	DC_{29}^{m}	17%	DC_{5}^{f}	12%
	DC_{30}^{m}	10%	DC_{27}^{f}	14%	DC_{6}^{m}	10%	DC_{21}^{f}	12%
	DC_{27}^{m}	9%	DC_{5}^{f}	13%	DC_{20}^{m}	8%	DC_{20}^{f}	10%
	DC_{10}^{m}	9%	DC_{34}^{f}	12%	DC_{2}^{m}	8%	DC_{31}^{f}	10%
	DC_{2}^{m}	9%	DC_{31}^{f}	10%	DC_{32}^{m}	8%	DC_{27}^{f}	10%

	B171: Acute hepatitis C				**B172: Acute hepatitis E**			
	♂:177		♀:73		♂:33		♀:11	
Top 5 clusters	DC_{29}^{m}	19%	DC_{27}^{f}	14%	DC_{20}^{m}	21%	DC_{29}^{f}	27%
	DC_{11}^{m}	14%	DC_{7}^{f}	12%	DC_{8}^{m}	15%	DC_{5}^{f}	18%
	DC_{27}^{m}	12%	DC_{6}^{f}	12%	DC_{12}^{m}	12%	DC_{27}^{f}	18%
	DC_{8}^{m}	10%	DC_{19}^{f}	12%	DC_{9}^{m}	12%	DC_{26}^{f}	18%
	DC_{31}^{m}	9%	DC_{25}^{f}	12%	DC_{26}^{m}	12%	DC_{31}^{f}	18%

DC_{2}^{m}	Diseases of male genital organs
DC_{5}^{m}	Malignant neoplasms
DC_{6}^{m}	Diseases of veins, lymphatic vessels and lymph...
DC_{8}^{m}	Behavioural syndromes associated with...
DC_{9}^{m}	Renal failure
DC_{10}^{m}	Noninfective enteritis and colitis
DC_{11}^{m}	Mood [affective] disorders
DC_{12}^{m}	Ischaemic heart diseases
DC_{19}^{m}	Disorders of choroid and retina
DC_{20}^{m}	Diabetes mellitus
DC_{24}^{m}	Metabolic disorders
DC_{26}^{m}	Hypertensive diseases
DC_{27}^{m}	Disorders of gallbladder, biliary tract and...
DC_{29}^{m}	Mental and behavioural disorders due to...
DC_{30}^{m}	Other viral diseases
DC_{31}^{m}	Other diseases of the digestive system
DC_{32}^{m}	Disorders of conjunctiva
DC_{34}^{m}	Diseases of arteries, arterioles and capillaries

DC_{5}^{f}	Disorders of breast
DC_{6}^{f}	Diseases of oral cavity, salivary glands and jaws
DC_{7}^{f}	Influenza and pneumonia
DC_{10}^{f}	Aplastic and other anaemias
DC_{14}^{f}	Obesity and other hyperalimentation
DC_{18}^{f}	Disorders of thyroid gland
DC_{19}^{f}	Other degenerative diseases of the nervous system
DC_{20}^{f}	Diseases of veins, lymphatic vessels and lymph...
DC_{21}^{f}	Noninflammatory disorders of female genital tract
DC_{22}^{f}	Osteopathies and chondropathies
DC_{25}^{f}	Pediculosis, acariasis and other infestations
DC_{26}^{f}	Other nutritional deficiencies
DC_{27}^{f}	Disorders of gallbladder, biliary tract and...
DC_{28}^{f}	Renal failure
DC_{29}^{f}	Diseases of arteries, arterioles and capillaries
DC_{30}^{f}	Other diseases of blood and blood-forming organs
DC_{31}^{f}	Viral infections characterized by skin and mucous...
DC_{34}^{f}	Other viral diseases

external ear) leading to DC_{18}^{m} (cerebrovascular diseases). Therefore, further follow-up analyses are needed in order to determine whether such associations are truly meaningful, or they are hypotheses that are not supported by medical evidence.

Lastly, there have been various developments in methods related to our study that can provide multiple new directions for future works. For instance, deep learning's success in the past few years has led to "deep phenotyping" research on EHR; while such models can help the study of MPs, their use has been limited to learning disease representations (or embeddings) for disease/event predictions. The earlier works in this space, despite not taking time into account [55, 56], have shown that meaningful DCs can be learned from EHR. Similar results have been shown using CNN [57] and RNN [58, 59, 60]. Furthermore, there have been multiple neural solutions for matrix factorisation [61, 62]. Another novel methodology that has the potential to improve such research is Temporal Regularised Matrix Factorisation (TRMF) [63], which has the ability to regularise the temporal aspect of the factors so it is influenced by prior knowledge/assumptions such as smoothness.

Acknowledgements

This research was funded by the Oxford Martin School (OMS) and supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the OMS, the UK National Health Service (NHS), the NIHR or the Department of Health and Social Care. This work uses data provided by patients and collected by the NHS as part of their care and support and would not have been possible without access to this data. The NIHR recognises and values the role of patient data, securely accessed and stored, both in underpinning and leading to improvements in research and care.

Appendix A. Supplementary figures

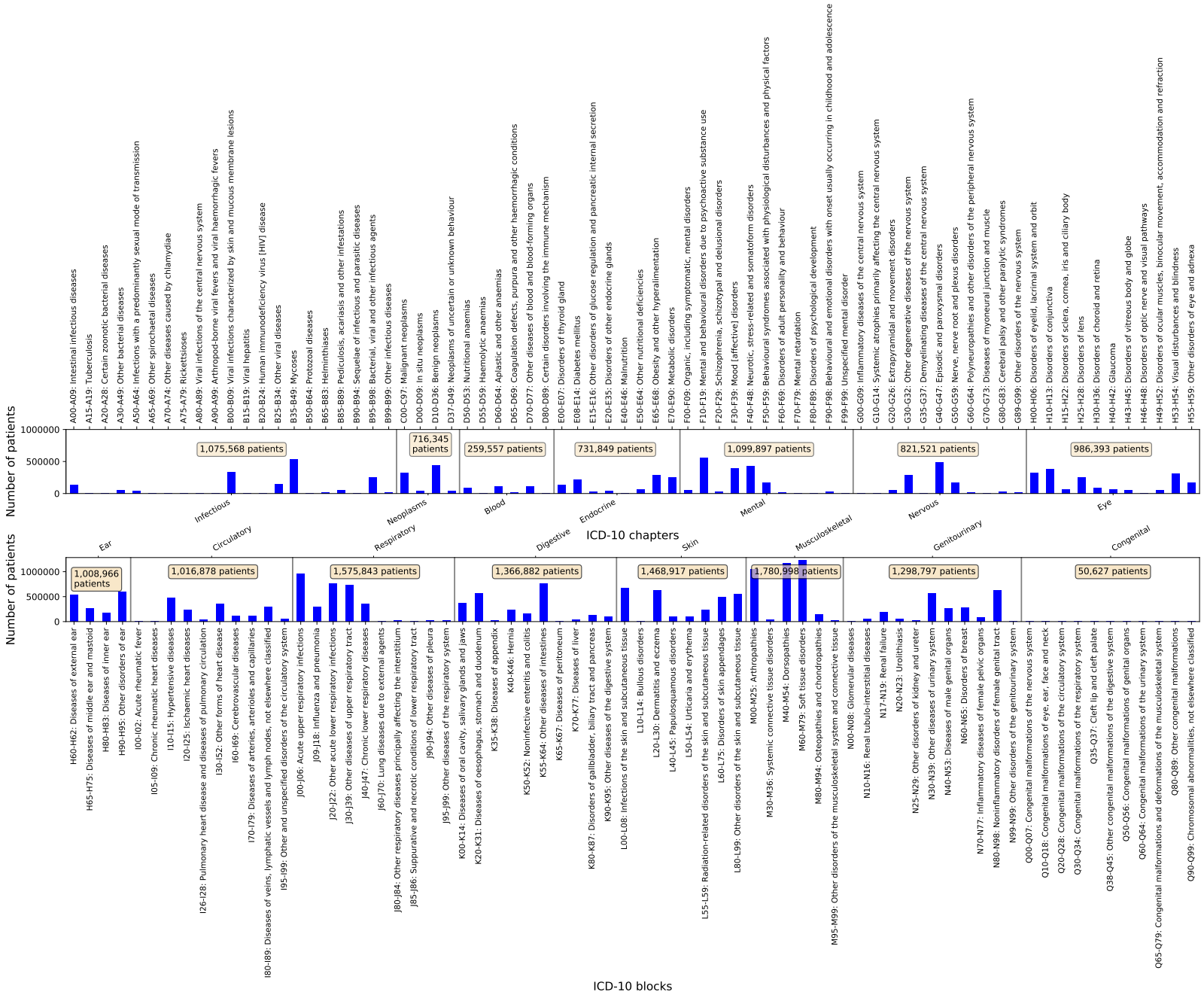


Figure A.1: Number of unique patients per ICD-10 block in our data cut.

Figure A.2 shows the heatmap in which each column corresponds to DCs that result from NMF (i.e., matrix $\mathbf{B}$), an online interactive version of this figure is available at `https://deepmedicine.github.io/JBI_NMF/ICD_block_clusters.html`. Figure A.3 shows association measures (κ) and ascendancy measures (τ) between each pair of DCs. High values of κ indicate that the diseases are associated. Positive values of τ indicate an ascendancy relationship in a given direction, whereas negative values indicate an ascendancy relationship in the opposite direction.

References

[1] M. Van den Akker, F. Buntinx, J. F. Metsemakers, S. Roos, J. A. Knottnerus, Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases, Journal of clinical epidemiology 51 (5) (1998) 367–375.

[2] J. Tran, R. Norton, N. Conrad, F. Rahimian, D. Canoy, M. Nazarzadeh, K. Rahimi, Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the uk between 2000 and 2014: A population-based cohort study, PLoS medicine 15 (3) (2018) e1002513.

[3] T. A. of Medical Sciences, Multimorbidity: a priority for global health research, `https://acmedsci.ac.uk/policy/policy-projects/multimorbidity` (2018).

[4] M. Goldacre, L. Kurina, D. Yeates, V. Seagroatt, L. Gill, Use of large medical databases to study associations between diseases, Qjm 93 (10) (2000) 669–675.

[5] C. A. Hidalgo, N. Blumm, A.-L. Barabási, N. A. Christakis, A dynamic network approach for the study of human phenotypes, PLoS computational biology 5 (4) (2009) e1000353.

[6] A. B. Jensen, P. L. Moseley, T. I. Oprea, S. G. Ellesøe, R. Eriksson, H. Schmock, P. B. Jensen, L. J. Jensen, S. Brunak, Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients, Nature communications 5 (2014) 4022.

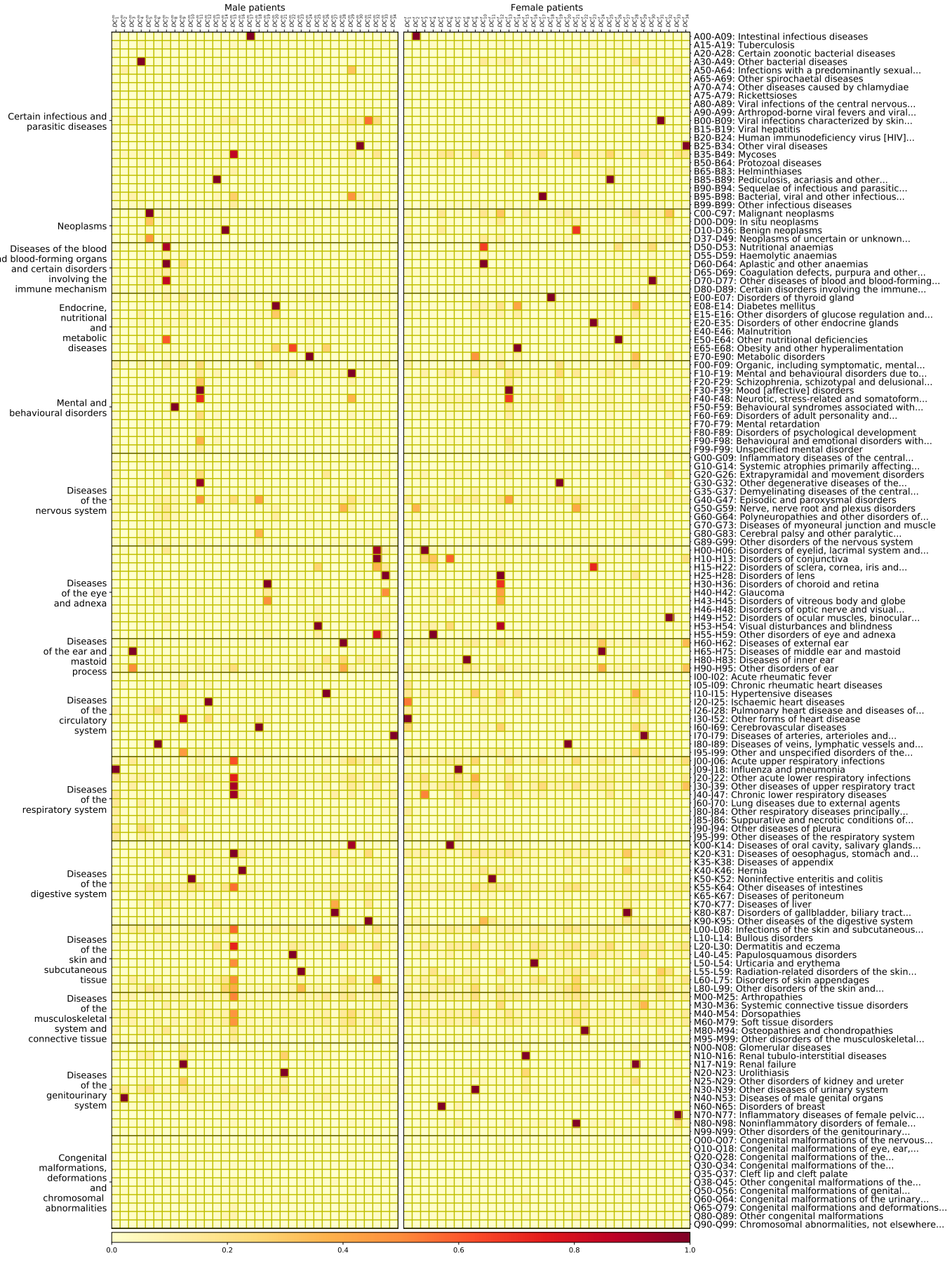
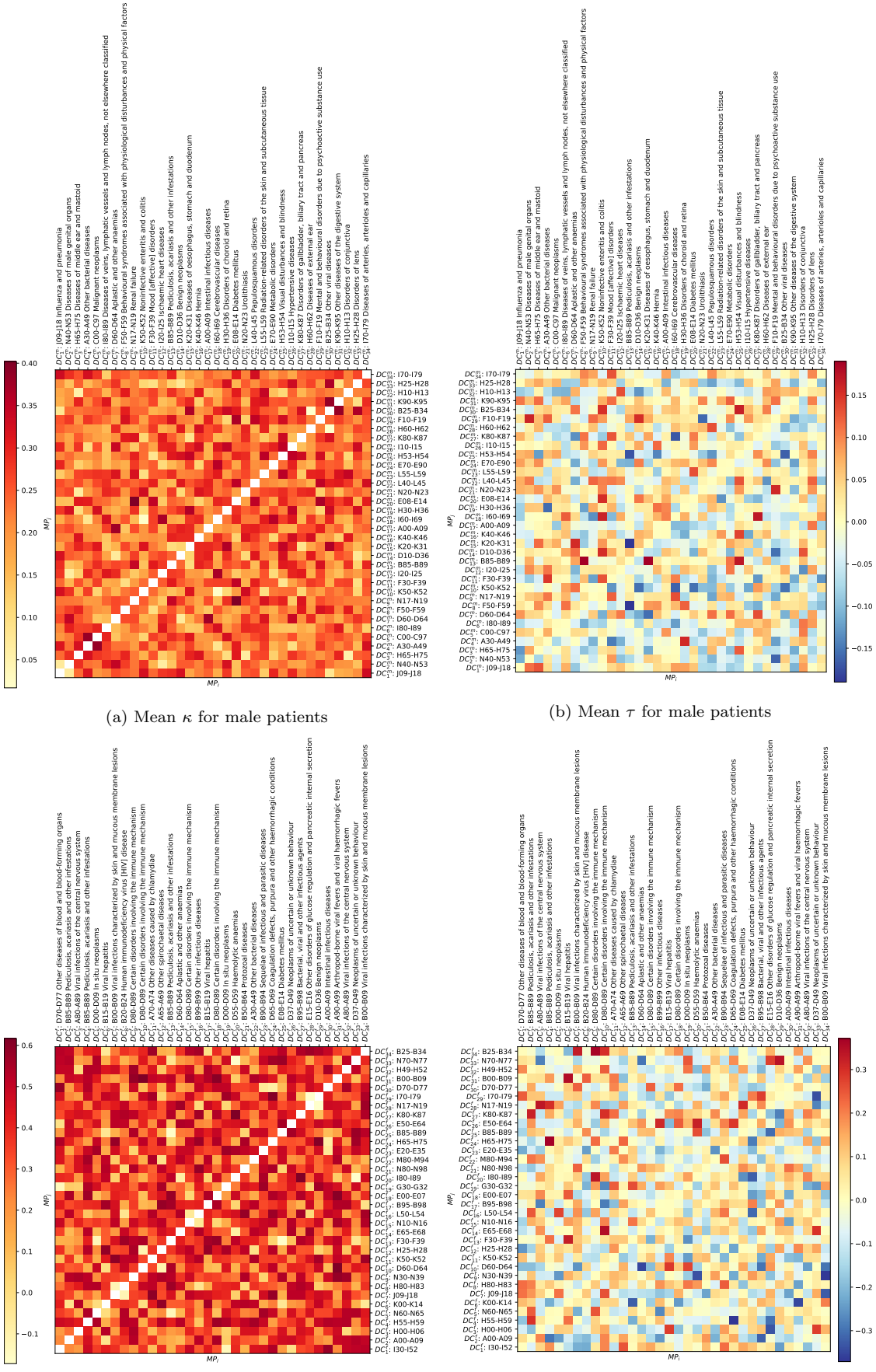


Figure A.2: Disease clusters for male and female patients (on the left and right sides, respectively). The figure shows the transposed version ($\mathbf{B}^T$) of $\mathbf{B}$ matrices, after gamma correction (so that small values are visible). An interactive version of this figure is available at `https://deepmedicine.github.io/JBI_NMF/ICD_block_clusters.html`



(a) Mean κ for male patients

(b) Mean τ for male patients

(c) Mean κ for female patients

(d) Mean τ for female patients

Figure A.3: Mean κ and τ for all pairs of MP components for both male and female patients. Each row/column corresponds to a DC, while the label shows the DC index plus the ICD-10 block with the strongest weight in that cluster. Note that, κ is symmetric while τ is asymmetric (for visual clarity, κ values along the diagonal are shown in white to make other values easier to observe).

[7] V. Y. Strauss, P. W. Jones, U. T. Kadam, K. P. Jordan, Distinct trajectories of multimorbidity in primary care were identified using latent class growth analysis, Journal of clinical epidemiology 67 (10) (2014) 1163–1171.

[8] C. Liu, F. Wang, J. Hu, H. Xiong, Temporal phenotyping from longitudinal electronic health records: A graph based framework, in: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2015, pp. 705–714.

[9] J. Pearl, Causality: Models, Reasoning and Interference, Cambridge University Press, 2009.

[10] X. Wang, D. Sontag, F. Wang, Unsupervised learning of disease progression models, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014, pp. 85–94.

[11] L. Holden, P. A. Scuffham, M. F. Hilton, A. Muspratt, S.-K. Ng, H. A. Whiteford, Patterns of multimorbidity in working australians, Population health metrics 9 (1) (2011) 15.

[12] I. Schäfer, E.-C. von Leitner, G. Schön, D. Koller, H. Hansen, T. Kolonko, H. Kaduszkiewicz, K. Wegscheider, G. Glaeske, H. van den Bussche, Multimorbidity patterns in the elderly: a new approach of disease clustering identifies complex interrelations between chronic conditions, PloS one 5 (12) (2010) e15941.

[13] A. Marengoni, D. Rizzuto, H.-X. Wang, B. Winblad, L. Fratiglioni, Patterns of chronic multimorbidity in the elderly population, Journal of the American Geriatrics Society 57 (2) (2009) 225–230.

[14] I. Kirchberger, C. Meisinger, M. Heier, A.-K. Zimmermann, B. Thorand, C. S. Autenrieth, A. Peters, K.-H. Ladwig, A. Döring, Patterns of multimorbidity in the aged population. results from the kora-age study, PloS one 7 (1) (2012) e30556.

[15] A. Roso-Llorach, C. Violán, Q. Foguet-Boreu, T. Rodriguez-Blanco, M. Pons-Vigués, E. Pujol-Ribera, J. M. Valderas, Comparative analysis of methods for identifying multimorbidity patterns: a study of 'real-world' data, BMJ open 8 (3) (2018) e018986.

[16] J. Zhou, F. Wang, J. Hu, J. Ye, From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014, pp. 135–144.

[17] J. C. Ho, J. Ghosh, J. Sun, Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014, pp. 115–124.

[18] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, J. Sun, Rubik: Knowledge guided tensor factorization and completion for health data analytics, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1265–1274.

[19] I. Perros, E. E. Papalexakis, R. Vuduc, E. Searles, J. Sun, Temporal phenotyping of medically complex children via parafac2 tensor factorization, Journal of biomedical informatics 93 (2019) 103125.

[20] J. Zhao, Y. Zhang, D. J. Schlueter, P. Wu, V. E. Kerchberger, S. T. Rosenbloom, Q. S. Wells, Q. Feng, J. C. Denny, W.-Q. Wei, Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study, Journal of biomedical informatics 98 (2019) 103270.

[21] A. Cichocki, A.-H. Phan, Fast local algorithms for large scale nonnegative matrix and tensor factorizations, IEICE transactions on fundamentals of electronics, communications and computer sciences 92 (3) (2009) 708–721.

[22] C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the β-divergence, Neural computation 23 (9) (2011) 2421–2456.

[23] Clinical practice research datalink, available at: https://www.cprd.com/ (2015).

[24] E. Herrett, A. M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. van Staa, L. Smeeth, Data resource profile: clinical practice research datalink (cprd), International journal of epidemiology 44 (3) (2015) 827–836.

[25] T. Walley, A. Mantgani, The uk general practice research database, The Lancet 350 (9084) (1997) 1097–1099.

[26] C. A. Emdin, S. G. Anderson, T. Callender, N. Conrad, G. Salimi-Khorshidi, H. Mohseni, M. Woodward, K. Rahimi, Usual blood pressure, peripheral arterial disease, and vascular risk: cohort study of 4.2 million adults, Bmj 351 (2015) h4865.

[27] C. A. Emdin, S. G. Anderson, G. Salimi-Khorshidi, M. Woodward, S. MacMahon, T. Dwyer, K. Rahimi, Usual blood pressure, atrial fibrillation and vascular risk: evidence from 4.3 million adults, International journal of epidemiology 46 (1) (2016) 162–172.

[28] L. Smeeth, S. L. Thomas, A. J. Hall, R. Hubbard, P. Farrington, P. Vallance, Risk of myocardial infarction and stroke after acute infection or vaccination, New England Journal of Medicine 351 (25) (2004) 2611–2618.

[29] F. Lee, H. Patel, M. Emberton, The 'top 10'urological procedures: a study of hospital episodes statistics 1998–99, BJU international 90 (1) (2002) 1–6.

[30] H. Mohseni, A. Kiran, R. Khorshidi, K. Rahimi, Influenza vaccination and risk of hospitalization in patients with heart failure: a self-controlled case series study, European heart journal 38 (5) (2017) 326–333.

[31] WHO International Classification of Diseases, ICD-10 version 2016, available at: https://icd.who.int/browse10/2016/en (2016).

[32] NHS-Digital, Read codes, Available at: https://digital.nhs.uk/services/terminology-and-classifications/read-codes.

[33] NHS-Digital, Read-ICD10 cross map, `https://nhs-digital.citizenspace.com/uktc/crossmaps/`.

[34] NHS-Digital, Snomed codes, `https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct`.

[35] NHS-Digital, Read v2 to SNOMED CT Mapping Lookup (October 2018), `https://hscic.kahootz.com/connect.ti/t_c_home/view?objectId=407588`.

[36] N. L. of Medicine, Snomed ct to icd-10-cm map, https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html.

[37] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: Advances in neural information processing systems, 2001, pp. 556–562.

[38] S. Z. Li, X. Hou, H. Zhang, Q. Cheng, Learning spatially localized, parts-based representation, CVPR (1) 207 (2001) 212.

[39] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788.

[40] Y. W. Y. Jia, C. H. M. Turk, Fisher non-negative matrix factorization for learning local features, in: Proc. Asian conf. on comp. vision, Citeseer, 2004, pp. 27–30.

[41] J.-P. Brunet, P. Tamayo, T. R. Golub, J. P. Mesirov, Metagenes and molecular pattern discovery using matrix factorization, Proceedings of the national academy of sciences 101 (12) (2004) 4164–4169.

[42] Z. Zhang, T. Li, C. Ding, X. Zhang, Binary matrix factorization with applications, in: Seventh IEEE International Conference on Data Mining (ICDM 2007), IEEE, 2007, pp. 391–400.

[43] M. Arngren, M. N. Schmidt, J. Larsen, Bayesian nonnegative matrix factorization with volume prior for unmixing of hyperspectral images, in: 2009 IEEE International Workshop on Machine Learning for Signal Processing, IEEE, 2009, pp. 1–6.

[44] M. Zitnik, B. Zupan, Nimfa: A python library for nonnegative matrix factorization, Journal of Machine Learning Research 13 (2012) 849–853.

[45] M. Tepper, G. Sapiro, Compressed nonnegative matrix factorization is fast and accurate, IEEE Transactions on Signal Processing 64 (9) (2016) 2269–2283.

[46] M. Kapralov, V. Potluru, D. Woodruff, How to fake multiply by a gaussian matrix, in: International Conference on Machine Learning, 2016, pp. 2101–2110.

[47] A. Rajaraman, J. D. Ullman, Mining of massive datasets, Cambridge University Press, 2011.

[48] R. S. Patel, F. D. Bowman, J. K. Rilling, A bayesian approach to determining connectivity of the human brain, Human brain mapping 27 (3) (2006) 267–276.

[49] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, M. W. Woolrich, Network modelling methods for FMRI, Neuroimage 54 (2) (2011) 875–891.

[50] A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. P. Palmer, X. Shi, T. Cai, I. S. Kohane, Clinical concept embeddings learned from massive sources of multimodal medical data (2018). arXiv:1804.01486.

[51] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, S. M. Smith, Fsl, Neuroimage 62 (2) (2012) 782–790.

[52] R. W. Robinson, Counting labeled acyclic digraphs, new directions in the theory of graphs (proc. third ann arbor conf., univ. michigan, ann arbor, mich., 1971) (1973).

[53] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, A.-L. Barabási, The human disease network, Proceedings of the National Academy of Sciences 104 (21) (2007) 8685–8690.

[54] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, S. H. Jensen, Theorems on positive data: On the uniqueness of nmf, Computational intelligence and neuroscience 2008 (2008).

[55] T. Tran, T. D. Nguyen, D. Phung, S. Venkatesh, Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm), Journal of biomedical informatics 54 (2015) 96–105.

[56] R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Scientific reports 6 (2016) 26094.

[57] P. Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh, Deepr: A convolutional net for medical records, IEEE Journal of Biomedical and Health Informatics 21 (1) (2017) 22–30. doi:10.1109/JBHI.2016.2633963.

[58] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, in: Advances in Neural Information Processing Systems, 2016, pp. 3504–3512.

[59] M. Rafiq, G. Keel, P. Mazzocato, J. Spaak, C. Savage, C. Guttmann, Deep learning architectures for vector representations of patients and exploring predictors of 30-day hospital readmissions in patients with multiple chronic conditions, in: International Workshop on Artificial Intelligence in Health, Springer, 2018, pp. 228–244.

[60] C. Xiao, T. Ma, A. B. Dieng, D. M. Blei, F. Wang, Readmission prediction via deep contextual embedding of clinical concepts, PloS one 13 (4) (2018) e0195024.

[61] O. Levy, Y. Goldberg, Neural word embedding as implicit matrix factorization, in: Advances in neural information processing systems, 2014, pp. 2177–2185.

[62] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, B. Ramabhadran, Low-rank matrix factorization for deep neural network training with high-dimensional output targets, in: 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, 2013, pp. 6655–6659.

[63] H.-F. Yu, N. Rao, I. S. Dhillon, Temporal regularized matrix factorization for high-dimensional time series prediction, in: Advances in neural information processing systems, 2016, pp. 847–855.