Co-morbidity Representation in Artificial Intelligence: Tapping into Unused Clinical Knowledge

William J. Bolton, Pantelis Georgiou, Alison Holmes, and Timothy M. Rawson

Abstract Co-morbidities or long-term medical conditions are an essential piece of clinical information that can drastically alter a patient's treatment, management and outcome. However, such data is difficult to apply to artificial intelligence (AI) systems due to its heterogeneity, sparsity and combinatorial complexity. This text proposes a novel pipeline to address such pitfalls by utilizing the structure of Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), the most extensive clinical vocabulary in the world, to create disease and co-morbid patient embeddings which can be used in downstream AI applications. We demonstrate that our methodologies outperform existing approaches in classification and similar patient retrieval tasks. This research highlights that appropriate extraction and representation of clinical knowledge through innovative approaches can enable AI systems to advance personalized healthcare.

Keywords Comorbidity representation · Clinical knowledge representation · Artificial intelligence in healthcare · SNOMED CT and patient embeddings · Addressing complexities with healthcare data in AI

W. J. Bolton (✉)
Department of Computing, AI for Healthcare CTD, Centre for Antimicrobial Optimisation, NIHR HPRU in HAI and AMR, Imperial College London, London, UK
e-mail: william.bolton@imperial.ac.uk

P. Georgiou
Centre for Bio-inspired Technology, Centre for Antimicrobial Optimisation, NIHR HPRU in HAI and AMR, Imperial College London, London, UK
e-mail: pantelis@imperial.ac.uk

A. Holmes
Faculty of Health Life Sciences, University of Liverpool, UK. Centre for Antimicrobial Optimisation, NIHR HPRU in HAI and AMR, Department of Infectious Diseases, Imperial College London, London, UK
e-mail: alison.holmes@imperial.ac.uk

T. M. Rawson
Centre for Antimicrobial Optimisation, NIHR HPRU in HAI and AMR, Imperial College London, London, UK
e-mail: timothy.rawson07@imperial.ac.uk

A. Shaban-Nejad et al. (eds.), *AI for Health Equity and Fairness*, Studies in Computational Intelligence 1164, https://doi.org/10.1007/978-3-031-63592-2_14

1 Introduction

As the capabilities of artificial intelligence (AI) systems advance the significance of the data used for training and inference, as well as its representation, becomes increasingly important. This is especially apparent in healthcare, where numerous key challenges exist including data quality, heterogeneity and biases [1, 2]. Developing methods to effectively extract and integrate data into healthcare AI applications is thus of paramount importance. Particularly for those medical domains that exhibit significant data complexity and have an unmet clinical need. This is true for co-morbidities defined in this text as chronic long-term medical conditions which are a major challenge in healthcare [3, 4]. They can drastically alter a patient's treatment plan, are associated with increased mortality, place a huge financial strain on healthcare services and their prevalence is increasing due to the ageing population [5–7]. Co-morbidities are therefore a fundamental part of illustrating a patient in both a clinical and AI setting. However, challenges exist with representing and using co-morbidity data in AI systems. The large number of unique diseases, their sparsity, missingness and frequent confounding means learning useful co-morbidity relationships even through powerful machine or deep learning methods is difficult. This combined with limited dataset sizes, particularly for those with rare diseases or complex co-morbidity combinations can hinder models generalizability. Furthermore, heterogeneity in how medical information is recorded across different healthcare systems and institutions introduces additional complexities when training and deploying AI models. Developing methods and pipelines to mitigate these problems and appropriately represent co-morbidities is therefore of great importance to maximise the utility of healthcare focused AI systems.

In this text, we propose a novel approach of utilizing the underlying knowledge and structure contained within a healthcare ontology to generate disease and patient embeddings, and demonstrate their utility in downstream AI applications. Specifically, Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) the worlds most comprehensive clinical healthcare terminology [8] is used. Originally developed by the College of American Pathologists and the National Health Service (NHS), SNOMED CT is used extensively around the world in electronic health records (EHRs) to represent clinically relevant information. In England, all NHS healthcare providers are now mandated to use SNOMED CT for documentation [9]. It enables standardized record keeping and for information to be seamlessly exchanged between systems which ultimately improves patient care. SNOMED CT contains a wealth of known clinical information about how diseases are related and interact, by utilizing this publicly available expert curated knowledge graph we can create embeddings that contain implicit information on the relationship between diseases. This overcomes some of the previously stated challenges with co-morbidity data that exist when only training models on disease data from EHRs. For example, the embeddings are not influenced by dataset size, the number of diseases or their rareness and they are adaptable to variation in clinical documentation when different SNOMED disorder codes are used to record the same chronic condition (e.g.,

Diabetes vs Diabetes Type 1 or 2), as these diseases will be closely located in the embedding space. Furthermore, by exploiting a well adopted common clinical coding language our approach is inherently generalizable, and we mitigate bias concerns [10] and integration issues which are common in healthcare due to systems lacking interoperability. Note that the terms co-morbidity, chronic condition and disease are used synonymously in this text, as this research currently focuses on long-term conditions and not acute diagnoses. By addressing the problem of co-morbidity representation in AI systems, we envisage that our approach could be useful to AI applications that use case-based reasoning or decision support to provide personalized care.

2 Related Work

Knowledge representation in healthcare is a broad topic that covers the formalization and organization of medical information for AI systems [4]. Most relevant to this work is research using ontologies and graphs. For terminologies such as SNOMED CT, AI techniques including natural language processing have previously been utilized to automatic identify diseases from free text [11]. However, as far as we are aware, no work has focused on investigating patients long-term health conditions through SNOMED CTs architecture. Graphs are a way of structuring data using nodes and edges [12]. A number of methods have been developed to embed graphical information into a latent space. Node2vec [13], subgraph2vec [14] and graph2vec [15] all convert high-dimensional graph structured data into low-dimensional continuous vector spaces, while trying to maintain graphical structure. Such methods have previously been applied to medical claim and co-occurrence data [16]. Knowledge graphs are a type of graph that contain entities as nodes and their relationships as edges. In healthcare many knowledge graphs such as SNOMED CT exist, which are created by pooling a broad spectrum of data including expert consensus, EHRs, guidelines, and published literature [17]. For example, PrimeKG [18], is a large multimodal knowledge graph that contains connections from diseases to a wide range of medically relevant factors, including drugs, phenotypes and genotypes. AI research on such data is diverse, ranging from drug design to clinical decision support [17]. Pertinent to this work, recent research looked into creating embeddings from knowledge graphs with the aim of understanding drug-drug interactions [19, 20].

When focusing on co-morbidities a wide variety of research exists. The Charlson co-morbidity index [21, 22] is a widely adopted tool that quantifies the burden of chronic conditions on patients and predicts mortality risk through weighted scores. It assigns some specific co-morbidities to 17 different categories and was developed through retrospective analysis and expert consensus. More recently methods such as clustering [23, 24] or finite mixture models [25] have been applied to EHRs to elucidate groups of patients. Subsequent analysis then aims to improve disease understanding and management [26], for example by preventing adverse events. Such approaches are generally disease specific rather than focusing on predictions for a

given patient. More relevant to this research is AI based clinical decision support systems that use co-morbidity information alongside other data to make clinically useful, patient specific predictions. Often disease data is fed into such models via one hot encodings [27]. While modern machine and deep learning methods are sophisticated and can have a substantial number of parameters, they still suffer from the combinatorial explosion of many diseases and a lack of data on those with rare diseases or a high degree of multi-morbidity (two or more long-term health conditions). Case-based reasoning is an alternative approach to clinical decision support where historical past examples are collated to help solve current problems. It has seen great use in healthcare [28], but ensuring patients recalled are similar to the current patient in question with regards to their chronic conditions is essential. Simple approaches have defined patient similarity based on their overlapping diagnosis codes [29] and the Jaccard index [30] when encoded using the one hot methodology. Rocheteau et al. on the other hand approached the challenges associated with patients diagnoses data by utilizing clinical reasoning to define a patient similarity score and subsequently generating a patient graph for downstream AI applications [31]. All of these patient similarity approaches, however, require exact diseases matches between patients, and hence will suffer from coding quality issues and miss similar patients with slightly different diagnoses. As such, they are not generalizable and are sensitive to data heterogeneity. Existing AI research on co-morbid patients therefore lacks appropriate representation. By incorporating external knowledge and creating meaningful embeddings from medically grounded data, we can help overcome the problems of sparsity and hetrogeneity, enabling powerful deep learning methods to be utilized and personalized healthcare to become a reality. The main contributions of this research include a novel patient pair similarity metric, a pipeline to generate disease and co-morbid patient embeddings from the previously untapped fundamental structure of a common clinical coding system (SNOMED CT), and demonstration of the utility of such embeddings to healthcare AI applications.

3 Methods and Tasks

3.1 Generating SNOMED Disease Embeddings

EHR data from Imperial College Healthcare NHS Trusts (ICHT) Clinical Analytics, Research and Evaluation (iCARE) environment was used in this research. iCARE encompasses de-identified real-world clinical data from those with inpatient encounter at ICHT. In particular the 'problem' table in iCARE was used as it contains data on patients chronic diseases (i.e., their co-morbidities) and their date of diagnosis. SNOMED CT International Edition (version released on 31/07/2022) was downloaded from the NHS Digital Technology Reference Update Distribution. The structure of SNOMED CT is highly complex [32]. A high level overview is provided in the appendix, so the methodologies employed in this research can be understood.

SNOMED CT was filtered for concepts classified as 'disorders' and relationships of type 'is a', to capture all diseases, their relationships to each other, and to remove all other information. The 'SNOMED distance' was defined and calculated for each unique disease pair as the shortest path length between the two concepts in the SNOMED CT hierarchy. Patients and their historic co-morbidity diagnoses were then extracted from ICHT EHR and the SNOMED data was filtered to only those co-morbidities present in the EHR. This filtering was applied as it ensures the disease embeddings are relevant to those patients in the EHR, and from experimentation returned significantly better embeddings than simply applying node2vec directly to the whole hierarchical structure of SNOMED CT. Because diagnoses within the EHR were only used as a filter our approach is adaptable to any number of diagnosis and different healthcare institutions. Based on the filtered SNOMED data a connected undirected graph was created where each co-morbidity in the EHR dataset with > 10 edges with a SNOMED distance of ≤ 2 was a node and a weighted edge existed between each node if their SNOMED distance was < 6. These preprocessing steps were hyperparameters tuned to optimize the metric discussed below and to prevent diseases with few strongly weighted edges connected to them from skewing the resulting embeddings. Edges were weighted (w) based on the following formula $w = \frac{1}{s_{d_i,d_j} \cdot f_s}$ where s_{d_i,d_j} is the SNOMED distance between two diseases (d_i and d_j) and f_s is the number of edges with that SNOMED distance. This ensured that random walks were more likely to traverse edges with smaller SNOMED distances compared to the more common longer range connections. Node2vec [13] was then applied to this graph to generate embeddings for each disease, the graphical structure meaning the embeddings were learnt based on the fundamental clinical knowledge and architecture of SNOMED CT. The mean SNOMED distance between each node and their nearest neighbor was utilized as a metric for hyperparameter optimization. Minimizing this metric enables us to optimize node2vec as we aim to create embeddings where by those diseases nearby in SNOMED CT are close in the embedding space. Therefore, preserving the graphical structure of SNOMED and creating meaningful disease embeddings. Optuna was used for hyperparameter optimization [33]. The final hyperparameters for node2vec were an embedding dimension of size 128, p = 0.1, q = 6.0, window size = 10, 10 walks per node and a walk length of 1000.

3.2 Supervised Learning Task

SNOMED disease embeddings can be utilized as feature inputs for machine learning and deep learning models, which when trained in a supervised way can make clinically useful predictions. Given a particular patient may have any number of co-morbidities, and that their relationship is unordered, a set transformer [34] was chosen as an appropriate permutation invariant attention-based neural network model. We undertook two classification prediction tasks, firstly for in-hospital mortality within a year (assuming no loss to follow-up) and secondly for a long in-patient length of

stay (LOS), defined as >7 days. For both tasks only patients' historical co-morbidity information was used as features, no information from a patient's stay or any acute conditions were included to ensure we directly evaluated the utility of our generated embeddings. Such predictive models could be useful for planning and preventative medicine before a patient enters a hospital. Area Under the Receiver Operating Characteristic curve (AUROC) results were generated for the whole population and those without the 75% most common co-morbidities using 10 fold stratified cross-validation. Stratification preserves the percentage of samples with each label across dataset folds, while cross-validation ensures a robust performance estimation. We compared our method to logistic regression with patients' co-moribities encoded using the traditional one hot methodology, logistic regression based on patients' Charlson co-morbidity categories, and a set transformer with random SNOMED disease embeddings. Details on our models, baselines and statistical analysis can be found in the appendix.

3.3 Similar Patient Retrieval Task

Co-morbid patient embeddings were intuitively defined as the mean of all the SNOMED disease embeddings for that particular individual. This then allows for similar patients to be retrieved through simple nearest neighbor lookup [35] based on euclidian distance. We compared our method to nearest neighbor lookup with traditional one hot encodings, and the patient similarity score from Rocheteau et al. [31] (details in the appendix). A similar patient retrieval task was defined as finding the most comparable patient within the EHR for any given unique co-morbid patient, where no identical match was possible. This tests the methods under difficult circumstances where similarity is a more meaningful and important clinical question, as opposed to simple cases when similarity is more trivial. This task was run for 1000 randomly selected patients due to time and computational restrictions within the iCARE environment. Determining if patients retrieved by similarity methods are indeed comparable is challenging [36]. To evaluate this task we propose a novel metric called 'SNOMED similarity'. To determine SNOMED similarity for a particular patient pair ($p1$, $p2$) their co-morbidities and respective SNOMED distances are utilized to create a distance matrix ($S_{p1,p2}$). Specifically within $S_{p1,p2}$ the ith and jth entry is $s_{d_i^{p1},d_j^{p2}}$, i.e., the SNOMED distance between $p1's$ ith disease and $p2's$ jth disease. We then match each disease of $p1$ to the diseases of $p2$ so that the matching minimized this equation $f(A) = \sum_{i=1}^{n} \min_{j\in\{1,\ldots,m\}}(1 - \frac{1}{A_{ij}+1})$ where $A \in \mathbb{R}^{n\times m}$. The same operation was applied to $p2$, with the resulting numbers summed to give the final SNOMED similarity score: $\text{sim}_{p1,p2} = f(S_{p1,p2}) + f(S_{p2,p1})$. This methodology ensures identical diseases return 0 while more unalike co-morbidities add to the score proportionally. Furthermore, by applying the operation to each individual, a mismatch in the number of co-morbidities will serve as a penalty. As such, patients with similar diseases will return low SNOMED similarity scores while

those that are dissimilar will return higher scores. Given the potential bias of using SNOMED distance in our pipeline and assessment, we defined the Charlson Jaccard index as another evaluation metric. Here we determined the presence of 17 Charlson co-morbidity categories [21, 22] for each patient through the Unified Medical Language System (UMLS) metathesaurus International Classification of Diseases (ICD) to SNOMED map [37]. The Jaccard index between the categories for the particular patient in question and the similar patient identified by each method was then calculated, where a higher score indicates a more similar match. By grouping multiple diagnostic codes into categories, this approach overcomes the exact disease match limitation of a straightforward Jaccard index on the one hot encodings. However, a drawback is that the Charlson co-morbidity index [21, 22] does not cover all co-morbidities and hence some information is discarded. Additional analysis was conducted to understand how robust the methods are to variation in the degree of multi-morbidity and the rarity of patients conditions.

To complement the automatic nature of the SNOMED similarity score and Charlson Jaccard index we employed expert human evaluation for a random subset of these unique co-morbid patient examples. More specifically a 10 question multiple-choice survey was created (Appendix Table 2) and 20 clinicians recruited. They were requested to select the patient (A, B or C) they believed was the most similar to the patient in the question with regards to their co-morbidities. Each answer represented a patient identified as most similar by the three methodologies, allowing us to evaluate which method clinicians perceived as the most appropriate in those scenarios. In two cases (question 4 and 10) one hot encodings and Rocheteau's similarity score returned the same patient and so these questions were presented as a binary choice to avoid duplication. Additional details on the survey are provided in the appendix.

4 Results and Discussion

4.1 SNOMED Disease Embeddings

2,133 diseases from 95,157 co-morbid patients were extracted from the EHR. The most prevalent disease was hypertension followed by asthma, varicella (historical diagnosis) and diabetes mellitus type 2. Figure 1 shows the frequency distribution for diseases and a histogram of the number of patient co-morbidities. Both show an exponentially decreasing trend, highlighting that combinatorial complexity makes using co-morbidity information in AI difficult. An advantage of our approach of generating disease embeddings through the SNOMED ontology is that EHR data is not required beyond filtering for diseases of interest, this overcomes limitations frequently seen with such data including a lack of examples on rare diseases and highly multi-morbid patients. When generating SNOMED disease embeddings node2vec optimization returned a mean SNOMED distance between each node and their nearest neighbor of 1.23. Analysis of the t-SNE visualisation (Fig. 2) with disease groups

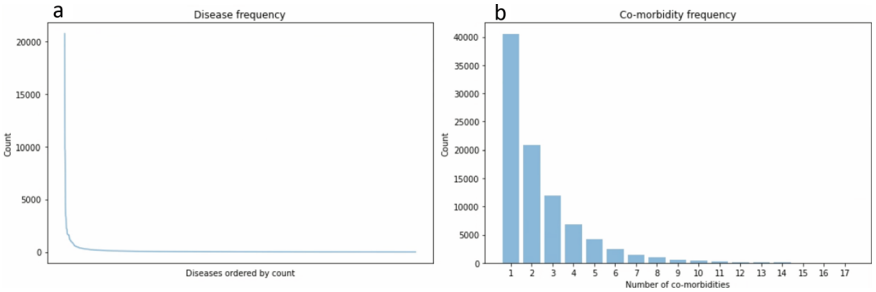

Fig. 1 **a** The frequency distribution for chronic diseases in the EHR data. **b** Histogram displaying the number of patient co-morbidities. The exponential decay on both graphs indicates the complexity of using co-morbidities in AI. Note to preserve privacy for both plots only those with counts greater than 5 were included

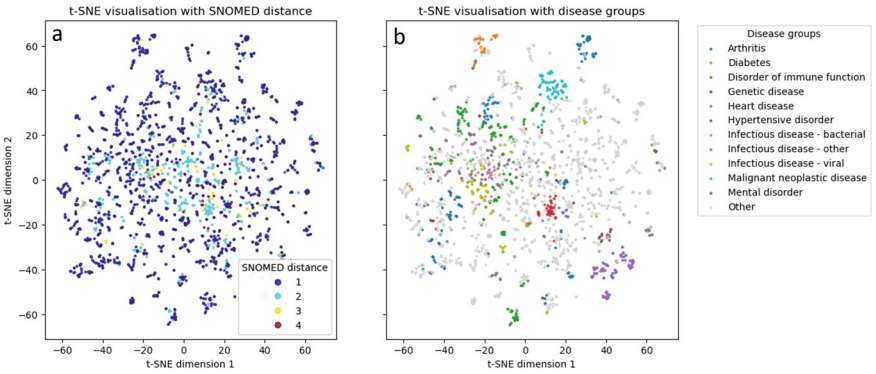

Fig. 2 t-SNE visualisations of the SNOMED disease embeddings with (**a**) the nearest neighbor SNOMED distance displayed and (**b**) high-level disease groups highlighted. Most diseases are embedded next to their neighbours in SNOMED CT, with related diseases groups clustered together

highlighted demonstrates related conditions are located close in the embedding space. For example, viral and bacterial infectious diseases are clustered together and heart diseases are adjacent to hypertensive disorders.

4.2 Supervised Learning Task

Results for the supervised learning tasks are shown in Table 1. Our method of utilizing SNOMED disease embeddings obtained the highest mean AUROC in all scenarios. The difference in performance between the SNOMED disease embeddings and the random embeddings on the rarest co-morbidities, demonstrates that the positional information of our embeddings is useful and enables the model to effectively approximate those co-morbidity instances with a low frequency in the

Table 1 Mean unseen test set AUROC results for classification tasks in different populations

Features	Year mortality		Long length of stay	
	Overall	Rarest co-morbidities	Overall	Rarest co-morbidities
Charlson co-moribdity categories	0.65 (SD 0.01)	0.50 (SD <0.01)	0.60 (SD 0.01)	0.50 (SD 0.03)
One hot encodings	0.79 (SD 0.02)	0.80 (SD 0.23)	0.72 (SD 0.01)	0.55 (SD 0.11)
Random SNOMED disease embeddings	0.80 (SD 0.03)	0.56 (SD 0.33)	0.74 (SD 0.02)	0.52 (SD 0.23)
SNOMED disease embeddings	**0.82 (SD 0.02)**	**0.85 (SD 0.14)**	**0.75 (SD 0.01)**	**0.61 (SD 0.20)**

training data. The random embeddings performed reasonably well when evaluated on the overall dataset indicating that the set transformer can learn useful relationships between diseases and a label even if the information within each item in the set is uninformative. Results from the one hot encodings were good considering the simplicity of the approach. However its behaviour would likely not scale with increasing number of diseases (SNOMED CT contains >70,000 'disorders' [38]). Charlson co-morbidity categories contain less nuanced disease information and so as expected achieved the lowest AUROC results, with random performance on those with rare co-morbidities that don't fall into one of the 17 Charlson co-morbidity categories [21, 22]. Statistically significant differences (all p-values <0.03) were observed between all methods on the long LOS prediction task for the overall population using the Wilcoxon Signed-rank test, while for mortality a paired t-test only indicated statistical differences between the Charlson co-morbidity categories and all other methods (all p-values <0.01), and the SNOMED disease embeddings and one hot encodings (p-value <0.01). The relationship between co-morbidities and death is firmly established and hence the superior performance on the mortality prediction task throughout is expected. Length of stay prediction is a much more nuanced task [39] and would likely require additional data regarding the patient's stay to see further improvements. Both supervised tasks were challenging with a large label imbalance. The year mortality rate was 2.53% and only 5.80% of patients had a long LOS. Overall, these results show our SNOMED disease embedding-set transformer method is generalizable and has utility for clinically relevant predictions.

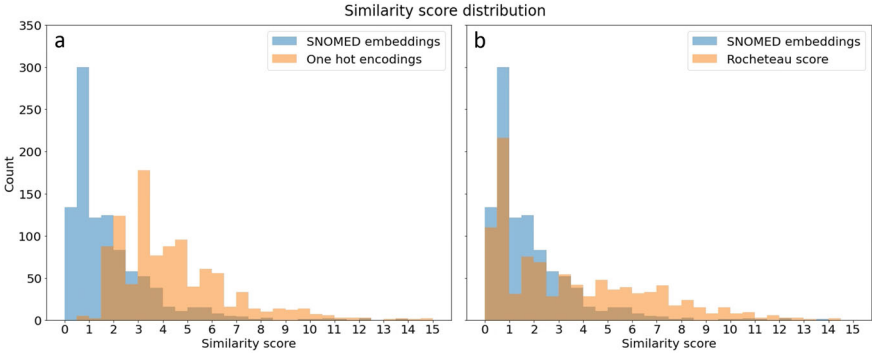

Fig. 3 Distribution of SNOMED similarity scores from the similar patient retrieval task. **a** Compares our co-morbid patient embeddings method to one hot encodings while **b** compares it to Rocheteau's score. Co-morbid patient embeddings returns patients with smaller scores

4.3 Similar Patient Retrieval Task

Using co-morbid patient embeddings to retrieve similar examples resulted in a mean SNOMED similarity score of 1.78 (SD 1.90) and Charlson Jaccard index of 0.84 (SD 0.34). One hot encodings obtained a mean score and index of 4.40 (SD 2.32) and 0.88 (SD 0.30) respectively, while Rocheteau's method achieved 3.52 (SD 3.26) and 0.69 (SD 0.20) respectively. Statistically significant results (all p-values <0.01) from the Wilcoxon Signed-rank test (alpha = 0.05) indicated that for both metrics all the methods come from different populations. In addition, our method was over 10^4 times faster (averaged 0.01s) at finding the most closely matched patient versus using Rocheteau's score or one hot encodings. The distribution of SNOMED similarity scores is shown in Fig. 3. Our method returned more similar patients and shows a narrower distribution. Results based on patients with different numbers of co-morbdities and the rarity of their co-morbidities are shown in the Appendix (Fig. 5). Co-morbid patient embeddings demonstrates consistent performance across patients with different degrees of rarity and multi-morbidity. In particular, our method is robust as the patients number of co-morbidities increases, while one hot encodings see a linear increase. For the mean Charlson Jaccard index metric the co-morbid patient embeddings and one hot encodings show superior coincident behaviour compared to Rocheteau's method. Note that this metric is not suitable for those with only very rare co-morbidites (right side of Fig. 5d) as the lack of diseases which fit into the Charlson categories causes scores to trend towards 1.

For expert human evaluation, the distribution of responses for each question in the multi-morbid patient similarity survey is shown in Fig. 4. Patients identified by our method were selected as the most similar in 60% of questions, and received the majority of the overall votes (53%). In the remaining 4 questions those patients identified by one hot encodings were preferred. For the 6 questions the co-morbid patient embedding method won, it received on average 68.33% (SD 16.93) of the

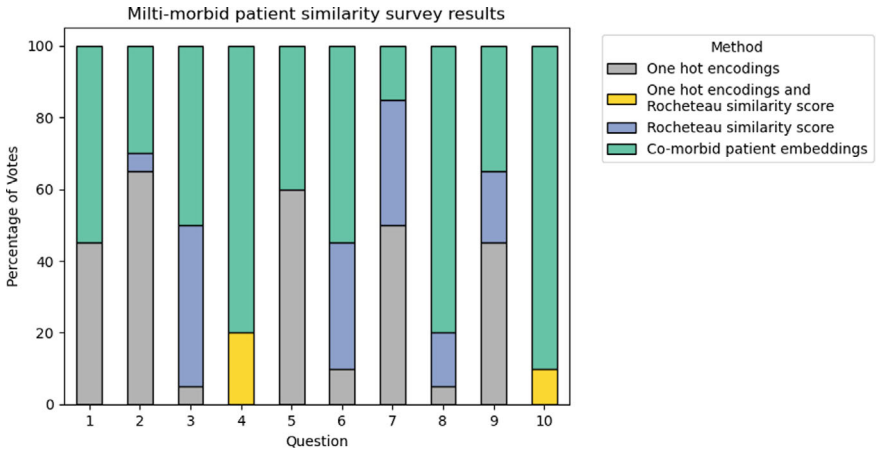

Fig. 4 Stacked bar plot showing the proportion of votes for patients identified by each method, for each question, in the survey. Co-morbid patient embeddings obtained the most votes for 6 questions with one hot encodings prevailing in the others

votes, and had a mean winning margin of 40% (SD 32.09). One hot encodings on the other hand were victorious by a smaller margin of 20% (SD 10.80) with less of the votes (mean 55%, SD 9.13). This method saw a highly variable number of votes across questions. For 4 questions it received $\leq$10% of the votes. Our method utilizing co-morbid patient embeddings on the other hand never got $\leq$15%. This consistency correlates well with the distribution of SNOMED similarity scores shown in Fig. 3. Co-morbid patient embeddings was selected by the majority of participants in cases ranging in complexity from 2 co-morbidities (question 10) to highly multi-morbid patients with 8 diseases (question 6). In the two binary choice questions where one hot encodings and Rocheteau's similarity score returned the same patient (question 4 and 10), co-morbid patient embeddings performed particularly well obtaining 80 and 90% of the votes respectively. Additional analysis on the variation in survey responses and clinicians free text answers is in the appendix. The results from the automatic SNOMED similarity score and expert human evaluators are aligned and demonstrate the ability of our method to identify analogous co-morbid patients.

Three patient similarity cases from the survey are shown in Appendix Fig. 6. Question 8 serves as a great example of the utility of our method as it is able to discern that 'Pregnancy-induced hypertension' is comparable to 'Hypertensive disorder' especially given the wider context of the patient in question, who had a number of pregnancy related co-morbidities. This enables a highly comparable patient to be returned, while the other methods exact matching approach means nuances are missed and hence compromises between co-morbidities must be made. For question 9 our method retrieves a patient whose main clinical difference is a lack of hypothyroidism. The most similar patient obtained using Rocheteau's method also has many overlapping conditions but these are alongside other diseases. For example, they have

rheumatoid arthritis which is associated with a different treatment to osteoarthritis and their other additional co-morbidities places them at a higher risk of other conditions, such as cardiovascular disease. The more simplistic matching of the one hot encoding method obtained the most of votes for this question, highlighting variability in clinicians opinions and expectations. Finally in Question 10 our method is able to find an almost identically matched patient, despite different coding terms being used. However, the similarity between alcoholism and alcohol dependence is missed by other methods, and as such a more chronically unwell patient is identified. These examples illustrate the advantage of co-morbid patient embeddings at accounting for slight distinctions between patients diseases. We believe our similar patient retrieval methodology could be useful to healthcare case-based reasoning technologies, by enabling more nuanced and relevant patients to be identified, particularly for cases with significant multi-morbidity or rarity.

4.4 Limitations and Future Work

The main limitation of this study is that we did not include additional patient information in our tasks. This was done deliberately to test how informative co-morbidity information can be through our methodological approach, rather than to create the best prediction model. However, as mentioned by the clinicians, many other important factors must be considered when undertaking clinical decision making, as well as, developing and evaluating AI systems in healthcare. Given the generalizable nature of our methodology, future work will look to address this by embedding additional clinically relevant information such as demographics and medications, while ensuring fairness. Another limitation of this work is that both our embedding method and SNOMED similarity score evaluation metric for the similar patient retrieval task use SNOMED distance, introducing a potential bias. This metric was developed and used as traditional approaches towards evaluating if patients are comparable such as the Jaccard similarity, require exact diseases matches which as discussed is not appropriate. To account for this we conducted additional evaluation with the Charlson Jaccard index and expert human evaluators on a subset.

5 Conclusion

In conclusion, we have developed a novel pipeline to extract and utilize the information contained within SNOMED CT to represent co-morbid patients and support healthcare focused AI systems. Such medical knowledge was previously untapped within the AI research community. We have demonstrated its utility in classification and similar patient retrieval tasks with automatic and human evaluation techniques. Our approach is able to overcome many of the problems associated with using disease data in AI, such as rarity and heterogeneity through creating standardized, informa-

tive, and clinically grounded embeddings, that allow for generalization across both patients and their diseases. We hope that this research can support AI applications that involve co-morbidity data, and ultimately advance such systems towards clinical benefit.

Competing Interests The authors have no conflicts of interest to declare that are relevant to the content of this chapter.

Acknowledgements William Bolton was supported by the UKRI center for doctoral training (CDT) in AI for Healthcare http://ai4health.io (Grant No. P/S023283/1). The authors would also like to acknowledge (1) the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infection and Antimicrobial Resistance (HAI and AMR) at Imperial College London. (2) The Department for Health and Social Care funded Centre for Antimicrobial Optimisation (CAMO) at Imperial College London. This study is independent research partly funded by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research the Department of Health and Social Care or the UK Health Security Agency. (3) The iCARE team, this research was enabled by the Imperial Clinical Analytics Research and Evaluation (iCARE) environment and used the iCARE team and data resources.

6 Appendix

6.1 SNOMED CT Overview

'Concepts' in SNOMED CT represent a specific piece of clinical information. Concepts are categorized into groups, including 'body structures', 'disorders', 'procedures' and 'pharmaceutical products'. All concepts are organized into a hierarchy where they are connected by 'relationships'. Relationships are also classified into groups by their 'type'. Relationships of type 'is a' are used to define the main hierarchy, for example, 'Diabetes mellitus is a disorder of glucose metabolism'. 'Attribute' relationships on the other hand, allow for more nuanced connections to be made as needed, such as between a disease and a side effect. In this research we focused on utilizing the main hierarchical structure of SNOMED CT ('is a' relationships) for those concepts that could be used to document patients co-morbidities ('disorders').

6.2 Additional Supervised Learning Task Details

Logistic regression was implemented with sklearn [40], and the 'class_weight' parameter set to balanced. This model was chosen as a baseline because it was the best performing sklearn model on the one hot encoded data. Note that passing this data format into a set transformer is computationally intractable due to the matrix size. The presence of 17 Charlson co-morbidity categories [21, 22] was determined for each patient using the YMLS metathesaurus ICD to SNOMED map [37].

Random SNOMED disease embeddings were created using NumPy's [41] random number generator, and constrained to within the maximum and minimum values of the original embeddings. While the information within the embedding was randomized, each disease was still associated with a particular embedding. This removes the disease location and relationship information and enables the capability of our generated embeddings to be evaluated. Our set transformer architecture included a single 'Induced Set Attention Block' followed by a 'Pooling by Multihead Attention' operation [34] and finally a simple linear decoder layer. It was created using Pytorch [42] and had 4 heads with a hidden dimension of size 160. Hyperparameter optimisation of the set transformer is tractable but was not conducted due to computational limits within the iCARE environment and because the objective of these tasks is to show our embeddings can be informative features rather than to create the best prediction model. For training a batch size of 512, the Adam optimiser [43] and binary cross entropy with logits loss were utilized. Statistical analysis used an alpha of 0.05 with a paired t-test or the Wilcoxon Signed-rank test to determine if groups came from the same population, depending on the normality of the data distribution, which was determined using the Shapiro-Wilk test.

6.3 Additional Similar Patient Retrieval Task Details

The similar patient retrieval task used Rocheteau et al's patient similarity metric [31] as one of the benchmarks. In that research they create a diagnosis matrix $D \in \mathbb{R}^{N \times m}$ where N is the number of patients and m is the amount of unique diagnosis. The similarity score M_{kl} between patients k and l is then defined as $M_{kl} = a \sum_{\mu=1}^{m} (D_{k\mu} D_{l\mu} (d_{\mu}^{-1} + c)) - \sum_{\mu=1}^{m} (D_{k\mu} + D_{l\mu})$ where d_{μ} is the frequency of diagnosis μ, and a and c are hyperparameters. Through grid search, c = 0.5, and a = 5 returned the best results on our dataset. This method rewards shared diseases, in-particular those that are rare, and penalises higher numbers of co-morbidities. The most closely matched patient is identified through having the greatest score.

For the survey, questions were presented in a plain text format, no information on the methods was included, and their positioning was changed for each question so as not to introduce any biases (Appendix Figure. 9). Clinicians were recruited through email and additional optional questions were presented after each patient similarity question and at the end of the survey, to understand how they made their decisions, their definition of patient similarity and their medical speciality. Beyond analysis of the votes and free text responses, the Chi-Square test of independence (alpha = 0.05) was used to determine if the distribution of votes was significantly different across medical specialities. Furthermore, additional investigations were conducted to understand the variability in choices through the Gini index, K-modes clustering [44] and multiple correspondence analysis [45] as a dimensionality reduction technique for plotting (Figs. 5 and 6).

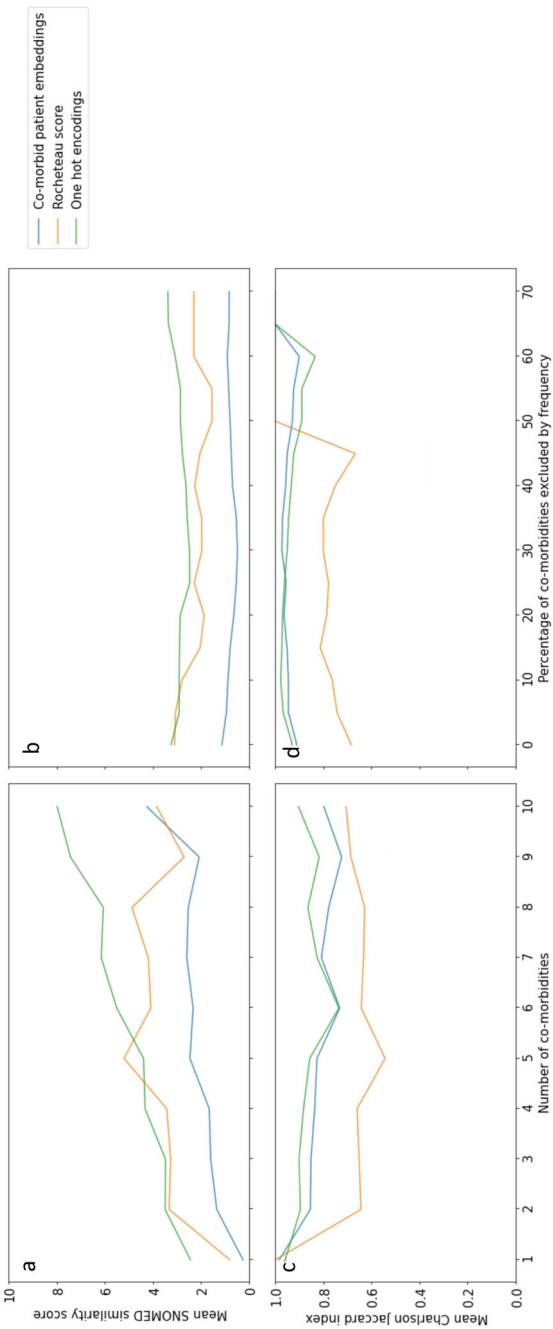

Fig. 5 Similar patient retrieval task evaluation by SNOMED similarity score (**a**, **b**) and Charlson Jaccard index (**c**, **d**) for patients with different numbers of co-morbdities (**a**, **c**) and those with co-morbidities of varying rarity (**b**, **d**) (co-morbidities were ordered based on their frequency in the EHR, and then starting from the most common, diseases and patients with those conditions were excluded based on percentages, as such, patients with all co-morbidities are included on the left hand side while only those with rare diseases are represented on the right hand side)

Fig. 6 Three survey examples of the patient in question and the similar patients retrieved by each method. The method that obtained the most votes for each question is shown in bold, with the similarity of co-morbidities indicated through a traffic light coloring scheme. Co-morbid patient embeddings identifies clinically appropriate examples with analogous diseases (e.g., alcoholism and alcohol dependence in question 10)

6.4 Additional Similar Patient Retrieval Task Analysis

Variation in clinical responses and decision making is to be expected [46]. In this survey reasonable heterogeneity was seen with a mean Gini index across questions of 0.46 (SD 0.14). The elbow method (Appendix Fig. 7) was applied to K-modes clustering [44] to find the most meaningful number of clusters within the data. 2 clusters was identified as the optimum. A multiple correspondence analysis [45] plot of these clusters is shown in the Appendix Fig. 8. Results indicate the 13 participants in cluster 0 are more concurrent, whereas those in cluster 1 showed greater variation in their responses. In particular, 8 clinicians in cluster 0 showed considerable agreement in their votes, which aligned with the overall results discussed above. 8 medical specialities from infectious diseases to general practitioners and internal medicine were represented in the survey, but no statistically significant differences were observed between these groups, which is likely due to a limited sample size. These investigations show that although heterogeneity in opinions exists for this

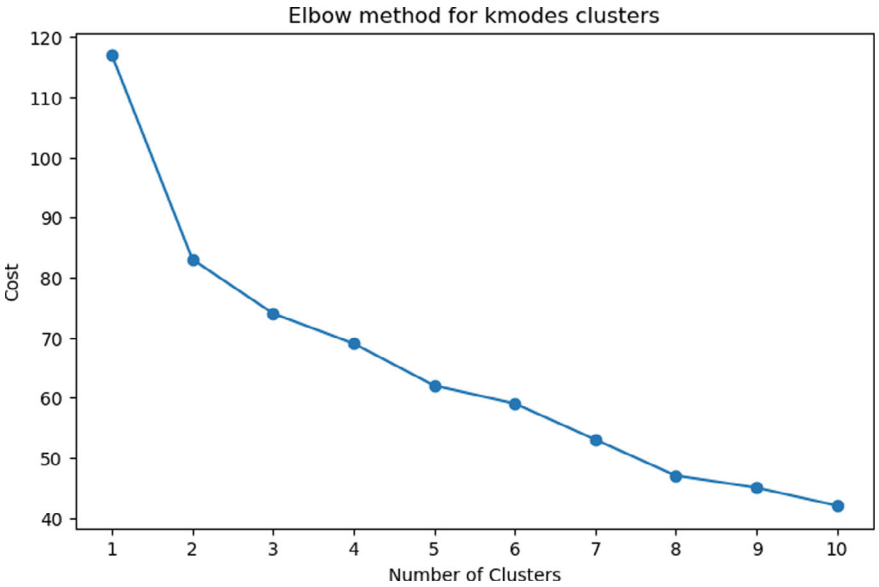

Fig. 7 Results for the elbow method on kmodes clustering of the survey data. Note that the infection point occurs at two clusters

type of clinical problem, appropriate research methodologies can be used to obtain a consensus and understand respondents differences.

Finally clinicians free text answers were examined. When asked to describe how and why they made their decisions a few key themes emerged, beyond the obvious matching of similar conditions. In particular 7 respondents noted that certain conditions, such as diabetes, cardiovascular and those that are more chronic, are of high-priority, while others are not required to be present in both patients. With one respondent stating "some diagnoses are core, and influence many aspects of care (diabetes) so must be included", while another said they "deprioritized 'benign' co-morbidities such as gastroesophageal reflux disease". Furthermore, 6 participants mentioned that it is crucial to include those diseases that are most life-threatening. Responding to "How would you define patient similarity?", 10 participants mentioned similar co-morbidities being a requirement as investigated in this research. However, other important factors were also mentioned including demographics (age, sex and ethnicity), as well as circumstance and contextual considerations. Respondents remarked that patient similarity can be useful to help inform treatment response and predict future risks, but emphasised the complex and multi-factorial nature of the task. Given the diversity in responses it is clear more work is needed for a gold standard definition and approach. Methodologies and research on clinically informed and data driven matching strategies, as employed in this study, progresses work towards this, with the aim of supporting AI applications that will ultimately benefit patients.

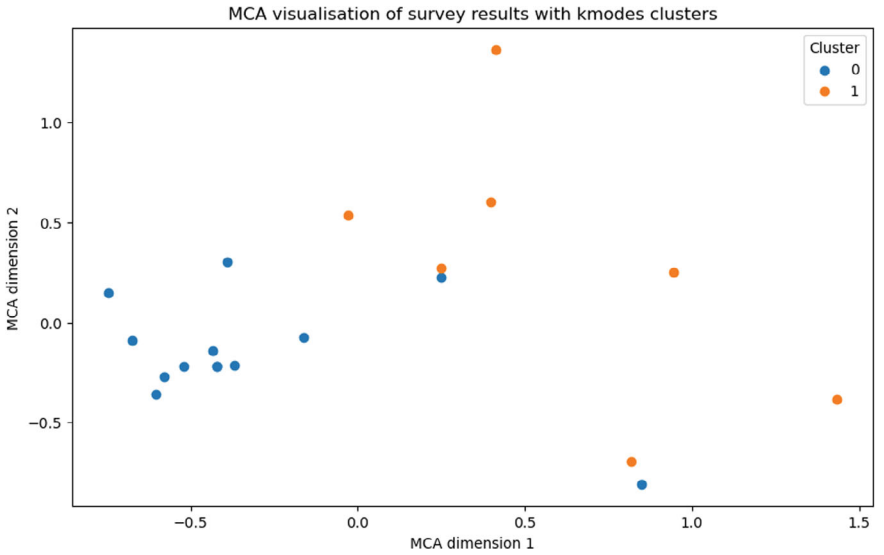

Fig. 8 Multiple correspondence analysis visualisations of the survey results with kmodes clusters highlighted. Notice how those in cluster 0 are more closely located implying agreement in the responses. In particular 8 respondents are highly clustered (note that two respondents were identical and hence overlap on this plot). Cluster 1 on the other hand is more heterogeneous

6.5 Additional Limitations and Future Work

One limitation for the similar patient retrieval task and survey is the sample size. Moving forwards we hope to collect more survey responses and conduct further primary research on such healthcare focused AI methodologies with clinicians and patients. For the supervised learning task one drawback of our approach is that set transformers using our SNOMED disease embeddings are much greater in complexity than logistic regression meaning they require greater computational resources and are potentially less interpretable. Finally, due to no data being present in iCARE on if or when patients recovered from their chronic conditions, this research does not consider the temporal evolution of a patients co-morbid conditions but this is an avenue for future research. Additional future work includes representing patients with more information than just the mean of their diseases. For example, by weighting conditions according to the Charlson co-morbidity index [21, 22], to ensure those co-morbidities considered a high-priority are suitably accounted for, or by using more sophisticated methods such as a set transformer variational autoencoder to prevent information such as the patients number of co-morbidities from being lost. Furthermore, the SNOMED embeddings themselves could potentially be improved by considering that edges should have different magnitudes at different depths of the SNOMED CT hierarchy [47] (Table 2 and Fig. 9).

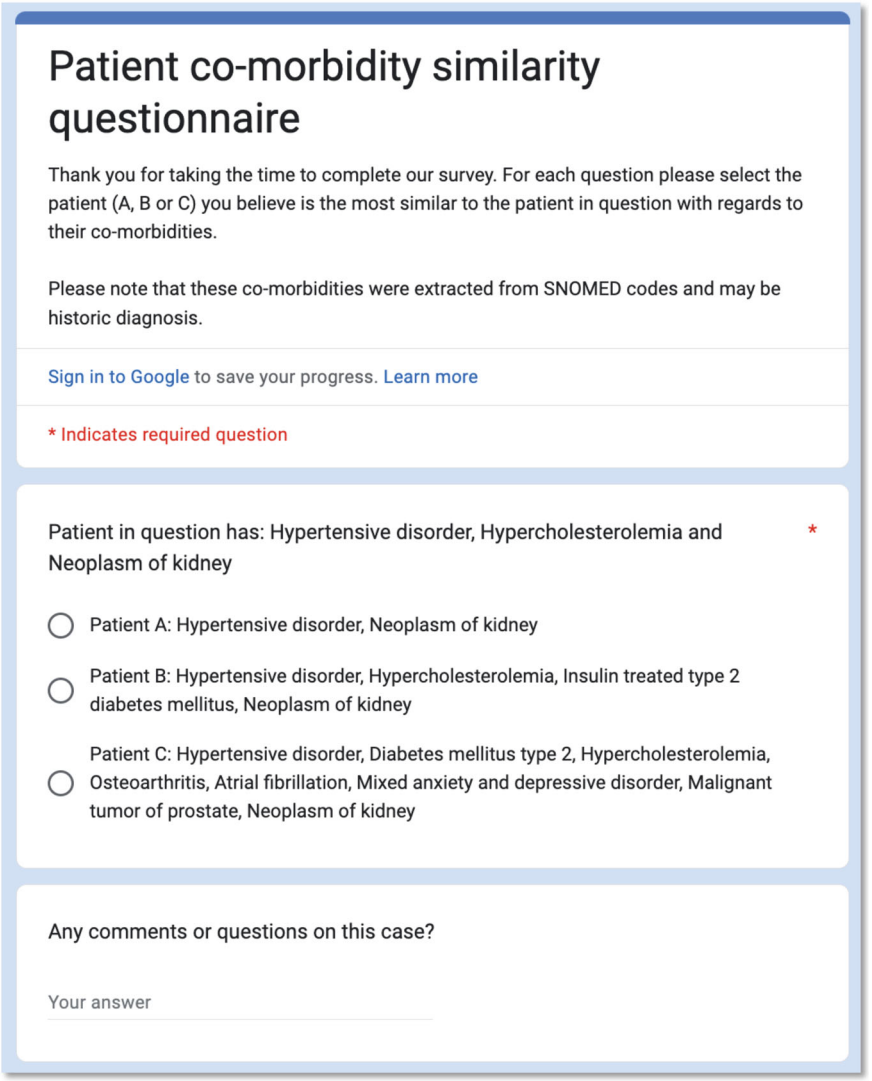

Fig. 9 Screenshot of a section from the multi-morbid patient similarity survey. Notice the context and instructions given to participants and that patient similarity questions were asked in a plain text format so as not to introduce any biases

Table 2 Co-morbidities for the patient in question and for those patients identified as most similar by each method for each question in the multi-morbid patient similarity survey

Question	Patient	Co-morbid patient embeddings	One hot encodings	Rocheteau similarity score
1	Hypertensive disorder, Hypercholesterolemia and Neoplasm of kidney	Hypertensive disorder, Neoplasm of kidney	Hypertensive disorder, Hypercholesterolemia, Insulin treated type 2 diabetes mellitus, Neoplasm of kidney	Hypertensive disorder, Diabetes mellitus type 2, Hypercholesterolemia, Osteoarthritis, Atrial fibrillation, Mixed anxiety and depressive disorder, Malignant tumor of prostate, Neoplasm of kidney
2	Asthma, Depressive disorder, Hypothyroidism and Fibromyalgia	Asthma, Depressive disorder, Hypothyroidism, Asperger	Asthma, Depressive disorder, Fibromyalgia	Asthma, Hypertensive disorder, Diabetes mellitus type 2, Gastroesophageal reflux disease, Cerebrovascular accident, Mixed anxiety and depressive disorder, Hyperlipidemia, Fibromyalgia
3	Ischemic heart disease, Hearing loss, Dementia, Essential hypertension, Alzheimer's disease and Diverticulosis of sigmoid colon	Hypertensive disorder, Myocardial infarction, Cataract, Dementia, Alzheimer's disease	Hearing loss, Dementia	Hypertensive disorder, Hypercholesterolemia, Transient ischemic attack, Osteoarthritis, Diverticular disease, Benign prostatic hyperplasia, Myocardial infarction, Alzheimer's disease, Retention of urine, Small vessel cerebrovascular disease
4	Hypertensive disorder, Varicella, Polycystic ovaries and Scoliosis deformity of spine	Hypertensive disorder, Varicella, Polycystic ovaries	Varicella, Polycystic ovaries, Scoliosis deformity of spine	

(continued)

Table 2 (continued)

Question	Patient	Co-morbid patient embeddings	One hot encodings	Rocheteau similarity score
5	Asthma, Hemorrhoids, Colitis, Viral hepatitis type B, Gastritis, Human immunodeficiency virus infection, Supraventricular tachycardia, Non-alcoholic fatty liver, Portal hypertension and Hemophilia	Asthma, Osteoporosis, Hypertensive disorder, Hypercholesterolemia, Vitamin D deficiency, Depressive disorder, Hemorrhoids, Atrial fibrillation, Chronic obstructive lung disease, Anemia, Rheumatoid arthritis, Inflammatory diseaser of liver, Dysphagia, Steatosis of liver, Aortic valve regurgitation	Asthma, Depressive disorder, Hemorrhoids, Human immunodeficiency virus infection	Hemorrhoids, Atrial fibrillation, Hypothyroidism, Viral hepatitis type B
6	Hypertensive disorder, Diabetes mellitus type 2, Hypercholesterolemia, Vitamin D deficiency, Benign prostatic hyperplasia, Angina, Old myocardial infarction and Chronic ischemic heart disease	Hypertensive disorder, Diabetes mellitus type 2, Ischemic heart disease, Angina	Hypertensive disorder, Diabetes mellitus type 2, Hypercholesterolemia, Vitamin D deficiency	Hypertensive disorder, Diabetes mellitus type 2, Benign prostatic hyperplasia, Gastroesophageal reflux disease, Iron deficiency anemia, Acute non-ST segment elevation myocardial infarction, Angina, Heart failure, Constipation, Hyperlipidemia, Retention of urine, Bilateral cataracts, Heart failure with normal ejection fraction
7	Varicella and Rheumatoid arthritis with arteritis	Varicella, Rheumatoid arthritis	Rheumatoid arthritis with arteritis	Asthma, Varicella, Transient ischemic attack, Cerebrovascular accident, Fibromyositis, Rheumatoid arthritis with aortitis, Seizures in response to acute event
8	Hypertensive disorder, Varicella, Gestational diabetes mellitus and Pre-eclampsia	Varicella, Gestational diabetes mellitus, Pre-eclampsia, Pregnancy-induced hypertension	Hypertensive disorder, Varicella, Gestational diabetes mellitus	Varicella, Gestational diabetes mellitus, Pre-eclampsia

(continued)

Table 2 (continued)

Question	Patient	Co-morbid patient embeddings	One hot encodings	Rocheteau similarity score
9	Asthma, Hypertensive disorder, Osteoarthritis, Type 2 diabetes, Hypercholesterolemia, Anemia, Gastroesophageal reflux disease and Hypothyroidism	Asthma, Hypertensive disorder, Osteoarthritis, Diabetes, Hyperlipidemia, Anemia, Gastroesophageal reflux disease, Obstructive sleep apnea	Asthma, Hypertensive disorder, Osteoarthritis, Type 2 diabetes, Hypercholesterolemia	Asthma, Hypertensive disorder, Rheumatoid arthritis, Diabetes, Hypercholesterolemia, Anemia, Gastroesophageal reflux disease, Hypothyroidism, Coronary arteriosclerosis, Pulmonary embolism, Chronic kidney disease
10	Osteoarthritis and Alcoholism	Osteoarthritis, Alcohol dependence	Osteoarthritis, Alcoholism, Peripheral nerve entrapment	

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019 Jan;25(1):44–56. Number: 1 Publisher: Nature Publishing Group.
2. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. The Lancet Digital Health. 2021 Apr;3(4):e260-5. Publisher: Elsevier.
3. Gijsen R, Hoeymans N, Schellevis FG, Ruwaard D, Satariano WA, van den Bos GAM. Causes and consequences of comorbidity: A review. Journal of Clinical Epidemiology. 2001;54(7):661–74.
4. Riaño D, Peleg M, ten Teije A. Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges. Artificial Intelligence in Medicine. 2019;100:101713.
5. Salive ME. Multimorbidity in older adults. Epidemiologic Reviews. 2013;35:75–83.
6. Divo MJ, Martinez CH, Mannino DM. Ageing and the epidemiology of multimorbidity. European Respiratory Journal. 2014 Oct;44(4):1055–68. Publisher: European Respiratory Society Section: State of the Art.
7. Skou ST, Mair FS, Fortin M, Guthrie B, Nunes BP, Miranda JJ, et al. Multimorbidity. Nature Reviews Disease Primers. 2022 Jul;8(1):1–22. Number: 1 Publisher: Nature Publishing Group.
8. Millar J. The Need for a Global Language-SNOMED CT Introduction. Studies in Health Technology and Informatics. 2016;225:683–5.
9. NHS. SNOMED CT; 2023.
10. Gerke S, Minssen T, Cohen G. Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. Academic Press; 2020. pp. 295–336.
11. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. Journal of the American Medical Informatics Association. 2021;28(9):2017–26.
12. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The Graph Neural Network Model. IEEE Transactions on Neural Networks. 2009 Jan;20(1):61–80. Conference Name: IEEE Transactions on Neural Networks.

13. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY, USA: Association for Computing Machinery; 2016. pp. 855–64.
14. Narayanan A, Chandramohan M, Chen L, Liu Y, Saminathan S. subgraph2vec: Learning Distributed Representations of Rooted Sub-graphs from Large Graphs. arXiv; 2016. ArXiv:1606.08928 [cs].
15. Narayanan A, Chandramohan M, Venkatesan R, Chen L, Liu Y, Jaiswal S. graph2vec: Learning Distributed Representations of Graphs. arXiv; 2017. ArXiv:1707.05005 [cs].
16. Choi Y, Chiu CYI, Sontag D. Learning Low-Dimensional Representations of Medical Concepts. AMIA Summits on Translational Science Proceedings. 2016;2016:41–50.
17. Cui H, Lu J, Wang S, Xu R, Ma W, Yu S, et al.. A Survey on Knowledge Graphs for Healthcare: Resources, Applications, and Promises. arXiv; 2023. ArXiv:2306.04802 [cs].
18. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Scientific Data. 2023 Feb;10(1):67. Number: 1 Publisher: Nature Publishing Group.
19. Yu Y, Huang K, Zhang C, Glass LM, Sun J, Xiao C. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. Bioinformatics. 2021;37(18):2988–95.
20. Su X, Hu L, You Z, Hu P, Zhao B. Attention-based Knowledge Graph Representation Learning for Predicting Drug-drug Interactions. Briefings in Bioinformatics. 2022 May;23(3):bbac140.
21. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Diseases. 1987;40(5):373–83.
22. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. Journal of Clinical Epidemiology. 1992;45(6):613–9.
23. Violán C, Foguet-Boreu Q, Fernández-Bertolín S, Guisado-Clavero M, Cabrera-Bean M, Formiga F, et al. Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a Mediterranean population. BMJ open. 2019;9(8):e029594.
24. Haug N, Deischinger C, Gyimesi M, Kautzky-Willer A, Thurner S, Klimek P. High-risk multimorbidity patterns on the road to cardiovascular mortality. BMC medicine. 2020;18(1):44.
25. Zador Z, Landry A, Cusimano MD, Geifman N. Multimorbidity states associated with higher mortality rates in organ dysfunction and sepsis: a data-driven analysis in critical care. Critical Care. 2019;23(1):247.
26. Majnarić LT, Babič F, O'Sullivan S, Holzinger A. AI and Big Data in Healthcare: Towards a More Comprehensive Research Framework for Multimorbidity. Journal of Clinical Medicine. 2021;10(4):766.
27. Nielsen AB, Thorsen-Meyer HC, Belling K, Nielsen AP, Thomas CE, Chmura PJ, et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. The Lancet Digital Health. 2019;1(2):e78-89.
28. Noll R, Schaaf J, Storf H. The Use of Computer-Assisted Case-Based Reasoning to Support Clinical Decision-Making - A Scoping Review. In: Keane MT, Wiratunga N, editors. Case-Based Reasoning Research and Development. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2022. pp. 395–409.
29. Diaz Ochoa JG, Mustafa FE. Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses. Artificial Intelligence in Medicine. 2022;131:102359.
30. Hu Z, Qiu H, Wang L, Shen M. Network analytics and machine learning for predicting length of stay in elderly patients with chronic diseases at point of admission. BMC Medical Informatics and Decision Making. 2022;22(1):62.
31. Rocheteau E, Tong C, Veličković P, Lane N, Liò P. Predicting Patient Outcomes with Graph Representation Learning. arXiv; 2021. ArXiv:2101.03940 [cs].
32. SNOMED_International. SNOMED CT Starter Guide; 2023.

33. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2019.
34. Lee J, Lee Y, Kim J, Kosiorek AR, Choi S, Teh YW. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. arXiv; 2019. ArXiv:1810.00825 [cs, stat].
35. Maneewongvatana S, Mount DM. Analysis of approximate nearest neighbor searching with clustered point sets. arXiv; 1999. ArXiv:cs/9901013.
36. Sharafoddini A, Dubin JA, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. JMIR Medical Informatics. 2017 Mar;5(1):e6730. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.
37. UMLS. Metathesaurus; 2023. Publisher: U.S. National Library of Medicine.
38. Lyudovyk O, Weng C. SNOMEDtxt: Natural Language Generation from SNOMED Ontology. Studies in health technology and informatics. 2019;264:1263–7.
39. Rocheteau E, Liò P, Hyland S. Temporal Pointwise Convolutional Networks for Length of Stay Prediction in the Intensive Care Unit. In: Proceedings of the Conference on Health, Inference, and Learning. CHIL '21. New York, NY, USA: Association for Computing Machinery; 2021. pp. 58–68.
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–30.
41. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357–62.
42. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. pp. 8024–35.
43. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization; 2014. Available from: https://arxiv.org/abs/1412.6980v9.
44. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery. 1998;2(3):283–304.
45. Greenacre M, Blasius J. Multiple Correspondence Analysis and Related Methods. CRC Press; 2006. Google-Books-ID: uAXOBQAAQBAJ.
46. Bate L, Hutchinson A, Underhill J, Maskrey N. How clinical decisions are made. British Journal of Clinical Pharmacology. 2012;74(4):614–20.
47. Wu Z, Palmer M. Verb Semantics and Lexical Selection. arXiv; 1994. ArXiv:cmp-lg/9406033.