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

William Bolton[1, 2, 3, 4], **Pantelis Georgiou**[3, 4, 5], **Alison Holmes**[3, 4, 6, 7], **Timothy Rawson**[3, 4]

[1]Department of Computing, Imperial College London, UK. [2]UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, UK. [3]Centre for Antimicrobial Optimisation, Imperial College London, UK. [4]National Institute for Health Research, Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Imperial College London, UK. [5]Centre for Bio-inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, UK. [6]Faculty of Health Life Sciences, University of Liverpool, UK . [7]Department of Infectious Diseases, Imperial College London, UK.

Why?

- **Co-morbidities** defined here as chronic long-term medical conditions are a **major challenge in healthcare**[1]
- **Challenges exist** with representing and using co-morbidity data in AI systems[1]:
  - **Combinatorial complexity** due to the large number of unique diseases
  - Sparsity, missingness and a **lack of data** particularly for those with rare diseases or complex co-morbidity combinations
  - **Heterogeneity** in how chronic conditions are recorded
- Existing AI research on co-morbid patients does not tackle these problems and therefore **lacks appropriate representation**

Aim

Creating **meaningful embeddings** from **external medically grounded knowledge**, to help **overcome such challenges** and **support downstream AI applications**

How?

- Processed **SNOMED CT**[2] a comprehensive clinical healthcare terminology into a connected undirected graph
- Generated **disease embeddings** through Node2vec[3] with optimization to reduce the mean SNOMED distance (shortest path length) between each node and their nearest neighbor
- Tested disease embeddings as sole features to **supervised learning models** for clinically relevant predictions
- Defined **co-morbid patient embeddings** as the mean of all the SNOMED disease embeddings for a particular individual
- Evaluated co-morbid patient embeddings through a **task to retrieve the most similar patient** for any given unique co-morbid patient, where no identical match was possible
  - Retrieved similar patients in our method through **nearest neighbor lookup** based on euclidian distance
  - Utilized two **novel metrics** and **human experts** as evaluators

Results

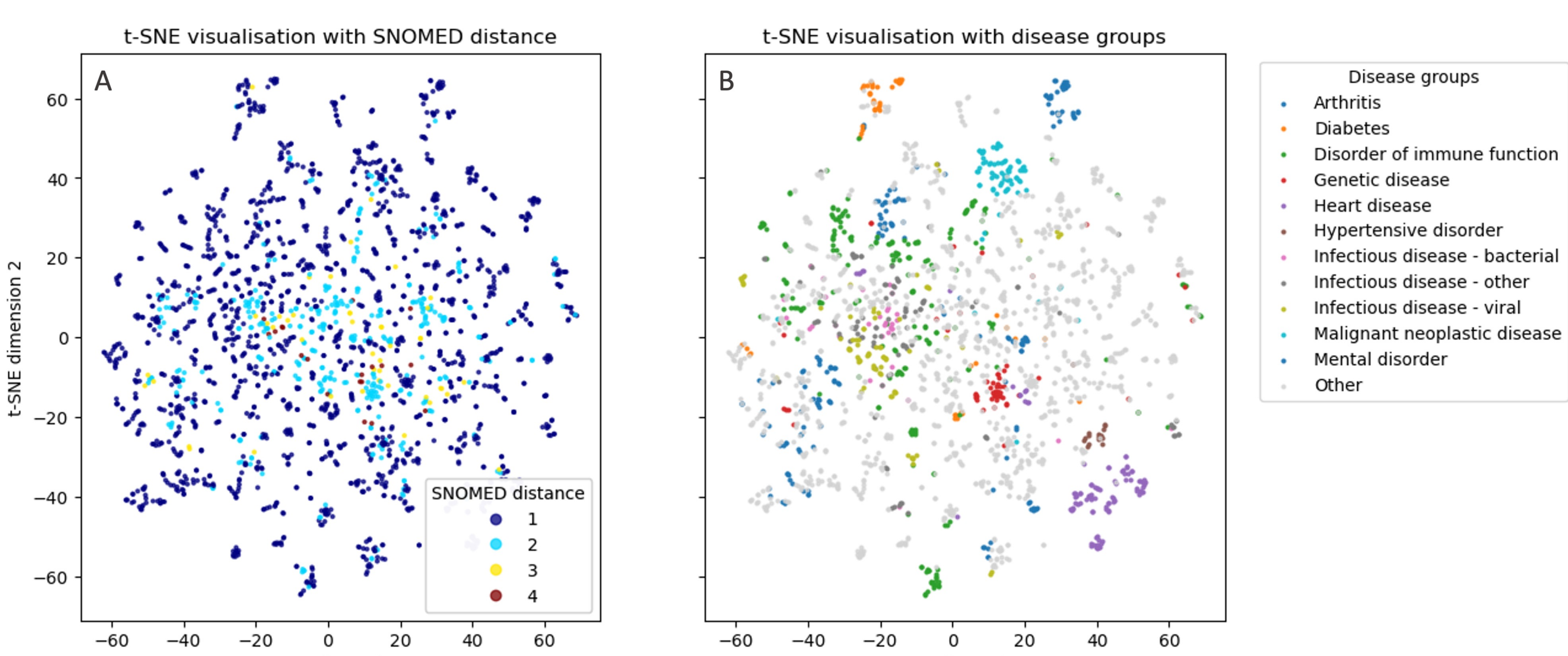

Figure 1: t-SNE visualisations of the SNOMED disease embeddings (2,133 diseases) with [A] nearest neighbor SNOMED distance (hyperparameter optimisation resulted in a mean of 1.23) and [B] high-level disease groups displayed.

Table 1: Mean unseen test set AUROC results for supervised learning classification tasks in different populations.

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

Table 2: Mean evaluation results for the similar patient retrieval task.

Method	SNOMED similarity score	Charlson Jaccard index
One hot encodings	4.40 (SD 2.32)	**0.88 (SD 0.30)**
Rocheteau's method[4]	3.52 (SD 3.26)	0.69 (SD 0.20)
Co-morbid patient embeddings	**1.78 (SD 1.90)**	0.84 (SD 0.34)

Figure 2: Equations used to determine SNOMED similarity score and Charlson Jaccard index.

$$SNOMED\ sim_{p1,p2} = f(S_{p1,p2}) + f(S_{p2,p1})$$

where $S_{p1,p2}$ is a SNOMED distance matrix for the patients co-morbidities, we match each disease of $p1$ to a disease of $p2$ so that the matching minimized the following equation:

$$f(A) = \sum_{i-1}^{n} min_{j\in\{1,\dots,m\}}\left(1 - \frac{1}{A_{ij}+1}\right)$$

where $A \in \mathbb{R}^{n\times m}$

$$Charlson\ Jaccard\ index_{p1,p2} = \frac{|C_{p1} \cap C_{p2}|}{|C_{p1} \cup C_{p2}|}$$

where Crepresents the set of Charlson co-morbidities[5] for a particular patient

The Charlson co-morbidity index[5] is a widely adopted clinical tool that classifies some specific co-morbidities to 17 different categories

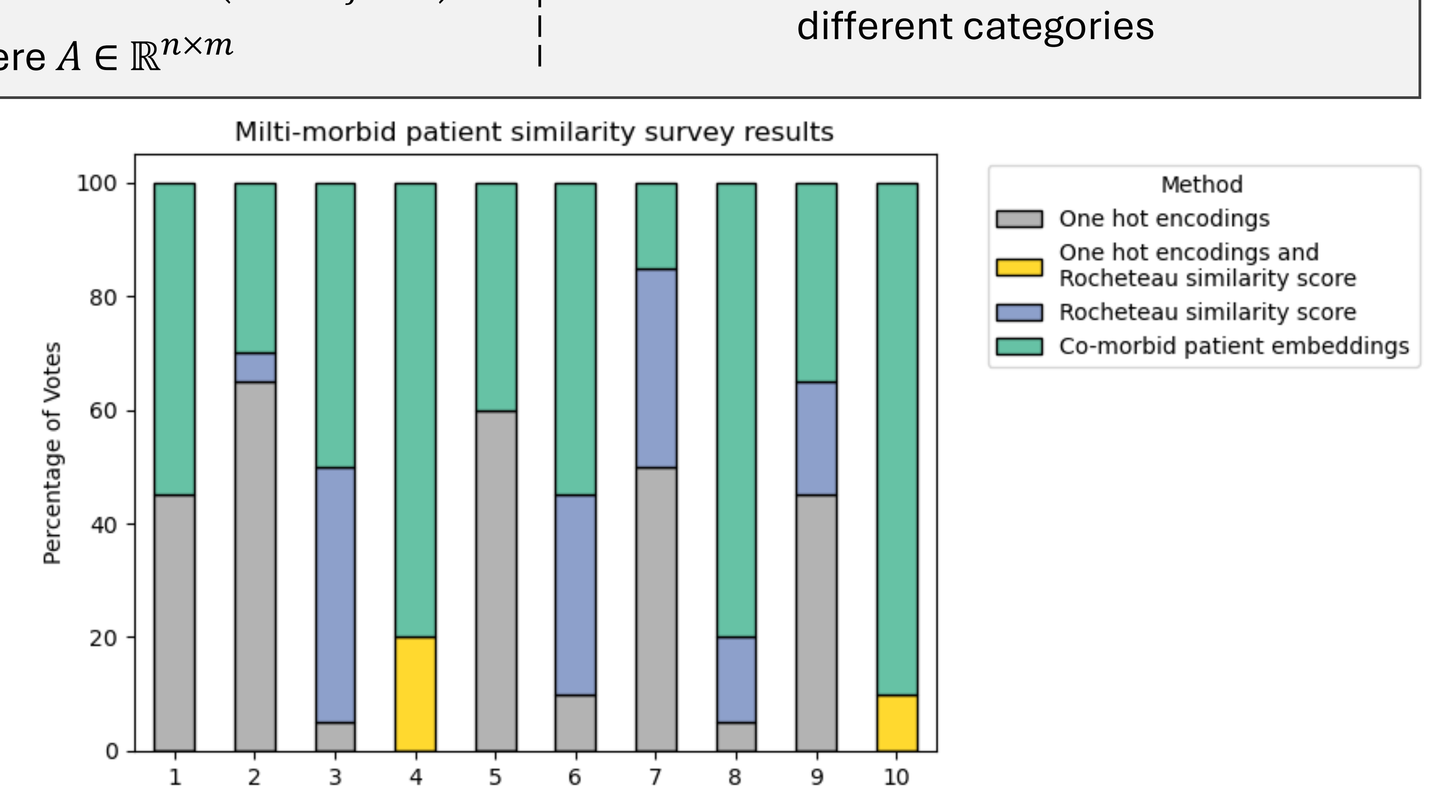

Figure 3: Proportion of human expert votes for patients identified by each method, for each question, in a survey. Co-morbid patient embeddings obtained the most votes for 6 questions.

Co-morbidities

Question 8 patient	Gestational diabetes mellitus	Hypertensive disorder	Pre-eclampsia	Varicella
Co-morbid patient embeddings	Gestational diabetes mellitus	Pregnancy-induced hypertension	Pre-eclampsia	Varicella
Rocheteau score	Gestational diabetes mellitus	Hypertensive disorder	-	Varicella
One hot encodings	Gestational diabetes mellitus	-	Pre-eclampsia	Varicella

Question 10 patient	Osteo-arthritis	Alcoholism	
Co-morbid patient embeddings	Osteo-arthritis	Alcohol dependence	
Rocheteau score	Osteo-arthritis	Alcoholism	Peripheral nerve entrapment
One hot encodings	Osteo-arthritis	Alcoholism	Peripheral nerve entrapment

Identical
Similar
Dissimilar

Figure 4: Two examples of the patient in question and the similar patients retrieved by each method. The similarity of co-morbidities is indicated through a traffic light coloring scheme.

Discussion

- We developed a **novel pipeline** to extract and utilize **untapped medical knowledge** and demonstrated its utility in classification and similar patient retrieval tasks with automatic and human evaluation
- Our approach is **generalizable** and can overcome some problems with using disease data in AI systems as the **embeddings are not influenced by dataset size, the number of diseases or their rareness** and are **adaptable to variation in clinical documentation**
- Future work includes, considering **temporal** aspects and **embedding additional clinical data** such as demographics and medications

Acknowledgements

William Bolton was supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. P/S023283/1), the National Institute for Health, Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at Imperial College London and the Department of Health and Social Care, Centre for Antimicrobial Optimisation. This research was enabled by the Imperial Clinical Analytics Research and Evaluation (iCARE) environment and used the iCARE team and data resources.

References

[1] Rian õ 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 Sep;100:101713. [2] Millar J. The Need for a Global Language – SNOMED CT Introduction. Studies in Health Technology and Informatics. 2016;225:683-5.; [3] 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. p. 855-64. [4] Rocheteau E, Tong C, Velickovic P, Lane N, Lio P. Predicting Patient Outcomes with Graph Representation Learning. arXiv; 2021. ArXiv:2101.03940 [cs]. [5] Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. Journal of Clinical Epidemiology. 1992 Jun;45(6):613-9.