Creating Disease-Symptom Health Knowledge Graphs

Health knowledge graphs (HKGs)[1/2/3] particularly symptom-disease graphs[4/5/6] are popular knowledge representations of medical data. These can be used as self-diagnostic tools by patients, or to aid physicians on difficult cases. Rotmensch [7] demonstrated the benefits of an automated approach to generating a HKG from medical reports; it is faster than manual construction, requires no medical knowledge, and can be kept up-to-date with real world information. The study's dataset however was very limited, and other HKGs also utilise only one data source. The goal of our project was therefore to extend the work on symptom-disease HKGs by comparing HKGs created from three disparate data sources, and explore the viability of combining them.

The first dataset[8] [Fig1] was taken from a paper studying disease-gene relationships[9]. Term frequency-inverse document frequency metrics quantified the association between diseases and symptoms co-occurring in articles from PubMed.

Figure 1: The "PubMed" dataset

MeSH Symptom Term	MeSH Symptom Code	MeSH Disease Term	MeSH Disease Code	TFIDF score
Chest Pain	D002637	Heart Diseases	D006331	119.852131
Apraxias	D001072	Ataxia	D001259	90.583415
Fever	D005334	Obstetric Labor Complications	D007744	25.995381
Reflex, Abnormal	D012021	Cerebral Infarction	D002544	53.470976
Muscular Atrophy	D009133	Pulmonary Disease, Chronic Obstructive	D029424	64.276158

For the second dataset[10] [Fig3] NLP system MedLEE was used to extract data from hospital medical records, and co-occurrence statistics calculated the association between diseases and symptoms[11]. As the published dataset did not contain the specific weights, we created our own custom metric, where the weight of an edge is higher if a symptom rarely occurs, and the strength of the ranked association between disease and symptom is high [Fig2].

Figure 2: The weight w_{ij} of the edge between symptom i and disease j is given as:
$w_{ij} = 10\,[\,0.4\,(\frac{symptomFrequency_{max} - symptomFrequency_i}{symptomFrequency_{max}}) + 0.8\,(associationStrength_{ij})\,]$

Figure 3: The "Medical Records" dataset

symptom	symptom_umls	disease	disease_umls	symptom_frequency	association_strength	weight
pain chest	C0008031	hypertensive disease	C0020538	21	0.966667	10.053333
shortness of breath	C0392680	hypertensive disease	C0020538	49	0.933333	7.546667
dizziness	C0012833	hypertensive disease	C0020538	8	0.900000	10.560000
asthenia	C0004093	hypertensive disease	C0020538	24	0.866667	9.013333
fall	C0085639	hypertensive disease	C0020538	9	0.833333	9.946667

Figure 4: A SPARQL query on DBpedia for diseases and their related symptoms

```
SELECT ?disease ?disease_mesh ?symptom ?symptom_mesh WHERE {
 ?disease a dbo:Disease .
 ?disease dbo:meshId ?disease_mesh .
 ?disease dbo:symptom ?symptom .
 ?symptom dbo:meshId ?symptom_mesh .
 }
 ORDER BY ?disease
```

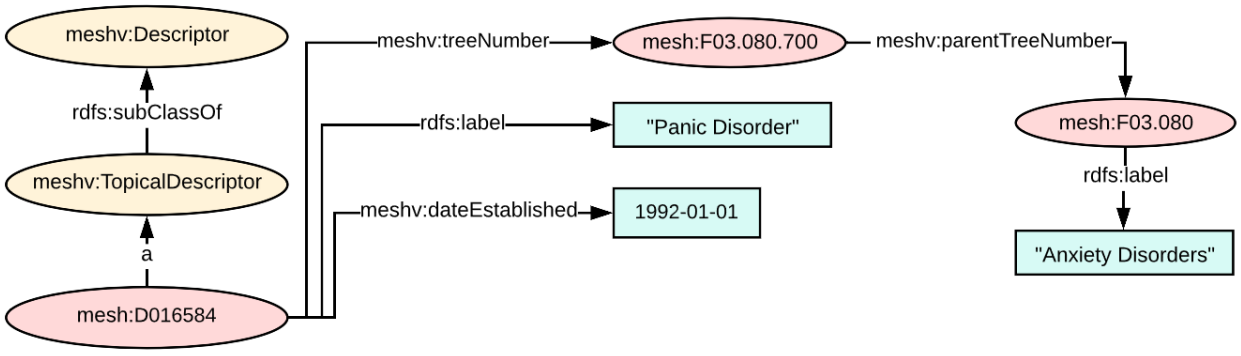
We extracted the third dataset [Fig5] from the Wikipedia knowledge graph DBpedia[12], a previously unexploited source for a HKG. A SPARQL query in RDF triple form [Fig4] extracts entities of class *disease* and the subjects of the *symptom* and *meshId* properties, as defined in the DBpedia ontology.

Figure 5: The "DBpedia" dataset

symptom	symptom_mesh	disease	disease_mesh
Fatigue	D005221	Pernicious anemia	D000752
Nausea	D009325	Hepatitis A	D006506
Fatigue	D005221	Cardiovascular disease	D002318
Fever	D005334	Bone tumor	D001859
Cough	D003371	Allergy	D006967

A MeSH code is a unique identifier for conditions that is part of a controlled medical vocabulary, which can be represented as linked data[13] [Fig6]. In order to standardise the representation of conditions across all of the datasets, the codes were mapped to UMLS identifiers via requests to the UMLS API[14].

Figure 6: A subset of the MeSH vocabulary linked data graph



We encountered issues with incorrect codes, particularly in the DBpedia dataset, as Wikipedia editors had often added the wrong identifier. If two different diseases have been assigned the same code, it must be determined whether this was a mistake, or if they should refer to the same concept, e.g., *panic disorder* and *panic attack* both had identical codes.

As per [9], each disease can be represented as a feature vector with binary values indicating whether each symptom is associated with that disease [Fig7].

Figure 7: A matrix of feature vectors for each disease (row) associated with each symptom (column)

	Fever	Shortness of breath	Fatigue	Abdominal pain	Headache	Vomiting	Nausea	Diarrhea	Chest pain	Jaundice	...
Adenovirus infection	1	0	1	1	1	1	0	0	0	0	...
Liver failure	0	0	0	0	0	1	1	0	0	1	...
Peanut allergy	0	0	0	1	0	0	0	1	0	0	...
Lupus	1	0	1	0	0	0	0	0	1	0	...
Influenza	1	0	1	0	1	0	0	0	0	0	...

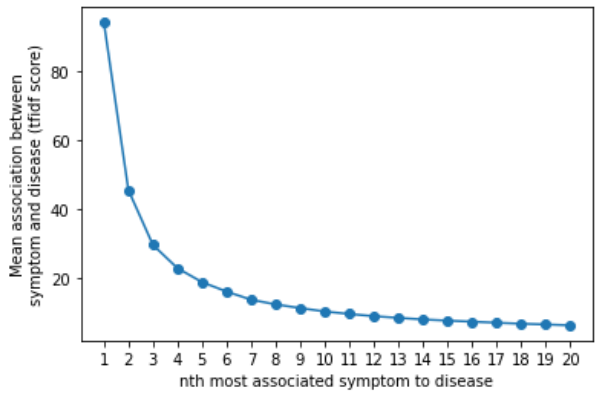
Distance metrics can then be used to examine the potential equivalency between diseases; the Jaccard index was used for the DBpedia dataset to emphasise the symptoms that diseases have in common, as in this dataset each disease was associated with few symptoms. [Fig8] shows how *panic attack* is more similar to *panic disorder* than it is to *tetanus.* In the project incorrect codes were fixed manually, however this approach demonstrates a potential automated solution, if appropriately and contextually adjusted to the dataset.

Figure 8: The distance between "Panic Attack" and other diseases

	Panic attack
Panic disorder	0.500
Cardiogenic shock	0.600
Fever	0.750
Chemical burn	0.833
Tetanus	0.833

After the codes had been standardised, the datasets could then be combined to form one large HKG. To prevent one dataset from dominating the others due to the different edge weight scales, all of the weights were standardised to 1. The weights of edges shared between datasets were summed, so that a weight of 2 means 2 of the datasets contained that association. Only the 10th most associated symptoms to each disease were retained from the larger PubMed dataset, so that the edges reflected stronger connections [Fig9].

Figure 9: The average tfidf score for the nth most associated symptom to diseases. Beyond n=10 the mean score is greatly reduced.



The datasets were visualised in Gephi[15] as directed graphs [Fig10]. Gephi is able to produce structured, explorable network visualisations, and can also compute graph metrics.

The graphs are very dense with small network diameters. The size of a node reflects its degree centrality; calculating the in-degree and betweenness measures resulted in the same nodes being highlighted as important. In all of the HKGs, large common symptom nodes tend to form the centre of clusters, surrounded by diseases that are strongly associated with that symptom. Gephi's modularity algorithm utilising the Louvain method[16] was run with a high resolution to suppress the number of clusters discovered; these are distinguished by colour.

Figure 10a: The "PubMed" HKG visualised in Gephi

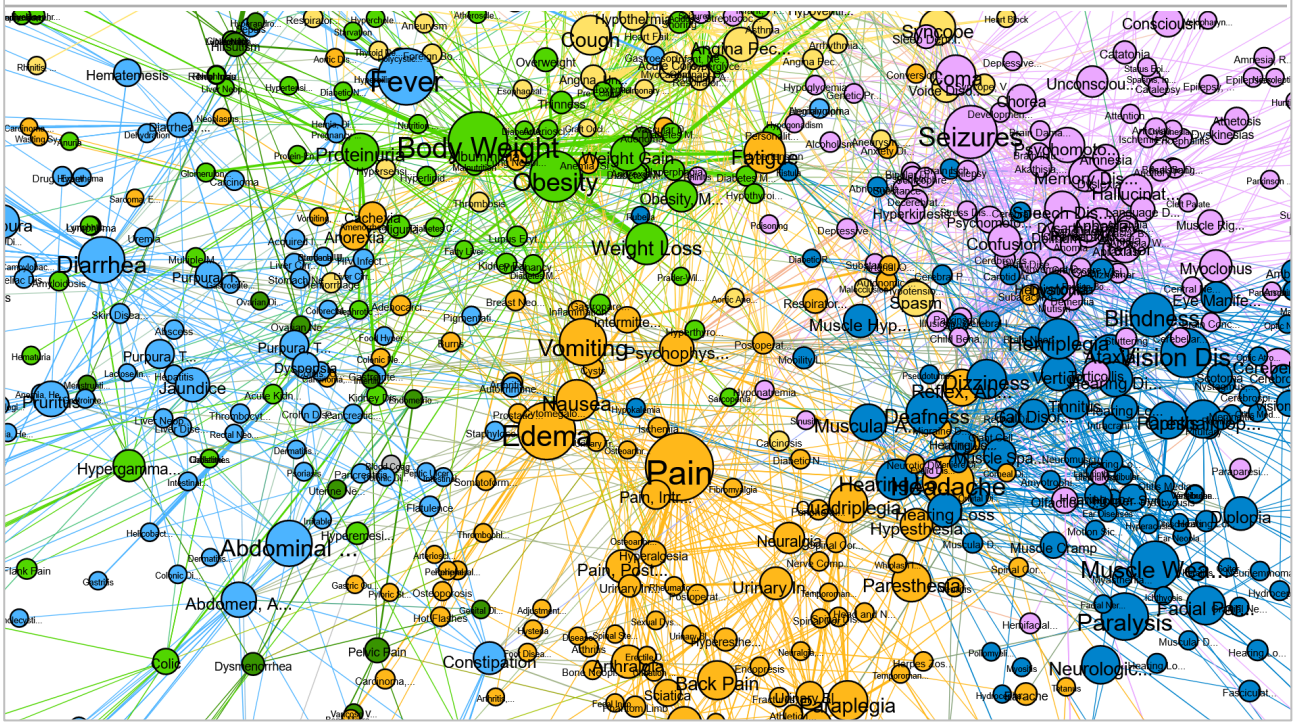


Figure 10b: The "Medical Records" HKG visualised in Gephi

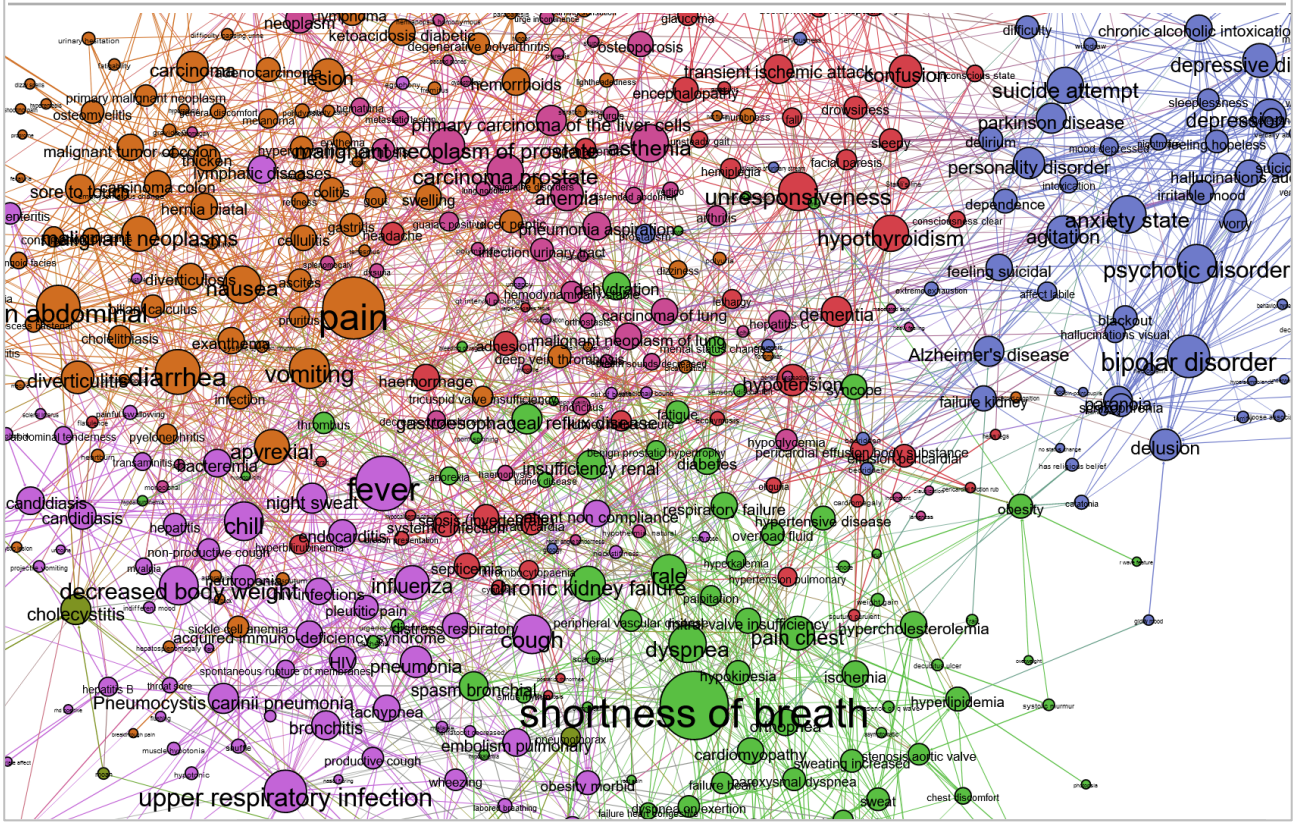
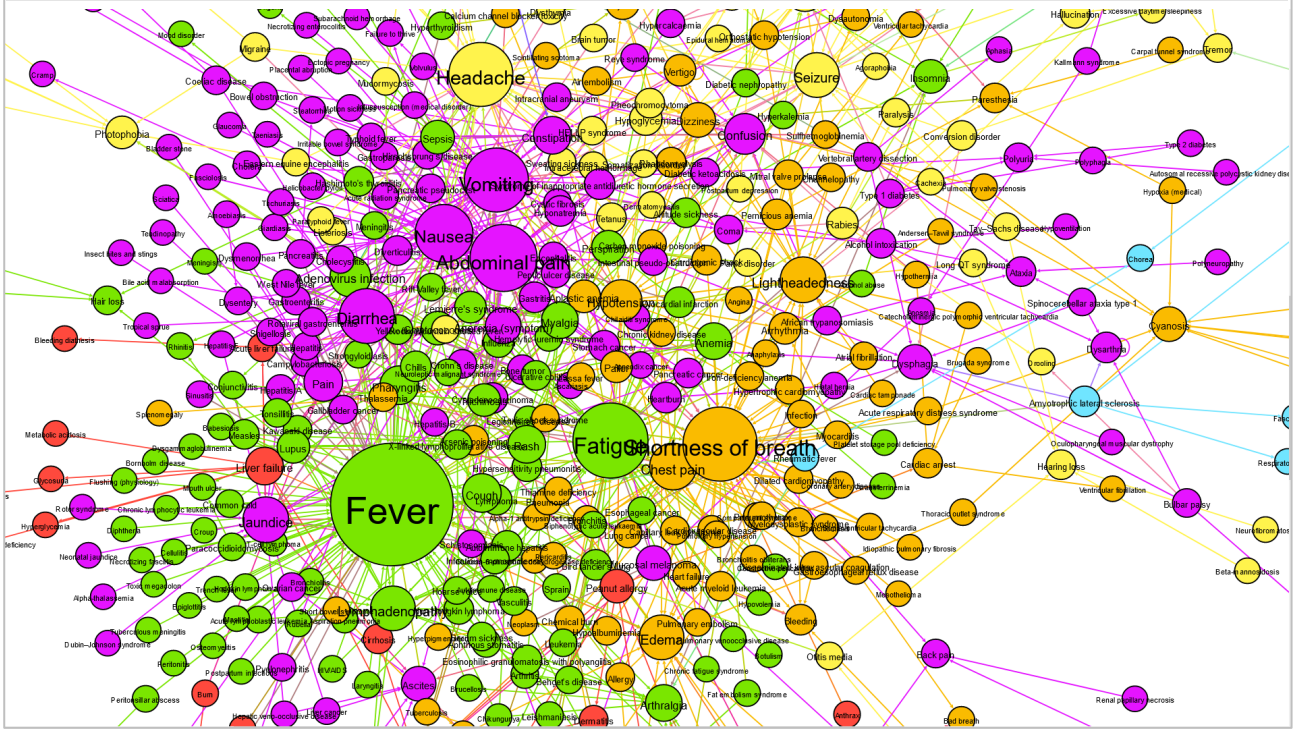


Figure 10c: The "DBpedia" HKG visualised in Gephi



The DBpedia HKG depicts 3 core clusters that are present in all of the graphs and reflect common medical conditions: centred around *pain*, associated with abdominal-based diseases, *fever*, linked to infections and influenza-based diseases, and *shortness of breath*, associated with cardiac-based diseases.

However, there are also clusters unique to each HKG:

The Medical Records HKG has a strongly separated *mental disorders* cluster, suggesting that hospitals handle a disproportionately higher number of cases of mental breakdowns.

The PubMed HKG's *body weight* cluster reflects the high frequency at which *weight loss* is discussed in journal articles, even though *obesity* does not appear as an important node in either of the other graphs. A cluster of *disabilities* reveals a potential downside of using literature as a data source; although diseases like *blindness* and *deafness* have little similarity with regards to shared symptoms, they are grouped together in the HKG, likely due to co-occurrences in papers related to disabilities.

The combined HKG [Fig11] can be depicted in full, representing an extensive but imprecise graph, useful for physicians and exploration. Alternatively, it can be filtered to edges with weights > 1, creating a smaller but arguably more reliable graph, more suitable as a self-diagnosis tool. Thus, the same HKG can be contextually altered for different use-cases.

Figure 11a: The full "Combined" HKG visualised in Gephi

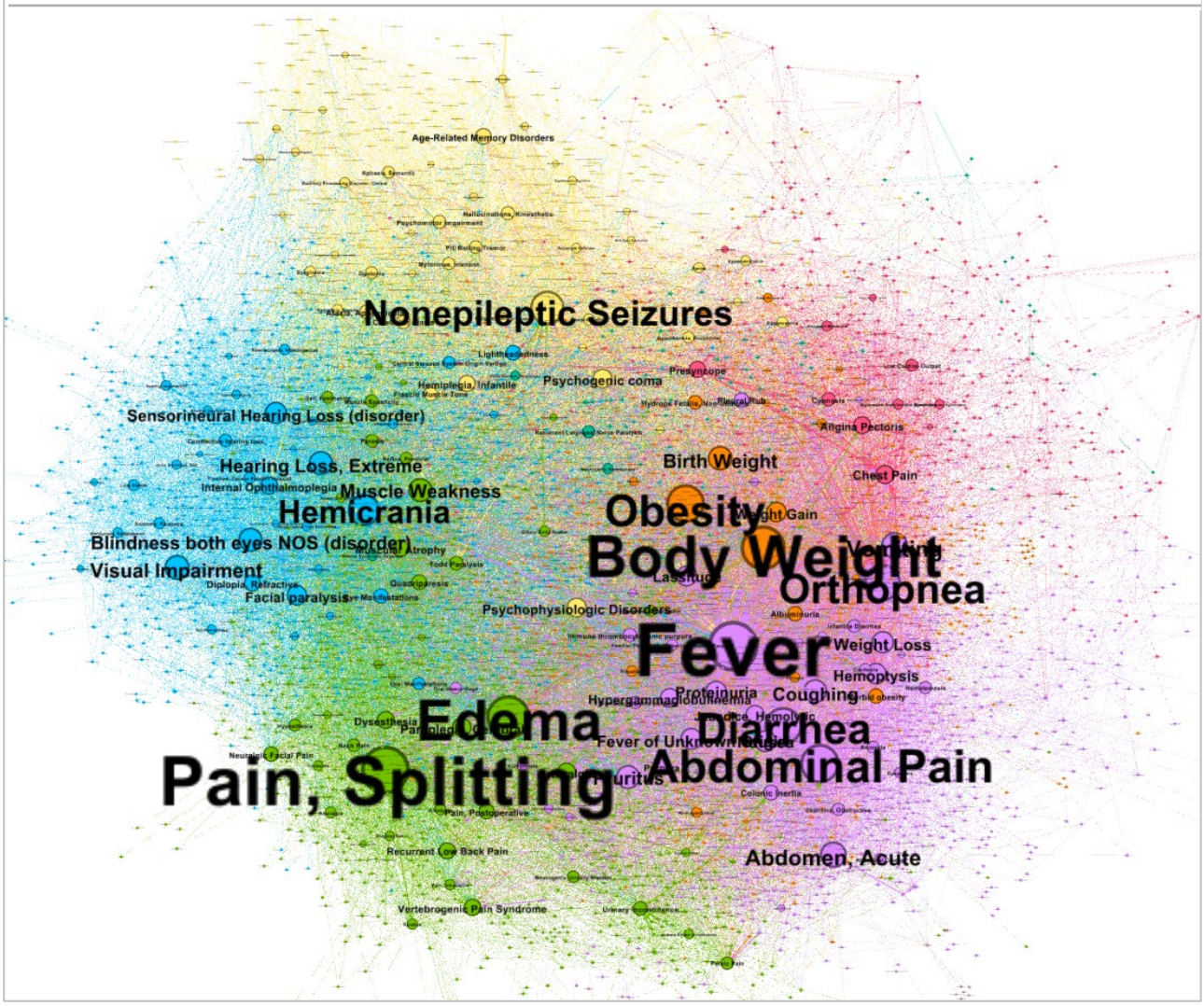
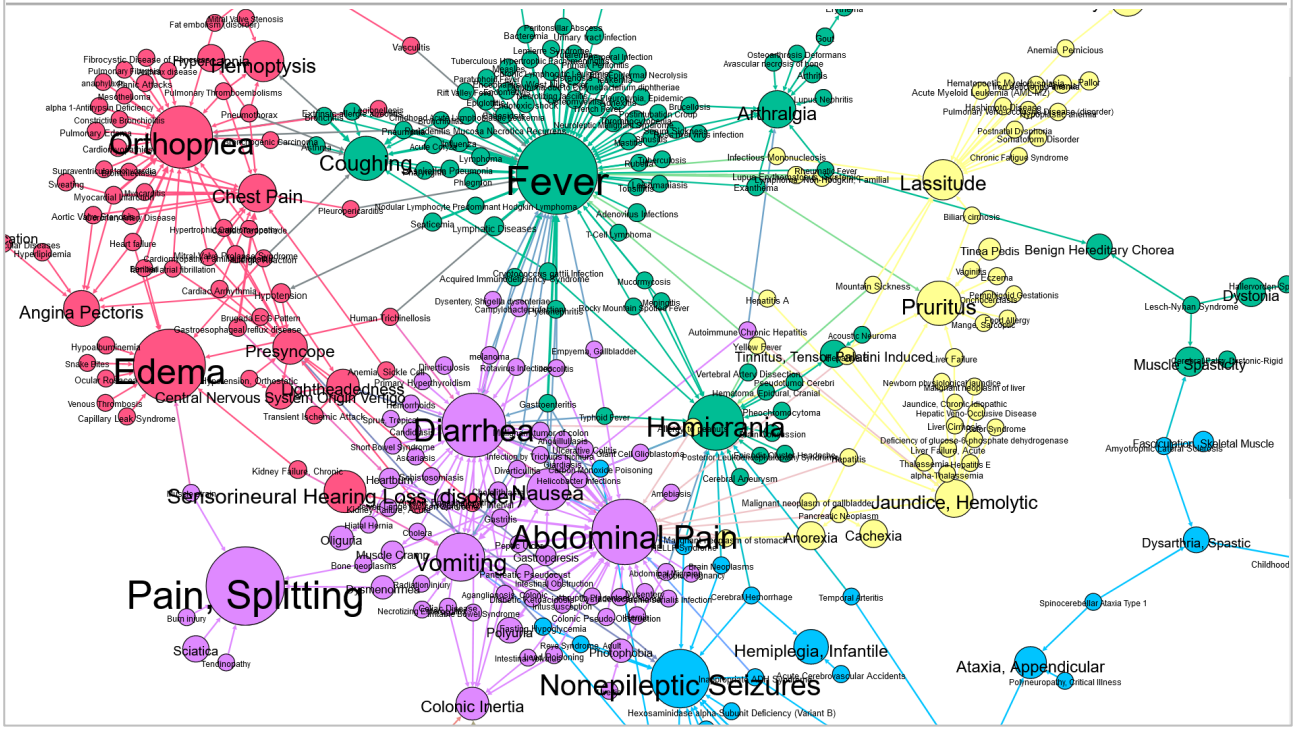


Figure 11b: The filtered "Combined" HKG visualised in Gephi



[Fig12] details the recall/precision metrics of the HKGs, revealing that the datasets share few edges in common, e.g., only ~10% of the associations in the Medical Records HKG also exist in the PubMed HKG. This indicates to the usefulness of combining the datasets, as they complement each other; the medical records capture emergency and common real-world diseases, whereas the literature describes more rare and chronic conditions.

Figure 12: A recall matrix indicating what % of associations from the "base" HKG also exist in the "comparison" HKG

		Base		
		PubMed	DBPedia	Medical Records
Comparison	PubMed	100.0%	57.25%	10.25%
	DBPedia	0.49%	100.0%	2.1%
	Medical Records	0.14%	3.37%	100.0%

Figure 13: A SPARQL query of the MeSH RDF API returning the more broad category of "Hearing Loss" for the specific disease "Hearing Loss, Sudden"

```
SELECT ?diseaseLabel ?treeNum
       ?ancestorTreeNum ?ancestorTreeLabel
FROM <https://id.nlm.nih.gov/mesh/sparql>
WHERE {
mesh:D003639 rdfs:label ?diseaseLabel .
mesh:D003639 meshv:treeNumber ?treeNum .
?treeNum meshv:parentTreeNumber ?ancestorTreeNum .
?meshCode meshv:treeNumber ?ancestorTreeNum  .
}
```

diseaseLabel	Hearing Loss, Sudden
treeNum	http://id.nlm.nih.gov/mesh/C09.218.458.341.900
ancestorTreeNum	http://id.nlm.nih.gov/mesh/C09.218.458.341
ancestorTreeLabel	Hearing Loss

However, the lack of shared associations may also be due to different physicians subjectively classifying the same condition slightly differently. Entity standardisation is a key problem with linked data, therefore we explored replacing specific conditions with broader categories [Fig13] to help alleviate this problem; however this method does cause some information loss.

In conclusion, our project demonstrated a proof-of-concept for creating an up-to-date, comprehensive medical knowledge graph that can be easily updated with new datasets, even from very heterogenous sources. Expert physicians would be needed to evaluate the HKGs' correctness however. Further research would include developing a more sophisticated approach to normalising and combining the datasets, including a bias for prior known reliable sources.

Character count: 6997 characters (with spaces).

References

[1] Chen, E. S., Hripcsak, G., Xu, H., Markatou, M., & Friedman, C. (2008). *Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study.* Journal of the American Medical Informatics Association, 15(1), 87-98.

[2] Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., & Liu, Y. (2020). *Real-world data medical knowledge graph: construction and applications.* Artificial intelligence in medicine, 103.

[3] Finlayson, S. G., LePendu, P., & Shah, N. H. (2014). *Building the graph of medicine from millions of clinical narratives.* Scientific data, 1(1), 1-9.

[4] Chen, I. Y., Agrawal, M., Horng, S., & Sontag, D. (2019). *Robustly extracting medical knowledge from EHRs: a case study of learning a health knowledge graph.* Pacific Symposium on Biocomputing 2020 (pp. 19-30).

[5] Cao, H., Markatou, M., Melton, G. B., Chiang, M. F., & Hripcsak, G. (2005). *Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics.* AMIA Annual Symposium Proceedings (Vol. 2005, p. 106).

[6] Shen, Y., Zhang, L., Zhang, J., Yang, M., Tang, B., Li, Y., & Lei, K. (2018). *Constructing a clinical Bayesian network based on data from the electronic medical record.* Journal of biomedical informatics, 88, 1-10.

[7] Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). *Learning a health knowledge graph from electronic medical records.* Scientific reports, 7(1), 1-11.

[8] Zhou, X., Menche, J., Barabási, AL., & Sharma, A. (2014). *Human symptoms–disease network: Supplementary Data 3.* https://static-content.springer.com/esm/art%3A10.1038%2Fncomms5212/MediaObjects/41467_2014_BFncomms5212_MOESM1045_ESM.txt

[9] Zhou, X., Menche, J., Barabási, A. L., & Sharma, A. (2014). *Human symptoms–disease network.* Nature communications, 5(1), 4212.

[10] Friedman, C. (2008). *Disease-Symptom Knowledge Database.* https://people.dbmi.columbia.edu/%7Efriedma/Projects/DiseaseSymptomKB/index.html

[11] Wang, X., Chused, A., Elhadad, N., Friedman, C., & Markatou, M. (2008). *Automated knowledge acquisition from clinical narrative reports.* AMIA Annual Symposium Proceedings (Vol. 2008, p. 783).

[12] *DBpedia.* (2023). https://www.dbpedia.org/

[13] Bushman, B., Anderson, D., & Fu, G. (2015). *Transforming the medical subject headings into linked data: creating the authorized version of MeSH in RDF.* Journal of library metadata, 15(3-4), 157-176.

[14] Bodenreider, O. (2004). *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Res. 2004 Jan 1;32 D267-70.

[15] Bastian, M., Heymann, S., & Jacomy, M. (2009, March). *Gephi: an open source software for exploring and manipulating networks.* Proceedings of the international AAAI conference on web and social media (Vol. 3, No. 1, pp. 361-362).

[16] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). *Fast unfolding of communities in large networks.* Journal of statistical mechanics: theory and experiment. 2008(10).

Zhao, C., Jiang, J., Guan, Y., Guo, X., & He, B. (2018). *EMR-based medical knowledge representation and inference via Markov random fields and distributed representation learning.* Artificial intelligence in medicine, 87, 49-59.

Shi, L., Li, S., Yang, X., Qi, J., Pan, G., & Zhou, B. (2017). *Semantic health knowledge graph: semantic integration of heterogeneous medical knowledge and services.* BioMed research international, 2017.

Yu, H. Q. (2019). *Mining Symptom and Disease Web Data with NLP and Open Linked Data.* The 5th World Congress on Electrical Engineering and Computer Systems and Science.

Datasets

All datasets were chosen due to their having FAIR properties, containing a substantial amount of workable data, and being open access:

- The "PubMed" dataset represented a data source of formal medical literature. It is released under the Creative Commons Attribution 4.0 International license, and was retrieved on the 17th December 2022 from: https://static-content.springer.com/esm/art%3A10.1038%2Fncomms5212/MediaObjects/41467_2014_BFncomms5212_MOESM1045_ESM.txt

- The "Medical Records" dataset represented a data source of real-world medical cases in a hospital. The licencing information does not provide a specific licence, but simply states: *"Open Access: verbatim copying and redistribution are permitted in all media for any purpose."* The dataset was retrieved on the 18th December 2022, from: https://people.dbmi.columbia.edu/%7Efriedma/Projects/DiseaseSymptomKB/index.html

- The "DBpedia" dataset represented a data source of manually curated medical information. It is released under the Creative Commons Attribution-ShareAlike 3.0 and GNU Free Documentation licences, and was retrieved on the 16th December 2022 from: https://dbpedia.org/

Code

The project code is openly available on GitHub: https://github.com/Natasha-R/DS_Linked_Open_Data_and_Knowledge_Graphs_2022_Natasha_Randall

All code was written in Python within fully documented, commented Jupyter notebooks, released under the MIT licence. The notebooks have the following structure:

Importing, processing and cleaning the three datasets:
1. DBpedia Dataset.ipynb
2. PubMed Dataset.ipynb
3. Medical Records Dataset.ipynb

A supplemental exploration of methods to fix missing MeSH codes and an approach to represent conditions by their "parent category":
4. Example Methods to Fix Discrepancies.ipynb

Combining the three datasets and comparing them:
5. Combining Datasets.ipynb
6. Comparing the Health Knowledge Graphs.ipynb

Software

All the software used are free with open access:

- DBpedia SPARQL endpoint: https://dbpedia.org/sparql/
- Gastrodon: https://github.com/paulhoule/gastrodon
- MeSH RDF API: https://id.nlm.nih.gov/mesh/
- MeSH Browser: https://meshb.nlm.nih.gov/search
- UMLS API: https://documentation.uts.nlm.nih.gov/rest/home.html
- Gephi: https://gephi.org/

The following statement is required in research making use of the National Library of Medicine data:

This product uses publicly available data from the U.S. National Library of Medicine (NLM), National Institutes of Health, Department of Health and Human Services; NLM is not responsible for the product and does not endorse or recommend this or any other product.