A Meta-Path-Based Prediction Method for Disease Comorbidities

Eduardo P. García del Valle
ETS Ingenieros Informáticos
Universidad Politécnica de Madrid
Pozuelo de Alarcón, Madrid, Spain
ep.garcia@alumnos.upm.es
ORCID: 0000-0002-3897-1684

Lucía Prieto Santamaría
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Pozuelo de Alarcón, Madrid, Spain
lucia.prieto.santamaria@alumnos.upm.es
ORCID: 0000-0003-1545-3515

Gerardo Lagunes García
Centro de Tecnología Biomédica
Universidad Politécnica de Madrid
Pozuelo de Alarcón, Madrid, Spain
gerardo.lagunes@ctb.upm.es
ORCID: 0000-0002-6780-675X

Massimiliano Zanin
Instituto de Física Interdisciplinar y Sistemas
Complejos IFISC (CSIC-UIB), Campus UIB,
Palma de Mallorca, Spain
massimiliano.zanin@upm.es
ORCID: 0000-0002-5839-0393

Ernestina Menasalvas Ruiz
Centro de Tecnología Biomédica, ETS Ingenieros Informáticos
Universidad Politécnica de Madrid
Madrid, Spain
ernestina.menasalvas@upm.es
ORCID: 0000-0002-5615-6798

Alejandro Rodríguez-González
Centro de Tecnología Biomédica, ETS Ingenieros Informáticos
Universidad Politécnica de Madrid
Madrid, Spain
alejandro.rg@upm.es
ORCID: 0000-0001-8801-4762

Abstract— **The simultaneous presence of diseases worsens the prognosis of patients and makes their treatment difficult. Identifying the co-occurrence of diseases is key to improving the situation of patients and designing effective therapeutic strategies. On the one hand, the increasing availability of clinical information opens new ways to unveil hidden relationships between diseases. On the other hand, heterogeneous information networks have been used in recent years to discover novel knowledge from disease data, including symptoms, genes or drugs. The use of meta-paths allows the complex semantics of the relationships between the different types of nodes to be included in heterogeneous networks. In this study, we propose a system to predict disease comorbidities through the use of meta-paths in a heterogeneous network of diseases and symptoms, built from textual sources of public access. The results obtained improve those of similar studies based on biological data, and the predictions calculated for diabetes and Crohn's disease are supported by medical literature. Both the used data and the obtained prediction model are publicly accessible.**

Keywords: disease comorbidity, heterogeneous disease networks, meta-paths, medical text mining, graph structure learning

I. INTRODUCTION

The occurrence of one or more additional conditions, known as comorbidity, is widespread among the patients admitted at multidiscipline hospitals. For instance, obese patients often develop type-2 diabetes and hypertension. A number of clinical studies show that disease comorbidity not only causes additional suffering to patients, but also compromises the success of standard treatments compared to patients who have a single disease. In the US, 80% of Medicare spending is dedicated to treating patients with multiple coexisting conditions [1]. For this reason, the accurate prediction of potential disease comorbidities is essential to design more efficient treatment strategies and improve the prognosis of patients.

In recent years, the increasing availability of clinical data has boosted the investigation of unknown relationships between diseases. Given the variety of sources and data, heterogeneous information networks have become a crucial tool for extracting novel knowledge [2], [3]. The identification of new disease-disease relationships using link prediction methods has not only improved our understanding of their etiology and pathogenesis, but has also made it possible to reuse existing treatments in new diseases [4]. Meta-paths, sequences of semantic relationships between nodes of

heterogeneous networks, provide a powerful mechanism for the training of link prediction models [5]. For example, two diseases can be connected via *disease-gene-disease* path, *disease-gene-compound-drug-disease* path, and so on. Intuitively, the semantics underneath different paths imply different similarities. Formally, a meta-path P is a path defined on the graph of network schema $T_G = (A, R)$ and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$ between type A_1 and A_{l+1}, where $\circ$ denotes the composition operator on relations [6].

The use of meta-path often involves a two-step process to solve the link prediction problem in heterogeneous networks. In the first step, the meta-path-based feature vectors are extracted. In the second step, a regression or classification model is trained to compute the existence probability of a link. For example, Sun et al. proposed *PathPredict* to solve the problem of co-author relationship prediction following this approach [7]. In [8], Dong et al. present the *Metapath2Vec* model to maximize the likelihood of preserving both the structure and semantics of a given heterogeneous network and apply its latent embeddings to various network mining tasks, such as node classification, clustering, and link prediction. In contrast to conventional meta-path-based methods, the advantage of latent-space representation learning lies in its ability to model similarities between nodes that are not connected through meta-paths. Recent studies have used heterogeneous networks and meta-paths for the prediction of comorbidities from biological data. Jin et al. built a *miRNA–gene–disease* network to uncover microRNA-mediated disease comorbidities and potential pathobiological implications [9]. Their method presented an accuracy, measured with the area under the curve of the Receiver Operating Characteristic (AUC-ROC), of 0.65 when inferring the clinically reported disease–disease pairs.

Despite the growing number of clinical texts and their potential as a source of new knowledge, their exploitation in the prediction of comorbidities through heterogeneous networks is limited, partly due to limited access to electronic health records imposed by privacy laws. In this paper, we present a method for predicting comorbidities from public clinical data, based on meta-paths. First, we built a heterogeneous network of diseases and symptoms, and defined the meta-paths. Next, we applied the Metapath2Vec model to tackle link prediction as a supervised learning problem on top of the network embeddings. The AUC-ROC obtained when evaluating the model was 0.74. Finally, we applied the prediction model to type-2 diabetes and Crohn's disease, and found that the results were supported by the medical literature. Figure 1 summarizes the methods schematically. Both the data used and the results obtained are published as supplementary materials, for their validation and reuse[1].

II. METHODS

A. Heterogeneous disease-symptom network

We extracted data on associations between diseases and symptoms from DISNET, a database that integrates phenotypic characteristics of diseases from Wikipedia,

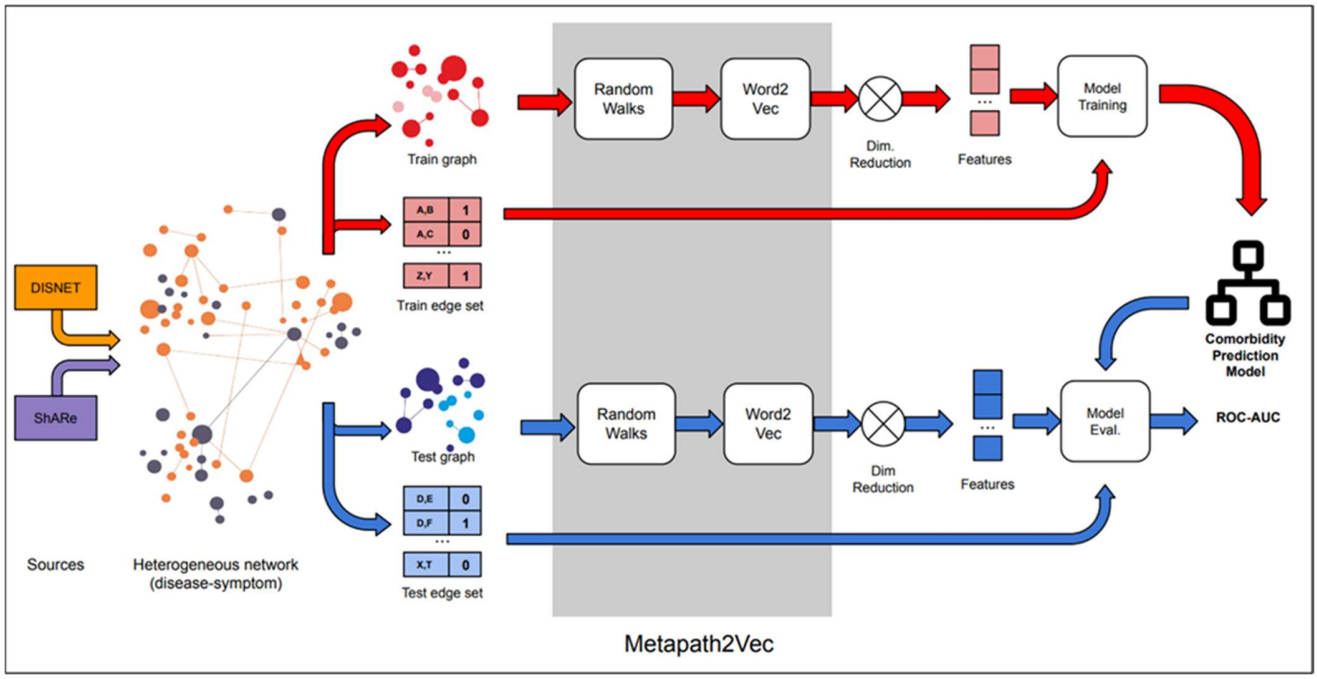

Figure 1. Visual summary of the methods applied to generate the comorbidity prediction model described in this paper.

[1] https://github.com/pantapps/cbms21

PubMed and MayoClinic, among others [10]. DISNET snapshot 2020-12-15 contains 7,193 diseases associated with 2,103 different symptoms. To extract the disease-disease relationships based on their co-occurrence in the same patient, we used the ShARe corpus published in SemEval / CLEF 2013–2015 evaluations, which contains 300 clinical notes with 12,095 annotated disorders and their attributes [11].

To connect the data from both sources, we used the Search API of the Unified Medical Language System (UMLS) to map the cross-referenced identifiers in DISNET to their Concept Unique Identifier (CUI) [12]. On the one hand, we only included DISNET diseases with a mapping in UMLS. On the other hand, we only selected diseases from the ShARe corpus that contain symptoms in DISNET.

Finally, we used the Stellargraph python library to build the heterogeneous network. Of the total of 5,147 nodes, 3,251 had disease type and 1,896 had symptom type. The 49,741 links were annotated as *disease-has_symptom-symptom* (46,333) and *disease-has_cooccurrence-disease* (3,408), according to their nature.

B. Link prediction model

We used *Metapath2Vec* to learn the embeddings, maximizing the likelihood of preserving both the structure and semantics of the heterogeneous network [8]. First, we split our network into a training graph and a test graph. From each graph, we set aside a sample (10%) of positive and negative edges into a training edge set and a test edge set, respectively. Negative edges are sampled at random by selecting two nodes in the graph and then checking if these edges are connected or not. If not, the pair of nodes is considered a negative sample. Otherwise, it is discarded and the process repeats.

Second, we applied uniform random walks to traverse the training graph and generate a corpus of sentences. A sentence is a list of node IDs, and each node ID is considered a unique word in a dictionary that has size equal to the number of nodes in the graph. The random walk is driven by meta-paths that define the node type order by which the random walker explores the graph. For example, the meta-path *disease-symptom-disease* defines a rule for the random walk to traverse the graph starting from a disease node, passing through a symptom node to end on a disease node. All meta-paths begin and end on disease type nodes. Figure 2 shows the node and edge types, and the meta-path schema applied for our random walk.

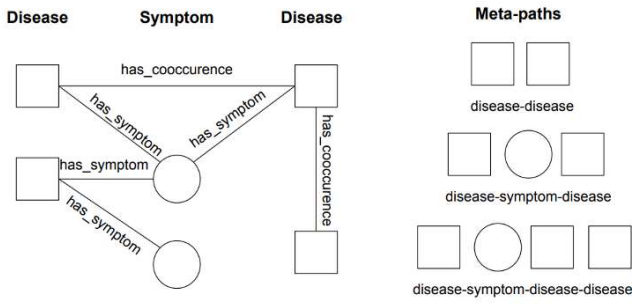

Figure 2. Extraction of the meta-path schema passed to the random walk algorithm to traverse the heterogeneous disease-symptom network.

Third, we fed the sentence corpus into a *Word2Vec* model to calculate an embedding vector for each node in the graph. Given a word (node ID), Word2Vec uses the skip-gram algorithm to predict the neighboring words within a specified window. This model gives more importance to words closer to the target word than to the distant ones [13].

Then we applied element-wise multiplication (Hadamard product) on the embeddings of the source and target nodes to calculate edge embeddings for positive and negative edge samples from the training edge set [14]. Finally, we trained a logistic regression classifier with the edge embeddings to predict a binary value indicating whether an edge between two nodes is expected to exist or not.

The heterogeneous network edge list and the trained model are available in the supplementary materials.

C. Model evaluation

To evaluate our predictor, we used the test graph to compute test node embeddings, and then computed AUC-ROC using the test edge set. In order to qualitatively evaluate its performance, we applied our model to predict the comorbidities of type-2 diabetes mellitus and Crohn's disease, and we contrasted the results with data available in the clinical literature [15]–[18].

III. RESULTS

The computed comorbidity prediction model showed an AUC-ROC=0.74. Figure 3 represents the AUC-ROC visually.

One of the advantages of using node embeddings in our approach is the possibility of representing the heterogeneous network in a low dimensional space, in which the graph structural information and graph properties are maximumly preserved. We used the t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the embeddings computed for the nodes and edges, by giving each datapoint a location in a two-dimensional map [19]. Figure 4 shows the t-SNE projection for node embeddings (A) and edge embeddings (B).

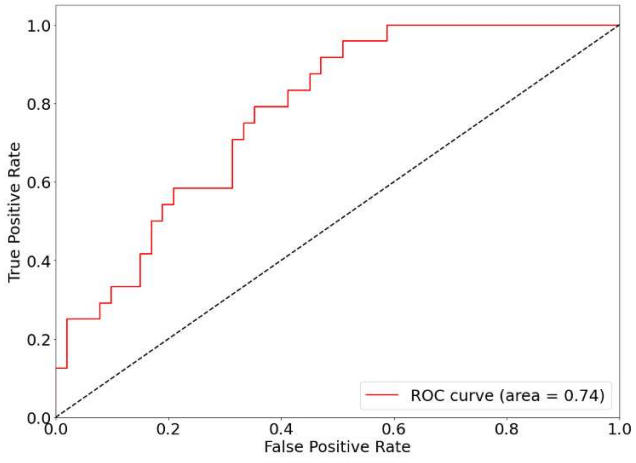

Figure 3. Area under the curve of the receiver operating characteristic obtained during the evaluation of the comorbidity prediction model.

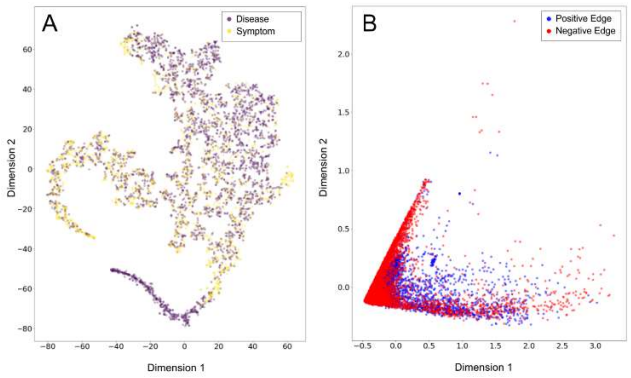

Figure 4. T-SNE 2D projection of the network embeddings. A) Embeddings of diseases (purple) and symptom (yellow) nodes; B) Embeddings of positive (blue) and negative (red) edges.

Table I and Table II contain the top 20 predicted disease-disease links (comorbidities) for type-2 diabetes mellitus and Crohn's disease, respectively. In the tables, diseases are sorted by the probability of no co-occurrence (P_0 column) in descending order.

An extended version with the top 100 predicted links is available in the supplementary materials.

IV. DISCUSSION

Results show that the presented method allows predicting co-occurrences between diseases from public data on symptoms and diseases, with reasonable accuracy (see Figure 3). The AUC-ROC of our model significantly improves that obtained by Jin et. by applying meta-paths to miRNA data, gene and proteins instead of symptom [9]. However, it is still lower than that of other more advanced models [20].

TABLE I. TOP 20 COMORBIDITIES PREDICTED FOR TYPE-2 DIABETES MELLITUS. P0 IS THE PROBABILITY THAT THE DISEASES ARE NOT CO-OCCURRENT.

UMLS CUI	Disease Name	P_0
C0029408	Degenerative polyarthritis	2.10e-11
C0032320	Pneumoperitoneum	4.36e-11
C0009938	Mitral Valve Insufficiency	8.40e-11
C0011581	Depressive disorder	1.31e-10
C0016169	Urethral Stenosis	2.16e-10
C0027543	Avascular necrosis of bone	2.26e-10
C0026266	Chronic Kidney Insufficiency	2.86e-10
C0003507	Aortic Valve Stenosis	5.55e-10
C0001418	Adenocarcinoma	6.74e-10
C0024633	Diabetic Retinopathy	8.27e-10
C0006826	Malignant Neoplasms	8.99e-10
C0262414	Acute Kidney Tubular Necrosis	1.03e-09
C0264912	White Coat Hypertension	1.05e-09
C1261287	Stenosis	1.91e-09
C0751523	Corn of toe	1.93e-09
C0021308	Infarction	4.51e-09
C0040961	Tricuspid Valve Insufficiency	5.15e-09
C0003855	Arteriovenous fistula	5.64e-09
C0011881	Diabetic Nephropathy	6.47e-09

TABLE II. TOP 20 COMORBIDITIES PREDICTED FOR CROHN'S DISEASE. P0 IS THE PROBABILITY THAT THE DISEASES ARE NOT CO-OCCURRENT.

UMLS CUI	Disease Name	P_0
C0016169	Pathologic Fistula	2.41e-06
C0032320	Pneumoperitoneum	5.24e-06
C0024633	Mallory-Weiss Syndrome	9.95e-06
C0011881	Multiple Sclerosis	1.18e-05
C0027543	Avascular necrosis of bone	1.20e-05
C0029408	Degenerative polyarthritis	1.61e-05
C0011581	Depressive disorder	2.22e-05
C0009938	Malignant tumor of colon	2.57e-05
C0040961	Tricuspid Valve Insufficiency	2.60e-05
C0003855	Arteriovenous fistula	3.88e-05
C0013481	Ebstein Anomaly	4.72e-05
C0021308	Infarction	5.81e-05
C0026266	Mitral Valve Insufficiency	6.01e-05
C0009324	Ulcerative Colitis	6.04e-05
C0156272	Enterovesical Fistula	6.76e-05
C0003507	Aortic Valve Stenosis	8.60e-05
C0019326	Ventral Hernia	1.39e-04
C0014175	Endometriosis	1.47e-04
C0264912	Left anterior fascicular block	1.74e-04

When applying the model to type-2 diabetes mellitus (see Table I), we obtained results that coincide with the most common comorbidities reported in the clinical literature, such as hypertension, chronic kidney diseases, cardiovascular diseases and visual problems [15], [16]. Other cases, such as degenerative polyarthritis, pneumoperitoneum, avascular necrosis of bone or corn of toe are not among the most common comorbidities, but are reported in the medical literature [21]–[24]. The extended results show numerous co-occurrences of diabetes with fractures (e.g., fracture of cervical spine, fracture of second cervical vertebra, rib fractures). The relationship between diabetes and bone fragility has also been studied [25].

In the case of Crohn's disease, the most common comorbidities are intestinal diseases (colon cancer, rectal cancer), respiratory diseases, vascular diseases, and arthritis. The results shown in Table II include diseases of these types [17], [18]. As in the case of diabetes, we find very specific cases such as pathologic fistula, Mallory-Weiss Syndrome and multiple sclerosis, described in the clinical literature [26], [27].

Notwithstanding the aforementioned results, our study presents some limitations. On the one hand, the number of diseases with a significant probability of comorbidity (> 0.95) is high, representing 17.32% and 11.05% for type-2 diabetes mellitus and Crohn's disease, respectively. This suggests that the classification is not specific enough. On the other hand, the data set contains common and/or unspecified diseases such as carcinoma, cancer or vitamin deficiency, which could affect the results. A pre-filtering of the data set to eliminate these types of entries could potentially improve the specificity of the system.

V. CONCLUSIONS

Improving our knowledge about disease comorbidities can improve the treatment of patients, saving not only suffering but also healthcare resources. In this paper, we propose the exploitation of data from open clinical texts through a meta-path-based network analysis to predict the probability of co-occurrence of two diseases. Both the used data and the obtained results are publicly available.

The main advantage of our approach is its good complexity-performance ratio. Methods based on meta-paths with random walks are intuitive and simple, describing the relationships between data in a semantic and interpretable way. However, they are less powerful than more complex methods, such as those based on graph neural networks (GNNs). GNNs are able to incorporate both latent and explicit features of the graph, demonstrating state-of-the-art performance on numerous problems, including link prediction [28].

As future work, we propose to apply methods based on GNNs to the prediction of comorbidities from textual data and compare the results with those obtained in the present study, considering the complexity-performance relationship.

ACKNOWLEDGMENT

The work is a result of the project "DISNET (Creation and analysis of disease networks for drug repurposing from heterogeneous data sources applied to rare diseases)", that is being developed under grant "RTI2018-094576-A-I00" from the Spanish Ministerio de Ciencia, Inovación y Universidades. Lucía Prieto Santamaría's work is supported by "Programa de fomento de la investigación y la innovación (Doctorados Industriales") from Comunidad de Madrid (grant IND2019/TIC-17159). Gerardo Lagunes-Garcia's work is supported by the Mexican Consejo Nacional de Ciencia y Tecnología (CONACYT) (CVU: 340523) under the programme "291114 - BECAS CONACYT AL EXTRANJERO". Massimiliano Zanin's work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 851255); and the Spanish State Research Agency, through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (MDM-2017-0711).

REFERENCES

[1] X. Ji, S. Ae Chun, and J. Geller, "Predicting Comorbid Conditions and Trajectories using Social Health Records," *IEEE Trans Nanobioscience*, vol. 15, no. 4, pp. 371–379, Jun. 2016, doi: 10.1109/TNB.2016.2564299.

[2] E. P. García del Valle, G. Lagunes García, L. Prieto Santamaría, M. Zanin, E. Menasalvas Ruiz, and A. Rodríguez-González, "Disease networks and their contribution to disease understanding: A review of their evolution, techniques and data sources," *Journal of Biomedical Informatics*, vol. 94, p. 103206, Jun. 2019, doi: 10.1016/j.jbi.2019.103206.

[3] J. Han, "Mining heterogeneous information networks: the next frontier," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, Aug. 2012, pp. 2–3, doi: 10.1145/2339530.2339533.

[4] D. S. Himmelstein *et al.*, "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *eLife*, vol. 6, p. e26726, Sep. 2017, doi: 10.7554/eLife.26726.

[5] X. Chen, M.-X. Liu, and G.-Y. Yan, "Drug-target interaction prediction by random walk on the heterogeneous network," *Mol Biosyst*, vol. 8, no. 7, pp. 1970–1978, Jul. 2012, doi: 10.1039/c2mb00002d.

[6] Y. Sun and J. Han, "Meta-path-based search and mining in heterogeneous information networks," *Tinshhua Sci. Technol.*, vol. 18, no. 4, pp. 329–338, Aug. 2013, doi: 10.1109/TST.2013.6574671.

[7] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks," *Proc. VLDB Endow.*, 2011, doi: 10.14778/3402707.3402736.

[8] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable Representation Learning for Heterogeneous Networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2017, pp. 135–144, doi: 10.1145/3097983.3098036.

[9] S. Jin *et al.*, "A network-based approach to uncover microRNA-mediated disease comorbidities and potential pathobiological implications," *npj Systems Biology and Applications*, vol. 5, no. 1, Art. no. 1, Nov. 2019, doi: 10.1038/s41540-019-0115-2.

[10] G. Lagunes García, A. González, L. Prieto Santamaría, E. García del Valle, M. Zanin, and E. Menasalvas, "DISNET: a framework for extracting phenotypic disease information from public sources," *PeerJ*, vol. 8, p. e8580, Feb. 2020, doi: 10.7717/peerj.8580.

[11] S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, and G. Savova, *SemEval-2014 Task 7: Analysis of Clinical Text*. 2014, p. 62.

[12] "Searching the UMLS." https://documentation.uts.nlm.nih.gov/rest/search/ (accessed Mar. 04, 2021).

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, Red Hook, NY, USA, Dec. 2013, pp. 3111–3119, Accessed: Apr. 20, 2021. [Online].

[14] P. R. Halmos, *Finite-Dimensional Vector Spaces: Second Edition*. Courier Dover Publications, 2017.

[15] J. Alwakeel *et al.*, "Diabetes Complications in 1952 Type 2 Diabetes Mellitus Patients Managed in a Single Institution," *Annals of Saudi Medicine*, vol. 28, pp. 260–266, Jul. 2008, doi: 10.5144/0256-4947.2008.260.

[16] H. F. Jelinek *et al.*, "Clinical profiles, comorbidities and complications of type 2 diabetes mellitus in patients from United Arab Emirates," *BMJ Open Diabetes Research and Care*, vol. 5, no. 1, p. e000427, Aug. 2017, doi: 10.1136/bmjdrc-2017-000427.

[17] C. Bernstein and A. Nabalamba, "Hospitalization-Based Major Comorbidity of Inflammatory Bowel Disease in Canada," *Canadian journal of gastroenterology = Journal canadien de gastroenterologie*, vol. 21, pp. 507–11, Sep. 2007, doi: 10.1155/2007/924257.

[18] C. N. Bernstein, A. Wajda, and J. F. Blanchard, "The Clustering of Other Chronic Inflammatory Diseases in Inflammatory Bowel Disease: A Population-Based Study," *Gastroenterology*, vol. 129, no. 3, pp. 827–836, Sep. 2005, doi: 10.1053/j.gastro.2005.06.021.

[19] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, Jan. 2002, pp. 857–864, Accessed: Apr. 20, 2021. [Online].

[20] P. Akram and L. Liao, "Prediction of comorbid diseases using weighted geometric embedding of human interactome," *BMC Medical Genomics*, vol. 12, no. 7, p. 161, Dec. 2019, doi: 10.1186/s12920-019-0605-5.

[21] F. Aiello *et al.*, "Molecular Links Between Diabetes and Osteoarthritis: The Role of Physical Activity.," *Current diabetes reviews*, 2017, doi: 10.2174/1573399812666151123104352.

[22] X. Xu, Y. Gong, Y. Zhang, J. Lang, and Y. Huang, "Effect of pneumoperitoneum pressure and the depth of neuromuscular block on renal function in patients with diabetes undergoing laparoscopic pelvic surgery: Study protocol for a double-blinded 2 × 2 factorial randomized controlled trial," *Trials*, vol. 21, Jun. 2020, doi: 10.1186/s13063-020-04477-x.

[23] A. Dima, A. B. Pedersen, L. Pedersen, C. Baicus, and R. W. Thomsen, "Association of common comorbidities with osteonecrosis: a nationwide population-based case–control study in Denmark," *BMJ Open*, vol. 8, no. 2, p. e020680, Feb. 2018, doi: 10.1136/bmjopen-2017-020680.

[24] "Diabetes Foot Problems Symptoms, Treatment & Complications," *MedicineNet*. https://www.medicinenet.com/foot_problems_diabetes/article.htm (accessed Mar. 04, 2021).

[25] K. F. Moseley, "Type 2 diabetes and bone fractures," *Curr Opin Endocrinol Diabetes Obes*, vol. 19, no. 2, pp. 128–135, Apr. 2012, doi: 10.1097/MED.0b013e328350a6e1.

[26] G. Bislenghi *et al.*, "P061 The molecular landscape of perianal fistula in Crohn's disease: opportunities for new therapeutic approaches," *Journal of Crohn's and Colitis*, vol. 14, no. Supplement_1, pp. S165–S165, Jan. 2020, doi: 10.1093/ecco-jcc/jjz203.190.

[27] R. Andre, P. Emonts, R. Kiekens, and M. Cremer, "[Ulcer of the colon and Mallory-Weiss syndrome (author's transl)]," *Acta Chir Belg*, vol. 75, no. 4, pp. 473–480, Jul. 1976.

[28] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Dec. 2018, pp. 5171–5181, Accessed: Apr. 20, 2021. [Online].