Review Study of Available Datasets Related to Disease-Disease Association

Qaiser Mehmood

PhD Student

Bioinformatics Final Project

Introduction

Many diseases have association with other diseases and often co-occur among the patients [31]. It is important to understand and analyze the association among different diseases for proper diagnostics and treatment of similar patients. In this study we focus on studying and analyzing different approaches and datasets used in literature for disease-disease association. We tend to study the available datasets, and the nature of data available in such datasets. Also, we tend to study how a particular dataset is used for disease-disease association and how it is compiled. Also, it is important to investigate that how such datasets are prepared and what other purposes they serve other than disease-disease association.

There are many studies related to disease-disease associations in literature in different perspectives, however, as to best of our knowledge, we could not find an evidence in recent studies that discuss the datasets in a way we propose in this document.

Problem Description

In literature, there are many datasets used for disease-disease association. These vary from case to case and different perspectives of disease-disease association are discussed in such studies. For example Disease Ontology (DO) [32], The Guideline Advantage (TGA) dataset [33] International Classification of Diseases (ICT) dataset [34] and other related data sets are under consideration.

As to best of our knowledge, there is no study in literature that compares these three datasets in perspective of the usage by other relevant literature in perspective of citations in recent years. We in tend to compare the datasets to investigate how these are used by other researchers for their studies and which of these datasets is used for what type of diseases mostly by the researchers.

This study will help the future perspective researchers to help make optimum decision for selection of datasets for their study as used by most of the researchers.

Methodology

In order to formulate the research, we used ‘survey’ research methodology by conducting a scientific survey of the datasets usage as described in literature.

We searched the literature using ‘scholar.google.com’ to find out how recently a chosen dataset is used in the literature. We used the official source of citation provided by the official website for Disease Ontology (DO) dataset and International Classification of Disease (ICD) dataset. However The Guideline Advantage (TGA) is not publicly available, therefore we used the quoted name in google scholar to analyze the results. Table 1. Shows the numbers obtained by this method.

The parameters we chosen were the ‘year of publication’ and ‘number’ of publications that use a particular dataset. We used these parameters because we want to investigate that how recently and frequently a particular dataset is used for a scientific publication and how the dataset is used by the researchers.

Our initial search showed that there were different numbers of citations for these three datasets during the course of different years. Table 1. Shows that number of citations per dataset during a particular time range.

	Number of Citations	Years Range
Disease Ontology (DO)	216	2018-2022
International Classification of diseases (ICD) *ICD-11 for Mortality and Morbidity Statistics*	379	2018-2022
The Guideline Advantage (TGA)	101	2014-2022

Table 1: Comparison of the Datasets in terms of Number of Citations

In this table, it is indicated that International Classification of Disease (ICD) was mostly used dataset in 379 scientific citations during the years 2018-2022. During the same interval of time, Disease Ontology (DO) was second highly used dataset. However, the usage of The Guideline Advantage (TGA) is very less comparatively. Since 2014 up to 2022, during 8 years, the dataset was used only by 101 researchers for their publications, as shown in Figure 1.

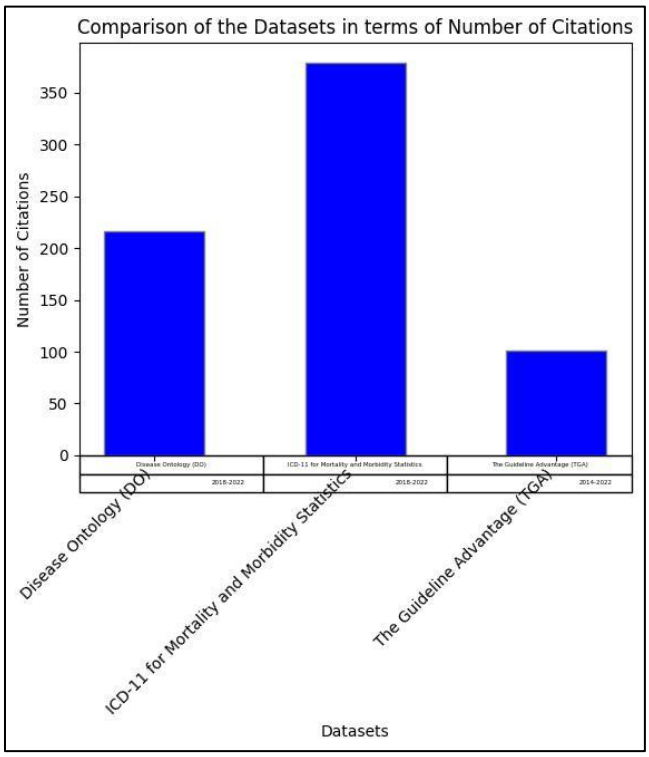

Figure 1 Comparison of the Datasets in terms of Number of Citations

Table 1 gives a rough idea of the usage of the datasets, however it does not tells us that what publications were titled and how those used these datasets. For example, 216 publications used DO but what were the titles of those publications and how these used this dataset is a crucial information that we need to investigate to get some trends in the dataset usage.

We might get best results if we use study all the publications for the dataset usage, but due to time limitations, we studied 10 most recent publication, resulted by google scholar, for studying the usage of relevant datasets. Table 2, 3, and 4 shows the relevant publications titles, the year of publication and the usage of dataset for each dataset.

	Publication Title	Year published	Cited 'DO' for; (Usage)
1	A functional cellular framework for sex and estrous cycle-dependent gene expression and behavior [1]	Jan-2022	Studying Sex hormones behaviors
2	How an Addiction Ontology can Unify Competing Conceptualizations of Addiction [2]	Jan-2022	Studying Addiction
3	Prediction of lncRNA-disease association based on a Laplace normalized random walk with restart algorithm on heterogeneous networks [3]	Jan-2022	Studying non-coding RNA-Diseases
4	A machine learning-driven approach for prioritizing food contact chemicals of carcinogenic concern based on complementary in silico methods [4]	Jan-2022	DO terms are used for feature encoding
5	Exploring Sentiment as a Potential Indicator of Bias in Disease Ontologies [5]	Dec-2021	Studying Bias based on sentiment
6	Ontology-based annotation and retrieval for large-scale VCF data [6]	Dec-2021	Studying as literature review for research gap
7	A network-based drug repurposing method via non-negative matrix factorization [7]	Dec-2021	Studying drug re-purposing
8	Etiology context of rare diseases in the Human Disease Ontology [8]	Nov-2021	Studying causes (Etiology) of rare diseases.
9	Development of a chemogenomics library for phenotypic screening [9]	Nov-2021	Preparing a chemogenomics Library
10	In silico Methods for Identification of Potential Therapeutic Targets [10]	Nov-2021	Studying Therapeutic diseases

Table 2: Disease Ontology (DO) Dataset

Table 2 shows the latest 10 publications, as sorted by google scholar, with their publication year and the reason of being cited, as indicated by the publication.

	Publication Title	Year published	Cited 'TGA' for; (Usage)

1	Discovering disease–disease associations using electronic health records in The Guideline Advantage (TGA) dataset [11]	2021	Disease-disease association
2.	Predicting cardiovascular health trajectories in time-series electronic health records with LSTM models [12]	2021	Studying cardiovascular diseases
3	Application of a time-series deep learning model to predict cardiac dysrhythmias in electronic health records [13]	2021	Studying cardiovascular diseases
4	Women and ethnoracial minorities with poor cardiovascular health measures associated with a higher risk of developing mood disorder [14]	2021	Studying cardiovascular diseases
5	Dissemination of a telehealth cardiovascular risk service: The CVRS live protocol [15]	2021	Studying cardiovascular diseases
6	A cluster randomized trial to evaluate a centralized remote clinical pharmacy service in large, health system primary care clinics [16]	2021	Studying clinical pharmacy service
7	Time-series cardiovascular risk factors and receipt of screening for breast, cervical, and colon cancer: The Guideline Advantage [17]	2020	Studying cardiovascular diseases
8	Abstract P177: Hypertension is Considerably More Prevalent, Yet Decreasing Over Time According to the New Guidelines [17]	2019	Studying Hypertension
9	Predicting Cardiovascular Health Trajectories in Time-series Electronic Health Records With Deep Learning [18]	2019	Studying cardiovascular diseases
10	Cardiovascular Health Trends in Electronic Health Record Data (2012–2015): A Cross-Sectional Analysis of The Guideline Advantage™ [20]	2019	Studying cardiovascular diseases

Table 3: The Guideline Advantage (TGA)

Similarly, we can view that data shown in Table 3 and Table 4.

	Publication Title	Year published	Cited ‘ICT’ for; (Usage)
1	Drug-induced poisoning during pregnancy: Four-year experience [21]	2022	Studying Drug-induced poisoning during pregnancy:
2	Validation of ICD-10-AM Coding for Myocardial Infarction Subtype in Hospitalization Data [22]	2022	Studying Myocardial Infarction Subtype

3	Drug Use and Brain Injury in Kentucky Acute Care Facilities in 2020 [23]	2022	Studying brain Injury
4	A Statewide Analysis of Pediatric Liver Injuries Treated at Adult Versus Pediatric Trauma Centers [24]	2022	Studying Liver Injuries
5	Congenital lung malformation patients experience respiratory infections after resection: A population-based cohort study [25]	2022	Studying respiratory infection
6	The Influence of Obesity Hypoventilation Syndrome on the Outcomes of Patients with Diabetic Ketoacidosis: An Analysis of National Inpatient Sample. [26]	2022	Studying patients of diabetic ketoacidosis
7	Evidence for an Inherited Contribution to Sepsis Susceptibility Among a Cohort of US Veterans [27]	2022	Studying Sepsis Susceptibility
8	Do Patient Demographic and Socioeconomic Factors Influence Surgical Treatment Rates After ACL Injury? [28]	2022	Studying Factors influencing surgical treatment
9	International Classification of Disease based Injury Severity Score (ICISS): a data linkage study of hospital and death data in Victoria, Australia [29]	2022	Studying disease based injuries
10	Frailty and In-Hospital Mortality Risk Using EHR Nursing Data [30]	2022	Studying in-hospital morality rates.

Table 4: International Classification of Diseases (ICT) dataset

Results and Analysis

The data shown in all the tables gives us a unique idea that how a particular datasets is used in literature by the scientists. We analyze the data by discussing each table one by one.

Table 2 refers to Disease Ontology (DO) dataset and indicates that out of latest 10 relevant publications, 4 are in year 2022 and 6 are in year 2021. The usage of this dataset is very diverse and does not indicates any consistency in these 10 publications in terms of the purpose of use.

However, Table 3, which refers to The Guideline Advantage (TGA) dataset, indicates 7 out of 10 publication that discuss the cardiovascular diseases. This indicates that most of the researchers, studying cardiovascular diseases tends to use this dataset for their research. But out of most recent 10 publications using TGA, 3 are in year 2019. This shows that very few researchers are able to access the dataset for their research as the dataset is not publically available. This also shows that if a dataset is not publically available, it cannot be used by most of the researchers.

Most interesting results are shown by Table 4. Not only all the top 10 publications are from this month (January 2022) but also the usage of dataset is for dealing with diseases mostly. Almost all the publication use this dataset for studying different injuries and diseases as shown in Table 4.

We compiled all the outcomes of Table 2, 3 and 4 to Table 5 to show what results we obtained in terms of the parameters. The different parameters that we discussed in this section are set in the table to obtain the data in the numerical form. Table shows that ICT is the best dataset in terms of the diversity of number of diseases and usage of the datasets for most of the scientists. However, for cardiovascular diseases, TGA dataset is used by the researchers in recent years, comparatively.

	Total Number of diseases discussed in 10 studies	Number of publications in 2022	Number of cardiovascular diseases discussed in 10 studies	Number of injuries based diseases discussed in 10 studies	Number of other diseases discussed in 10 studies	Number of publication discussing disease-disease association
ICD	7	10	0	3	4	0
OD	3	4	0	0	3	0
TGA	8	0	7	0	1	1

Table 5: Results with respect to data obtained

Conclusion and future work

Although we were only able to study 10 publications per dataset for the purpose of this research, however, the results show some trends towards what datasets is being mostly used recently and what datasets is used for particular disease, as shown in Figure 2.

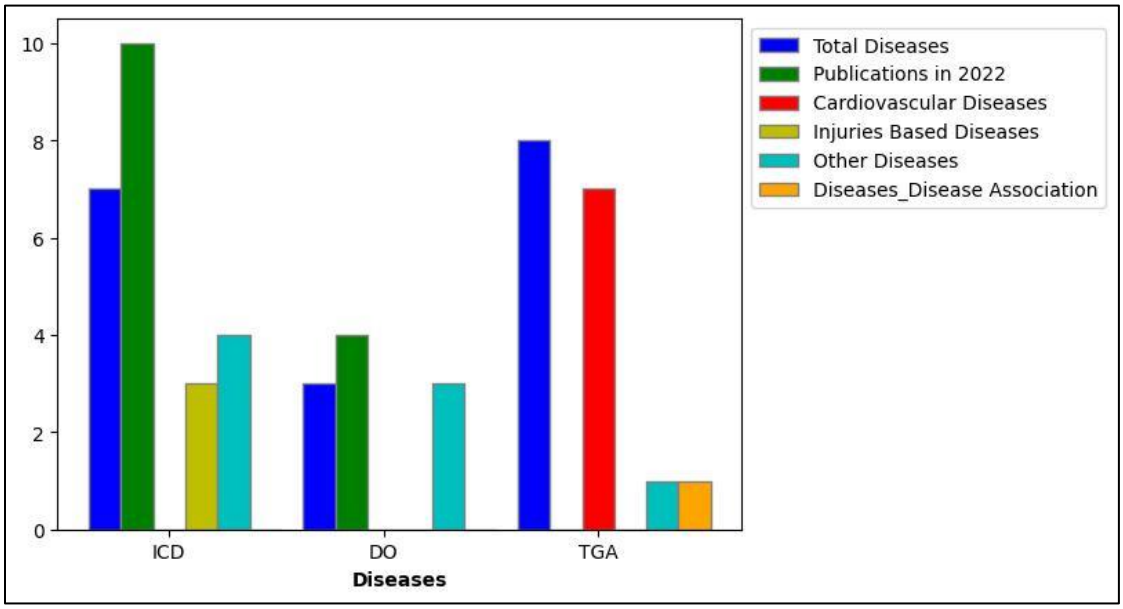

Figure 2 Results with respect to data obtained

So, we conclude that ICD and TGA are the datasets that actually are used by the researchers to discuss some particular diseases, However, DO is mostly used for research and literature review purposes. Cardiovascular diseases research papers we studied mostly use TGA dataset but this dataset is not publically available. So limited researchers used and cited TGA comparatively. ICD is the most frequently and recently used dataset and this study shows that it is used for studying multiple diseases and most of the researchers. So we can conclude that ICD, so far is the most usable, datasets overall for research in the multiple disease studies.

The time limitation for this project compelled us to use less number of research papers for study. But in future, if more numbers of research papers are studied, more accurate results might be obtained.

References

[1] Knoedler, J. R., Inoue, S., Bayless, D. W., Yang, T., Tantry, A., Davis, C. H., ... & Shah, N. M. (2022). A functional cellular framework for sex and estrous cycle-dependent gene expression and behavior. *Cell*.

[2] Kelly, R. M., Hastings, J., & West, R. (2022). How an Addiction Ontology can Unify Competing Conceptualizations of Addiction. In *Evaluating the Brain Disease Model of Addiction* (pp. 484-496). Routledge.

[3] Wang, L., Shang, M., Dai, Q., & He, P. A. (2022). Prediction of lncRNA-disease association based on a Laplace normalized random walk with restart algorithm on heterogeneous networks. *BMC bioinformatics*, *23*(1), 1-20.

[4] Wang, C. C., Liang, Y. C., Wang, S. S., Lin, P., & Tung, C. W. (2022). A machine learning-driven approach for prioritizing food contact chemicals of carcinogenic concern based on complementary in silico methods. *Food and Chemical Toxicology*, 112802.

[5] Slater, L. T., Williams, J. A., Schofield, P. N., & Gkoutos, G. V. (2021, December). Exploring Sentiment as a Potential Indicator of Bias in Disease Ontologies. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1826-1834). IEEE.

[6] Liu, J., Qu, Z., Li, Y., Sun, J., & Liu, Y. (2021, December). Ontology-based annotation and retrieval for large-scale VCF data. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 478-483). IEEE.

[7] Sadeghi, S., Lu, J., & Ngom, A. (2021). A network-based drug repurposing method via non-negative matrix factorization. *Bioinformatics (Oxford, England)*, btab826.

[8] Schriml, L. M., & Baron, J. A. (2021, December). Etiology context of rare diseases in the Human Disease Ontology. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2606-2608). IEEE Computer Society.

[9] Bryan, D., Natacha, C., Batiste, B., Anaelle, C., Ducrot, P., Thierry, D., ... & Olivier, T. (2021). Development of a chemogenomics library for phenotypic screening. *Journal of Cheminformatics*, *13*(1).

[10] Zhang, X., Wu, F., Yang, N., Zhan, X., Liao, J., Mai, S., & Huang, Z. (2021). In silico Methods for Identification of Potential Therapeutic Targets. *Interdisciplinary Sciences: Computational Life Sciences*, 1-26.

[11] Guo, A., Khan, Y. M., Langabeer, J. R., & Foraker, R. E. (2021). Discovering disease–disease associations using electronic health records in The Guideline Advantage (TGA) dataset. *Scientific reports*, *11*(1), 1-10.

[12] Guo, A., Beheshti, R., Khan, Y. M., Langabeer, J. R., & Foraker, R. E. (2021). Predicting cardiovascular health trajectories in time-series electronic health records with LSTM models. *BMC Medical Informatics and Decision Making*, *21*(1), 1-10.

[13] Guo, A., Smith, S., Khan, Y. M., Langabeer II, J. R., & Foraker, R. E. (2021). Application of a time-series deep learning model to predict cardiac dysrhythmias in electronic health records. *PloS one*, *16*(9), e0239007.

[14] Kennelty, K. A., Engblom, N. J., Carter, B. L., Hollingworth, L., Levy, B. T., Finkelstein, R. J., ... & Dorsey, K. K. (2021). Dissemination of a telehealth cardiovascular risk service: The CVRS live protocol. *Contemporary Clinical Trials*, *102*, 106282.

[15] Kennelty, K. A., Engblom, N. J., Carter, B. L., Hollingworth, L., Levy, B. T., Finkelstein, R. J., ... & Dorsey, K. K. (2021). Dissemination of a telehealth cardiovascular risk service: The CVRS live protocol. *Contemporary Clinical Trials*, *102*, 106282.

[16] Kennelty, K. A., Coffey, C. S., Ardery, G., Uribe, L., Yankey, J., Ecklund, D., ... & MEDication Focused Outpatient Care for Underutilization of Secondary Prevention (MEDFOCUS) trial investigators. (2021). A cluster randomized trial to evaluate a centralized remote clinical pharmacy service in large, health system primary care clinics. *Journal of the American College of Clinical Pharmacy*, *4*(10), 1287-1299.

[17] Guo, A., Drake, B. F., Khan, Y. M., Langabeer Ii, J. R., & Foraker, R. E. (2020). Time-series cardiovascular risk factors and receipt of screening for breast, cervical, and colon cancer: The Guideline Advantage. *PloS one*, *15*(8), e0236836.

[18] Rudy, J. E., Foraker, R. E., Harris, R. E., & Bower, J. K. (2019). Abstract P177: Hypertension is Considerably More Prevalent, Yet Decreasing Over Time According to the New Guidelines. *Circulation*, *139*(Suppl_1), AP177-AP177.

[19] Guo, A., & Foraker, R. E. (2019). Predicting Cardiovascular Health Trajectories in Time-series Electronic Health Records With Deep Learning. *Circulation*, *140*(Suppl_1), A11005-A11005.

[20] Rudy, J. E., Khan, Y., Bower, J. K., Patel, S., & Foraker, R. E. (2019). Cardiovascular Health Trends in Electronic Health Record Data (2012–2015): A Cross-Sectional Analysis of The Guideline Advantage™. *eGEMs*, *7*(1).

[21] Sert, Z. S., & Menekse, T. S. (2022). Drug-induced poisoning during pregnancy: Four-year experience. *The American Journal of Emergency Medicine*.

[22] Nedkoff, L., Lopez, D., Hung, J., Knuiman, M., Briffa, T. G., Murray, K., ... & Sanfilippo, F. M. (2022). Validation of ICD-10-AM Coding for Myocardial Infarction Subtype in Hospitalisation Data. *Heart, Lung and Circulation*.

[23] Bush, G. (2022). Drug Use and Brain Injury in Kentucky Acute Care Facilities in 2020.

[24] Pulido, O. R., Morgan, M. E., Bradburn, E., & Perea, L. L. (2022). A Statewide Analysis of Pediatric Liver Injuries Treated at Adult Versus Pediatric Trauma Centers. *Journal of Surgical Research*, *272*, 184-189.

[25] Markel, M., Derraugh, G., Lacher, M., Iqbal, S., Balshaw, R., Min, S. A. L., & Keijzer, R. (2022). Congenital lung malformation patients experience respiratory infections after resection: A population-based cohort study. *Journal of Pediatric Surgery*.

[26] Pattipati, M., & Gudavalli, G. The Influence of Obesity Hypoventilation Syndrome on the Outcomes of Patients with Diabetic Ketoacidosis: An Analysis of National Inpatient Sample. *Available at SSRN 4000534*.

[27] Kempker, J. A., Martin, G. S., Rondina, M. T., & Cannon-Albright, L. A. (2022). Evidence for an Inherited Contribution to Sepsis Susceptibility Among a Cohort of US Veterans. *Critical Care Explorations*, *4*(1).

[28] Testa, E. J., Modest, J. M., Brodeur, P., Lemme, N. J., Gil, J. A., & Cruz, A. I. (2022). Do Patient Demographic and Socioeconomic Factors Influence Surgical Treatment Rates After ACL Injury?. *Journal of racial and ethnic health disparities*, 1-6.

[29] Berecki-Gisolf, J., Fernando, D. T., & D'Elia, A. (2022). International Classification of Disease based Injury Severity Score (ICISS): a data linkage study of hospital and death data in Victoria, Australia. *Injury*.

[30] Lekan, D., McCoy, T. P., Jenkins, M., Mohanty, S., & Manda, P. (2021). Frailty and In-Hospital Mortality Risk Using EHR Nursing Data. *Biological research for nursing*, 10998004211060541.

[31]] Guo, A., Khan, Y. M., Langabeer, J. R., & Foraker, R. E. (2021). Discovering disease–disease associations using electronic health records in The Guideline Advantage (TGA) dataset. *Scientific reports*, *11*(1), 1-10.

[32] Disease Ontology Database. https://disease-ontology.org/.

[33] Bufalino, V. et al. Evolution of “The Guideline Advantage”: Lessons learned from the front lines of outpatient performance measurement. CA Cancer J. Clin. 64(3), 157–163 (2014).

[34] World Health Organization. (2018). ICD-11 for mortality and morbidity statistics (2018).