DisMaNET: A network-based tool to cross map disease vocabularies

Eduardo P. García del Valle[1], Gerardo Lagunes García[1,2], Lucía Prieto Santamaría[2], Massimiliano Zanin[3], Ernestina Menasalvas Ruiz[1,2], Alejandro Rodríguez-González[1,2].

[1] ETS de Ingenieros Informáticos. Universidad Politécnica de Madrid. Boadilla del Monte, Madrid, Spain.

[2] Centro de Tecnología Biomédica, ETS Ingenieros Informáticos. Universidad Politécnica de Madrid. Pozuelo de Alarcón, Madrid, Spain.

[3] Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Campus UIB, Palma de Mallorca, Spain.

Corresponding autor:

Eduardo P. García del Valle

Centro de Tecnología Biomédica.

Campus de Montegancedo. Pozuelo de Alarcón, 28223, Madrid.

Email: ep.garcia@alumnos.upm.es, Phone: +34 913364663

Abstract

Background and Objectives

The growing integration of healthcare sources is improving our understanding of diseases. Cross-mapping resources such as UMLS play a very important role in this area, but their coverage is still incomplete. With the aim to facilitate the integration and interoperability of biological, clinical and literary sources in studies of diseases, we built DisMaNET, a system to cross-map terms from disease vocabularies by leveraging the power and intuitiveness of network analysis.

Methods

First, we collected and normalized data from 8 disease vocabularies and mapping sources to generate our datasets. Next, we built DisMaNET by integrating the generated datasets into a Neo4j graph database. Then we exploited the query mechanisms of Neo4j to cross-map disease terms of different vocabularies with a relevance score metric and contrasted the results with some state-of-the-art solutions. Finally, we made our system publicly available for its exploitation and evaluation both through a graphical user interface and REST APIs.

Results

DisMaNET contains almost half a million nodes and near nine hundred thousand edges, including hierarchical and mapping relationships. Its query capabilities enabled the detection of connections between disease vocabularies that are not present in major mapping sources such as UMLS and the Disease Ontology, even for rare diseases. Furthermore, DisMaNET was capable of obtaining more than 80% of the mappings with UMLS reported in MonDO and DisGeNET. Our tool was used successfully to complete the missing mappings in DISNET, a web-based system designed to extract knowledge from signs and symptoms retrieved from medical databases.

Conclusions

DisMaNET is a powerful, intuitive and publicly available system to cross-map terms from different disease vocabularies. Its completeness and the potential of network analysis make it a competitive alternative to existing mapping systems. Expansion with new sources, versioning and the improvement of the search and scoring algorithms are envisioned as future lines of work.

Keywords

Disease vocabularies, data interoperability, disease cross-mapping, graph database, network analysis

1. Introduction

The increasing availability of large-scale biological, clinical and literary databases combined with the advances in computational methods are contributing to improve our understanding of diseases. However, the integration and interoperability of these sources pose a major challenge, particularly due to the use of different vocabularies and codifications of diseases [1]. This variety is explained because each vocabulary was originally created to meet a specific need. International Statistical Classification of Diseases (ICD) codes, for instance, are used by doctors, health insurance companies, and public health agencies across the world to represent diagnoses. In contrast, the Medical Subject Headings (MeSH) vocabulary is especially employed for the purpose of indexing journal articles and books in life sciences. A newer alternative to ICD and MeSH is the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT). It cross-maps to several revisions of ICD and has a considerably broader scope than just diseases. Other widely used disease classification systems, although of more specific use, are OMIM (genetic disorders), Orphanet (rare diseases) or NCI (carcinogenic diseases) [2–4].

For years, disease classifications evolved independently, making their interrelation difficult. However, the growing number of disease studies based on the integration of multiple biological and literary sources entailed the need to cross-map them. One of the most notable efforts in this direction is the Unified Medical Language System (UMLS). Created in 1986 and maintained by the National Library of Medicine (NLM), UMLS provides through its Metathesaurus a mapping structure of many controlled vocabularies in the biomedical sciences. As a result, numerous studies and tools have used UMLS as an authentic Rosetta stone of disease terms. Still, due to the disparity of scopes and the granularity of the disease vocabularies, the annotation and mapping of UMLS terms is a complex and unfinished task.

In the last decades, several initiatives have tried to improve and complete disease mappings by diverse methods. Already in 1998, a study by Bodenreider et al. proposed to use the semantic relationships between concepts to map terms of different vocabularies in UMLS with MeSH [5]. A later investigation contemplated the use of drug prescriptions to complete the mapping between MeSH and ICD-10-CM terms extracted from the UMLS Metathesaurus [6]. In 2017, Raje et al. leveraged the rich set of synonyms provided by the UMLS to identify lexical mappings for those concepts in the Disease Ontology (DO) without any mappings to SNOMED CT [7]. More recently, the emergence of massive source integration projects has driven the search for solutions to unify concepts. One remarkable example is the Monarch Merged Disease Ontology (MonDO), created by the Monarch Initiative to integrate multiple human disease resources into a single ontology by using a Bayes merging algorithm [8]. In the same line, MalaCards created a human disease database from existing categorization systems and applied a semantic algorithm to connect diseases from different sources [9].

As an alternative to automated mapping techniques, expert-curated sources can be exploited to complete disease mappings, avoiding error propagation and the need for additional validation [10, 11]. The authors applied this approach in a previous research, with the aim of facilitating the integration of data extracted from PubMed with other sources [12]. The present study continues this line of work, introducing a network-based methodology that allows researchers to cross-map disease terms when integrating heterogeneous sources. Network analysis has proven as an intuitive and powerful method to extract new knowledge out of previously existing information. Over the last decades, network-based methodologies have been applied to discover connections among apparently unrelated biomedical entities such as diseases, physiological processes, signaling pathways, and genes [1]. Following the same reasoning, our study proposes to apply network analysis in the detection of disease term mappings. To illustrate and evaluate this methodology, we developed DisMaNET, a network-based tool to cross map disease vocabularies.

2. Materials and methods

This section describes the implementation of DisMaNET. First, we give a detailed description of the vocabularies and mapping sources used to generate the datasets. Next, we integrate the datasets into a graph database and outline the parameters of the resulting disease term network. Then we exploit network analysis mechanisms to obtain mappings between disease terms in different vocabularies and contrast the results with other mapping projects. Finally, we explain how to access DisMaNET. Figure 1 summarizes the process.

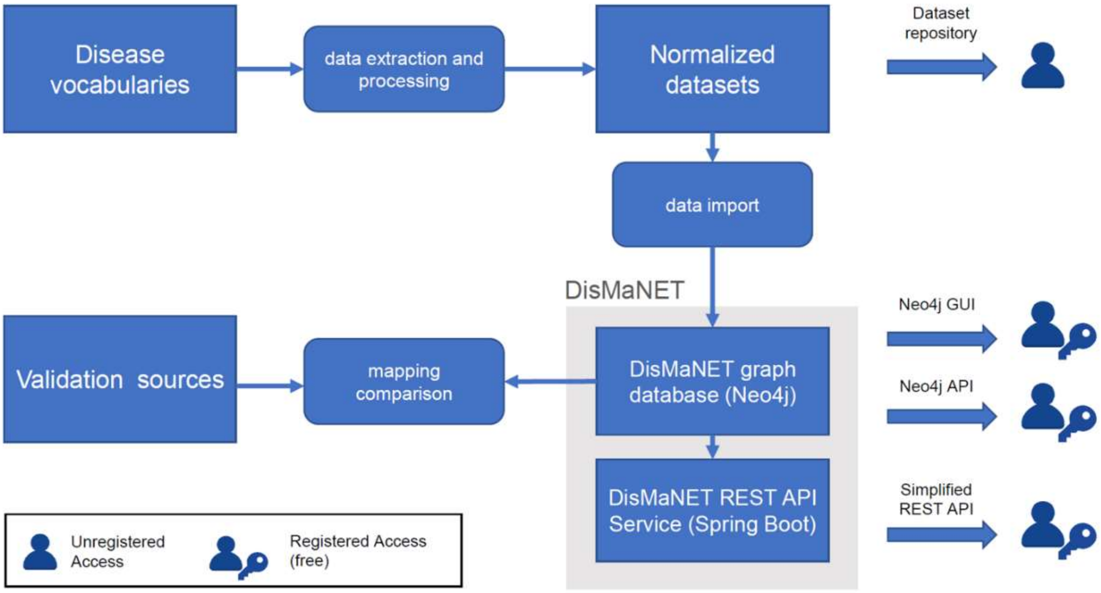

Figure 1. Block diagram with the implementation of DisMaNET.

2.1. Data sources

As a preliminary step to building our mapping tool, we collected data from several disease vocabularies. For the sake of comparability with previous results, the 2018 version of the data was used, when available. MeSH files were downloaded from NLM[1]. For descriptors, we considered only terms under categories C (Diseases) and F03 (Mental Disorders) [13–15]. To facilitate their connection with other vocabularies, we also included MeSH SCR of Class 3 (diseases). SNOMED CT (United States) 2018 version was downloaded from NLM[2]. In this case, we only considered concepts under the "Clinical Finding" top-level hierarchy. As for ICD-10-CM, the files containing the 2018 version of the code descriptions for this vocabulary were downloaded from the website of the Centers for Medicare & Medicaid Services[3]. In order to extend the set of disease vocabularies used in our previous research, for the present study we incorporated the OMIM[4] and Orphanet[5] as data sources. Both datasets were downloaded in June 2019. The UMLS Metathesaurus (version 2018AB) was used as the main source of disease code mappings. To obtain the UMLS concepts associated as synonyms with the terms of the other vocabularies in the study (i.e. share the same unique identifier in the Metathesaurus), we exploited the Search REST API of the UMLS Terminology Services[6]. The map of SNOMED CT and ICD-

[1] ftp://nlmpubs.nlm.nih.gov/online/mesh/2018/xmlmesh
[2] https://www.nlm.nih.gov/healthit/snomedct/archive.html
[3] https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs.html
[4] https://www.omim.org/downloads
[5] http://www.orphadata.org/cgi-bin/rare_free.html
[6] https://documentation.uts.nlm.nih.gov/rest/home.html

10-CM codes (September 2008 version) was downloaded from NLM[7]. Finally, the concepts of the DO, which contain mappings to 24 disease vocabularies, were downloaded from the code repository of the project[8] (version tag v2018-03-02).

Of the eight mined sources, four of them contain relationships with other vocabularies: UMLS, DO, Orphanet and SNOMED CT. For this reason, throughout the study we will refer to them as "mapping sources". Their mapping information will be helpful to connect the terms across different vocabularies. Additionally, in order to relate terms within each vocabulary, we also obtained data about their hierarchy, when available. Figure 2 depicts the data sources used in the study and the connections among them. Vocabularies and mapping sources provide information in different formats, ranging from XML to OBO or CSV. Therefore, in addition to the data extraction, as part of the dataset generation we had to standardize the format to CSV. The resulting dataset is publicly available in the project repository, where additional information on the data structure is provided[9].

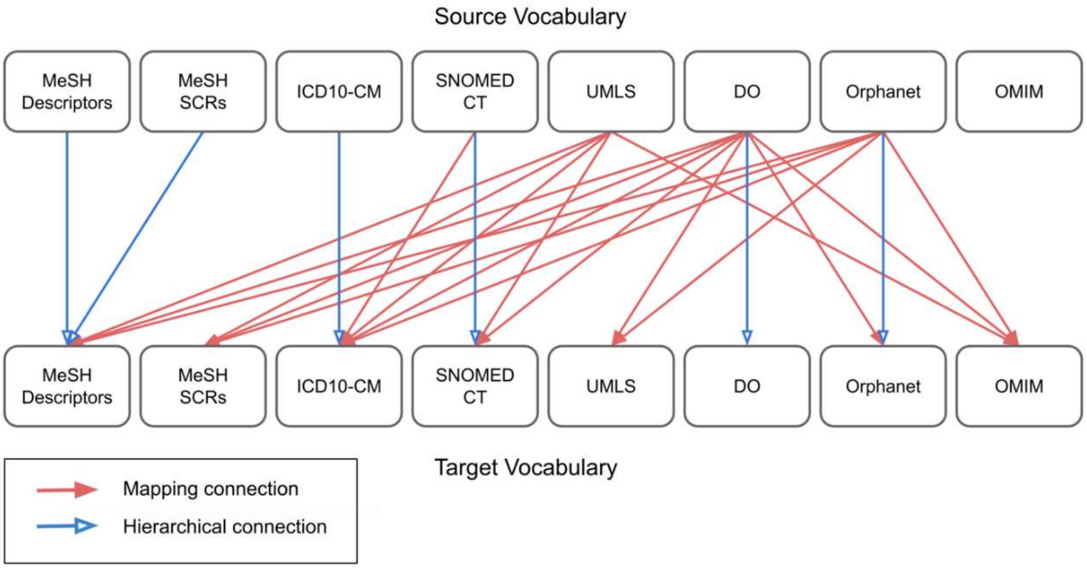

Figure 2. Data sources used in the study and the connections between them. The arrow indicates the direction of the relationship. For example, UMLS contains mappings to the SNOMED CT vocabulary, and SNOMED CT contains mappings to ICD-10-CM (via the IHTSDO project). MeSH descriptors point to their parent MeSH descriptor through a hierarchical connection.

2.2. Building DisMaNET

The next step to build our network-based mapping tool was to integrate the collected data into a graph database (GDB). Relationships can be intuitively visualized using GDBs, making them

[7] https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html
[8] https://github.com/DiseaseOntology/HumanDiseaseOntology
[9] https://github.com/dismanet/paper/tree/master/data

useful for heavily interconnected data [16]. Additionally, GDBs are often faster than relational databases for associative data sets and map more directly to the structure of object-oriented applications [17]. To implement our GDB of disease vocabularies we chose Neo4j[10] (v3.4.7), an open source solution which has been extensively used in network-based studies on diseases [18–20].

For each disease vocabulary, we imported its concepts into Neo4j as nodes, labeled with the vocabulary name. Each node contains the name and the ID of the disease, as unique properties. The connections between the diseases were imported as relationships with type MAP (for disease mappings across different vocabularies) or IS_A (for hierarchical links within the same vocabulary). Associations of MeSH descriptors and SCR were represented with the relation type HAS_DESCRIPTOR. All the relationships are directional: MAP points from the mapping source to the vocabulary term; IS_A, from the descendant to the ancestor; and HAS_DESCRIPTOR, from the SCR to the descriptor.

2.3. Querying DisMaNET

Once the collected datasets were integrated into the graph database, we leveraged network analysis to discover new mappings between diseases. By definition, two terms in DisMaNET are directly connected (distance 1) through a MAP type relationship if at least one of them belongs to a mapping source that contains this association. Alternatively, two terms that do not belong to any mapping sources may be connected with distance 2 through a third term of a mapping source. Up to this point, all mappings found through the analysis of our network are known, as there is at least one mapping source that contains these associations. However, by increasing the distance we can detect further indirect relationships, which involve the connection of two terms through two or more additional terms (distance 3 or more). These associations do not exist expressly in any mapping source, and therefore constitute new knowledge. Figure 3 illustrates this concept with an example.

[10] https://neo4j.com/

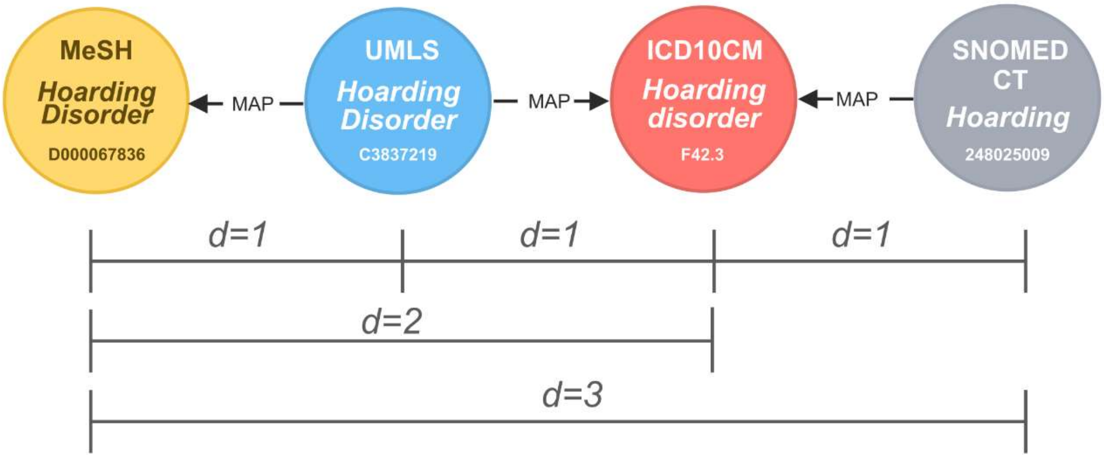

Figure 3. Finding new mappings of MeSH descriptors by exploring MAP relationships in the network. The distance (d) between two nodes connected via MAP relationships is 1 when (at least) one of the concepts belongs to a mapping source (e.g. UMLS maps C3837219 "Hoarding Disorder" with MESH D00067836 "Hoarding Disorder); d = 2 when two nodes are connected through a mapping source, but they might not belong to a mapping source themselves (e.g. the map between MESH and ICD-10-CM F42.3 "Hoarding disorder" is provided by UMLS); d = 3 when two nodes are connected through two sources (e.g. MeSH and SNOMED CT 248025009 "Hoarding").

We used Cypher[11], the graph query language of Neo4j, to find these indirect relationships and fill in the gaps in the mapping sources. For instance, the following query retrieves all related MeSH and SNOMED CT disease codes and names through a maximum of 3 MAP connections (i.e. distance 3), which are not listed in the UMLS mapping source:

```
MATCH (origin:MeSH)-[:MAP*1..3]-(target:SNOMEDCT)
WHERE NOT(origin:MeSH)-[:MAP]-(:UMLS)-[:MAP]-(:SNOMEDCT)
return distinct(origin.diseaseId), origin.diseaseName, target.diseaseId, target.diseaseName;
```

Likewise, to find any mappings between a specific term in a vocabulary with another source through the fewest relationships, we used Dijkstra's Shortest Path algorithm, included in Neo4j. For example, the query below returns the connections of the MeSH term "Hoarding Disorder" with any concept of SNOMED CT up to a distance of 3, following the path represented in Figure 3:

```
MATCH p=shortestPath((:MeSH{diseaseId:'D000067836'})-[:MAP *1..3]-(:SNOMEDCT))
RETURN *
```

These queries are easily generalizable, by adjusting the maximum distance value and replacing the labels of the origin and target vocabularies, and of the mapping source. Section 3 presents the results of applying these queries to all mapping sources.

[11] https://neo4j.com/developer/cypher/

2.4. DisMaNET relevance metric

As previously stated, our disease cross-mapping approach is based on the aggregation of reliable sources and their exploitation through network analysis. While this implies that the mappings obtained by querying DisMaNET have a solid base, for their correct assessment we must consider certain aspects. First, the main drawback of methods based on minimal paths is that their number grows quickly with the size of the system. For large networks, the increase of the number of paths leads to a combinatorial explosion [21]. Another network characteristic that may impact the quality of our approach is the fact that the strength of an indirect tie in a graph decreases with its length [22, 23]. Finally, to evaluate the obtained mappings, we must take into account the disparity in size, granularity and purpose of each source.

In the view of the above considerations, and with the aim of facilitating the assessment of DisMaNET results, we propose the following metric to score the relative relevance of a mapping between terms *a* and *b*:

$$score_{ab} = \frac{1}{2}\left(sim_{ab} + \frac{1}{d_{ab}}\right)$$

In the formula, *sim* is the cosine similarity of the terms, computed as:

$$sim_{ab} = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}}$$

where *A* and *B* are the vector representation of terms a and b, respectively [24]; and *d* is the distance (number of relationships) between the terms. The similarity measures the relevance of the relationship at a semantic level, while the inverse of the distance reflects the strength of the connection at a network level. Thus, a mapping between two nodes with short distance and high similarity will be more relevant in relative terms, compared to other relationships with greater distance and/or less similarity. Section 3 compares the mappings obtained with different thresholds of the score, while Section 4 addresses the interpretability of this metric with some examples.

2.5. Evaluation

In order to evaluate the capacity of our system to cross-map disease codes, we measured how many of the mappings between UMLS identifiers and other vocabularies in MonDO and DisGeNET were obtained through DisMaNET. We downloaded the MonDO ontology (v2019-

10-25) from the Monarch Initiative repository[12]. Only 13,043 out of 23,146 concepts in the ontology contain an association between an UMLS concept and at one term in DisMaNET. As a result, 31,308 mappings between UMLS and other vocabularies in MonDO were used for the evaluation. As for DisGeNET, we obtained the file with the mappings of UMLS CUIs to other disease vocabularies from the project website (v6.0)[13]. DisGeNET contains 22,201 unique CUIs, but only 17,921 have mappings to at least one vocabulary available in DisMaNET. OMIM identifiers starting with MTHU were excluded, as they do not represent exact matches. Overall, 37,916 mappings between UMLS and other vocabularies in DisGeNET were used for the evaluation. Section 3 presents the results of the evaluation.

2.6. Access to DisMaNET

DisMaNET is available for use by any researcher under the DISNET[14] project. Users must previously request a username and password at no cost, to keep a record of usage for statistical purposes. Once registered, they can query DisMaNET either via Neo4j's graphic user interface[15], or through its transactional HTTP API[16]. With the aim of abstracting DisMaNET users from the knowledge of the Cypher query language, we have developed and deployed a Spring Boot service to expose a simplified REST API[17]. Given a code and the label of its vocabulary in DisMaNET, this API returns all the mappings with other vocabularies, sorted by their score in descending order. The service leverages Neo4j Java driver[18] to query the database using the Shortest Path algorithm with MAP and HAS_DESCRIPTOR relationships, only. See the Supplementary Material for additional information.

3. Results

Table 1 contains the number of nodes and relationships in DisMaNET. In the case of relationships, in addition to the total, a distinction is made between those of hierarchical type (IS_A, HAS_DESCRIPTOR) and mapping type (MAP).

[12] https://github.com/monarch-initiative/mondo
[13] https://www.disgenet.org/downloads
[14] http://disnet.ctb.upm.es
[15] http://disnet.ctb.upm.es/dismanet-neo4j
[16] http://disnet.ctb.upm.es/dismanet-neo4j/db/data/transaction/commit
[17] http://disnet.ctb.upm.es/dismanet-api/mappings/
[18] https://neo4j.com/developer/java/

Vocabulary	*Label*	*Nodes*	*Relationships*	*Hierarchical*	*Mapping*
MeSH	MeSH	4,903	36,275	23,715	12,560
MeSH SCR	MeSHSCR	6,483	19,971	11,783	8,188
ICD-10-CM	ICD10CM	92,417	322,296	80,428	241,868
SNOMED CT	SNOMEDCT	150,352	607,679	368,746	238,933
Orphanet	ORPHANET	8,246	30,537	7,147	23,390
OMIM	OMIM	8,916	22,293	0	22,293
UMLS	UMLS	207,409	245,751	0	245,751
DO	DO	8,512	36,275	8,702	26,623
Total		487,238	898,541	488,738	409,803

Table 1. Nodes and relationships per vocabulary source in DisMaNET. In addition to the total count, the relationships are broken down by mapping and hierarchical type.

3.1. Finding missing mappings with DisMaNET

Thanks to the completeness of the database and the query capabilities described in Section 2, we could quantify the connections between disease vocabularies that are missing in the mapping sources. Table 2 contains the number of unique MeSH descriptors that are not related with terms of other vocabularies through the mapping sources. The number of MeSH descriptors not mapped with ICD-10-CM and SNOMED CT codes in UMLS coincide with those of the previous study [12].

Vocabulary	*UMLS*	*DO*	*Orphanet*
ICD-10-CM	2,458	3,395	4,228
SNOMED CT	702	2,606	4,903
OMIM	4,117	4,253	4,497
Orphanet	4,903	4,697	4,136

Table 2. Number of unique MeSH descriptors with missing mappings to the target vocabulary (first column) in each mapping source (header). As Orphanet is both a vocabulary and a mapping source, the Orphanet-Orphanet cell contains the number of missing MeSH descriptors in this source.

Table 3 contains the number of unique MeSH descriptors for which new mappings to other vocabularies were found through alternative sources, as described in Section 2. Again, these results are consistent with those obtained by the equivalent methods in the previous study. Additionally, they suggest that it is possible to complete mapping sources with the data extracted from others. However, it is noteworthy that the figures are considerably lower when we set a relevance threshold of 0.5. We analyze the significance of the score in Section 4.

	Total			*score >0.5*		
Vocabulary	*UMLS*	*DO*	*Orphanet*	*UMLS*	*DO*	*Orphanet*
ICD-10-CM	1,663	2,600	3,433	138	696	1,155
SNOMED CT	134	2,038	4,335	19	1,281	2,907
OMIM	1,219	1,355	1,599	227	274	400
Orphanet	1,318	2,378	2,584	126	821	965

Table 3. Number of unique MeSH descriptors for which new mappings to other vocabularies (first column) were found by querying DisMaNET, for a maximum distance of 4 MAP relationships between the MeSH node and the target vocabulary node. The vocabulary header refers to the mapping source in which the retrieved mappings are not available.

3.2. Interpretable cross-mappings of vocabularies

The previous results provide quantitative evidence of the power of DisMaNET to resolve missing mappings. However, in practice what researches need is to find the equivalent concept in a target vocabulary for a given concept in an origin vocabulary, with the closest relationship. Additionally, for indirect relationships it is essential to understand how they have been established and ensure that the result is valid for their investigation. To address this need, we used Neo4j's Shortest Path algorithm as described in Section 2. For example, the following query computes the shortest path between the MeSH descriptor D008577 ("Meningeal Neoplasms") and the ICD-10-CM vocabulary with a maximum of two relationships:

```
MATCH p=shortestPath((:MeSH{diseaseId:'D008577'})-[:MAP*1..2]-(:ICD10CM))
RETURN *
```

Figure 4 shows the result of the query in graph mode, when varying the maximum distance between 2 and 4.

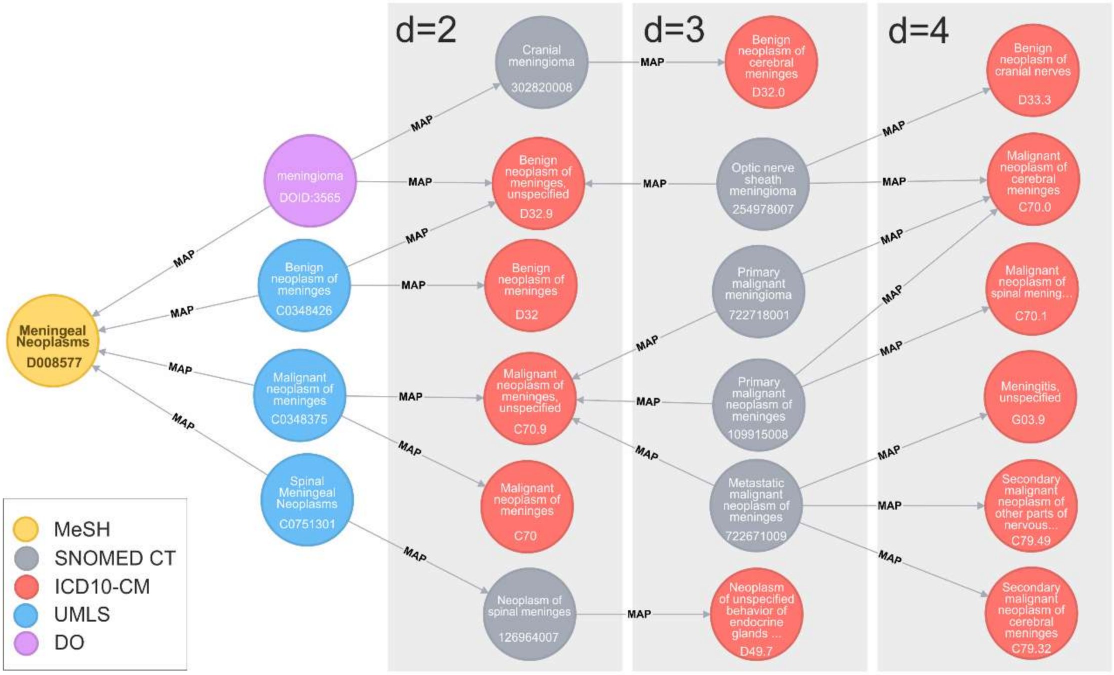
d=2
d=3
d=4
Meningeal Neoplasms
D008577
meningioma
DOID:3565
Benign neoplasm of meninges
C0348426
Malignant neoplasm of meninges
C0348375
Spinal Meningeal Neoplasms
C0751301
Cranial meningioma
302820008
Benign neoplasm of meninges, unspecified
D32.9
Benign neoplasm of meninges
D32
Malignant neoplasm of meninges, unspecified
C70.9
Malignant neoplasm of meninges
C70
Neoplasm of spinal meninges
126964007
Benign neoplasm of cerebral meninges
D32.0
Optic nerve sheath meningioma
254978007
Primary malignant meningioma
722718001
Primary malignant neoplasm of meninges
109915008
Metastatic malignant neoplasm of meninges
722671009
Neoplasm of unspecified behavior of endocrine glands ...
D49.7
Benign neoplasm of cranial nerves
D33.3
Malignant neoplasm of cerebral meninges
C70.0
Malignant neoplasm of spinal mening...
C70.1
Meningitis, unspecified
G03.9
Secondary malignant neoplasm of other parts of nervous...
C79.49
Secondary malignant neoplasm of cerebral meninges
C79.32
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
MAP
● MeSH
● SNOMED CT
● ICD10-CM
● UMLS
● DO

Figure 4. Connections between the Meningeal Neoplasms MeSH descriptor and ICD-10-CM concepts, obtained by using the Shortest Path algorithm with a maximum distance (d) of 2, 3 and 4 MAP relationships. Only MAP relationships are represented.

Thanks to the integration of OMIM, Orphanet and MeSH SCR in DisMaNET, we can use the same method to map rare diseases across different vocabularies. For instance, the following query retrieves the connections between "Osteopetrosis-hypogammaglobulinemia syndrome", an extremely rare primary bone dysplasia encoded in Orphanet as "178389", and MeSH descriptors by using not only MAP, but also HAS_DESCRIPTOR relationships [25]:

```
MATCH p=shortestPath((:ORPHANET{diseaseId:'178389'})-
[:MAP|:HAS_DESCRIPTOR*1..3]-(:MeSH))
RETURN *
```

Figure 5 shows the result of the query in graphical mode. Hierarchical IS_A relationships are also included for better interpretability. Section 4 describes how these graphs help to interpret the results.

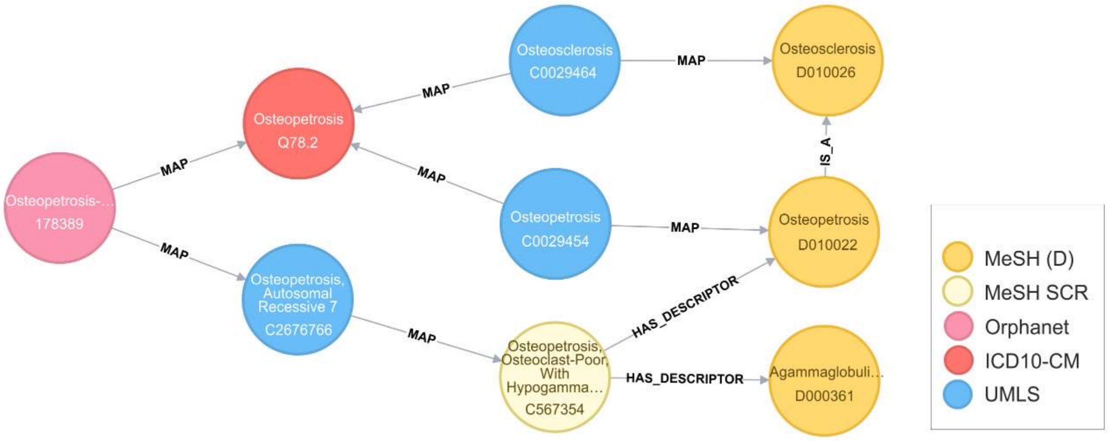

Figure 5. Connections between the rare disease Osteopetrosis-hypogammaglobulinemia in the Orphanet vocabulary and MeSH descriptors, obtained by using the Shortest Path algorithm with a maximum distance of 3 MAP relationships.

3.3. Evaluation

Table 4 contains the results of the evaluation of DisMaNET with MonDO. Of the 31,308 comparable mappings in this source, 27,349 (87.36%) were found by querying DisMaNET using the Shortest Path algorithm, as described previously. Filtering the results by their relevance, 85.17% and 53.88% of the comparable mappings had a score greater than 0.5 and 0.9, respectively. These results demonstrate that DisMaNET is capable of solving these mappings with a performance comparable to that of MonDO, where a semi-automatic approach is used. In addition to the matching mappings, 192,106 connections between UMLS concepts and other vocabularies that are not available in MonDO were obtained through DisMaNET. Of these, 31.58% had a high relevance (score > 0.5) and 0.80% a very high relevance (score > 0.9). For instance, MONDO:0005253 ("High output heart failure") was mapped to the homonymous ICD-10-CM code I50.83 with score 1.

	Existing in MonDo	*Missing in DisMaNET*	*MonDO and DisMaNET*			*New in DisMaNET*		
Vocabulary			*Total*	*s>0.5*	*s>0.9*	*Total*	*s>0.5*	*s>0.9*
DO	7,083	545	6,538	6,498	5,151	12,487	6,264	28
ICD-10-CM	1,121	102	1,019	808	629	10,229	4,794	584
MeSH	6,107	92	6,015	5,967	3,229	47,576	9,619	249
OMIM	6,062	2,339	3,723	3,702	2,256	25,478	18,221	83
Orphanet	4,519	349	4,170	4,049	3,137	25,795	6,710	62
SNOMED CT	6,416	532	5,884	5,643	2,468	70,541	15,064	532
Total	31,308	3,959	27,349	26,667	16,870	192,106	60,672	1,538

Table 4. Number mappings between UMLS concepts in MonDO and other vocabularies (first column) found by querying DisMaNET, for different relevance scores (s)

On the other hand, Table 5 contains the results of the evaluation with DisGeNET. In this case, 31,036 of the 37,916 comparable mappings in this source (81.86%) were found in DisMaNET. Connections with a score higher than 0.5 and 0.9 were found for 65.92% and 31.34% of the mappings, respectively. These results are comparable to those obtained by MalaCards [26]. Of the 279,932 new mappings between UMLS concepts in DisGeNET and other vocabularies in DisMaNET, 34.31% and 2.67% had high or very high relevance, respectively. It is worth mentioning that this large result set is due to the fact that mappings for vocabularies not available in DisGeNET, such as ICD-10-CM, Orphanet and SNOMED CT, are included. For instance, UMLS CUI C0409495 (“Protrusio acetabuli”) was mapped to the homonymous terms in ICD-10-CM (M24.7) and SNOMED CT (59606006). These results have been shared in the project repository[19].

[19] https://github.com/dismanet/paper/tree/master/results

	Existing in DisGeNET	*Missing in DisMaNET*	*DisGeNET and DisMaNET*			*New in DisMaNET*		
Vocabulary			*Total*	*s>0.5*	*s>0.9*	*Total*	*s>0.5*	*s>0.9*
DO	15,675	1,465	14,210	8,227	2,456	13,079	3,313	160
ICD-10-CM	N/A	N/A	N/A	N/A	N/A	17,369	7,934	1,511
MeSH	10,064	142	9,922	9,902	5,183	62,980	14,931	8
OMIM	12,177	5,273	6,904	6,867	4,245	40,689	28,068	36
Orphanet	N/A	N/A	N/A	N/A	N/A	34,608	12,299	1,922
SNOMED CT	N/A	N/A	N/A	N/A	N/A	111,207	29,494	3,817
Total	37,916	6,880	31,036	24,996	11,884	279,932	96,039	7,454

Table 5. Number mappings between UMLS concepts in DisGeNET and other vocabularies (first column) found by querying DisMaNET, for different relevance scores (s).

3.4. Contribution to DISNET

The DISNET database integrates phenotypic and genetic-biological characteristics of diseases and information on drugs from several sources [27]. In the case of Wikipedia, one of its textual sources, disease articles usually include a list of "Medical Resources" with the codes of disease vocabularies associated to the term. However, these resources do not include UMLS codes[20], making it difficult to integrate this source of phenotypic data with the biological layer in DISNET. To address this problem, we used DisMaNET to resolve the mappings missing in DISNET (as of January 2020) with UMLS and other vocabularies. The results are shown in Table 6. Mappings with the highest scores were contributed to Wikipedia, resulting in 950 disease articles extended with new medical resources[21].

[20] https://en.wikipedia.org/wiki/Template:Medical_resources

[21] https://en.wikipedia.org/wiki/Special:Contributions/Eduardo_P._Garc%C3%ADa_del_Valle

	Existing in DISNET	*New in DisMaNET*		
Vocabulary		*Total*	*s>0.5*	*s>0.9*
DO	N/A	5,668	4,163	1,778
ICD-10-CM	4,949	6,055	1,864	95
MeSH	4,179	1,849	794	52
OMIM	2,238	1,503	771	11
Orphanet	977	9,058	3,147	606
SNOMED CT	20	32,377	10,091	742
UMLS	N/A	16,035	11,658	3,051
Total	12,363	72,545	32,488	6,335

Table 6. Number mappings between DISNET disease codes and other vocabularies (first column) found by querying DisMaNET, for different relevance scores (s).

4. Discussion

Our disease vocabulary cross-mapping system consists of a complex network with around five hundred thousand nodes and approximately nine hundred thousand relationships. However, numbers vary significantly between vocabularies, as depicted in Table 1. In the case of MeSH descriptors, for example, the number of relationships, especially of hierarchical type, is relatively large with respect to the node count. This is due to the fact that the same descriptor is usually related to multiple parent descriptors, sometimes within different categories. In contrast, ICD-10-CM, whose elements have only one ancestor, presents a number of hierarchical relationships very close to that of nodes. Further queries to DisMaNET reveal that only 185 disease terms are connected through direct MAP type relationships in all 8 vocabularies. If we only consider UMLS, SNOMED CT, ICD-10-CM and MeSH, vocabularies, the number of connected concepts totals 2,728, which is slightly higher than in MalaCards [26].

The completeness of the database and the power of network analysis enabled us not only to quantify the number of MeSH descriptors which are not mapped with terms of other vocabularies in the mapping sources (Table 2), but also to solve a significant number of these missing mappings (Table 3). Orphanet has the largest contribution of new connections between MeSH descriptors and other vocabularies, with 81.20% and 88.41% of the missing mappings with ICD-10-CM and

SNOMED CT resolved, respectively. The reason is that Orphanet does not contain direct relationships with SNOMED CT, and relatively few with ICD-10-CM. If we look at the missing mappings between DO and SNOMED CT, 78.20% of them were resolved by applying the same techniques. As a result, 75.96% of all the DO terms in DisMaNET are connected to SNOMED CT concepts, which is comparable in relative terms to previous studies based on mapping through semantic and hierarchical characterization [7]. The comparison with MonDO and DisGeNET, with over 80% of coincidences in both cases, confirms the ability of DisMaNET to cross-map codes from different vocabularies with a performance comparable to that of the state of the art.

Despite these benefits, the extent of the network and the use of algorithms such as Shortest Path result in an overwhelming number of cross-mappings. To alleviate this effect and provide a quantitative reference of the mapping relevance, we introduced a score metric in Section 2. The results in Section 3 show that the ratio of relevant mappings (score > 0.5) ranges from 10% to 70%. The fact that most of the mappings in DisMaNET validated with MonDO and DisGeNET have a high score, demonstrates the validity of this metric to measure the relevance of the results and to prevent eventual false positives. Still, this metric is only a first approximation, and it is necessary to study more advanced alternatives to improve the reliability of our solution. To complement the quantitative assessment provided by the score, result visualization allows researchers to understand and validate mappings qualitatively. For example, Figure 4 shows the ICD-10-CM concepts associated with the "Meningeal Neoplasms" MeSH descriptor. In particular, the mapping with ICD-10-CM code "G03.9" ("Meningitis, unspecified") has a score of 0.125. Although not highly relevant, this relationship is meaningful when we look at the graph. Meningitis is typically caused by an infection with microorganisms, but in some cases it may occur as the result of several non-infectious causes, such as the spread of cancer to the meninges (malignant or neoplastic meningitis) [28].

DisMaNET has proven to be an effective cross-mapping tool even in the case of rare diseases, which often present additional difficulties. On the one hand, rare disease databases such as OMIM and Orphanet represent different perspectives of diseases, and as a result they are inconsistently cross-referenced [29, 30]. On the other hand, generic vocabularies have very limited coverage of this type of diseases, posing a major problem for health insurance reimbursement and research. Figure 5 shows the connections between the "Osteopetrosis-hypogammaglobulinemia" term in Orphanet with MeSH, obtained by querying DisMaNET. Interestingly, there is an indirect relationship with the MeSH descriptor "Agammaglobulinemia" through the Supplementary Concept Record "Osteopetrosis, Osteoclast-Poor, With Hypogammaglobulinemia", via HAS_DESCRIPTOR. As explained, MeSH SCR are often used to label rare diseases. This example demonstrates the value of introducing more sources in DisMaNET to solve non-direct mappings, especially for rare diseases.

5. Conclusion

The integration of data from biological, clinical and literary sources enables the study of diseases from a more comprehensive and holistic approach. However, the interoperability of these sources, particularly of the codes used to identify diseases, poses a major challenge. Despite the availability of cross mapping resources such as the Unified Medical Language System or the Disease Ontology, their coverage, especially in the case of rare diseases, is still insufficient. To address this problem, we built DisMaNET, a network-based system to cross-map terms from disease vocabularies in an intuitive and efficient way. Thanks to its completeness and the power of network analysis, our system is able to detect and solve a significant number of the mappings missing in the most commonly used resources, even for rare diseases. The study compares DisMaNET with other resources such as MonDO and DisGeNET, demonstrating that our tool is capable of obtaining more than 80% of their mappings. A relevance score and the possibility of visually analyzing the obtained mappings facilitate the understanding and validation of the results. DisMaNET and the datasets used to build it are publicly available for their exploitation and evaluation.

Given the continuous publication of new versions of the disease vocabularies, the update and versioning of DisMaNET is one of the next challenges to face. On the other hand, the exploitation of alternative network analysis algorithms to find disease mappings, the design of a more advanced relevance score and the integration of new data sources, such as NCI or GRAD, are envisioned as future lines of work. Finally, it is necessary to investigate and evaluate alternatives for the network algorithms and the relevance score, in order to improve the precision of the system.

Funding

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). This paper has been supported by Fundación AECC and Instituto de Salud Carlos III (grant AC19/00034), under the frame of ERA-NET PerMed.

Competing Interests

The authors have no competing interests to declare.

References

1. García del Valle EP, Lagunes García G, Prieto Santamaría L, Zanin M, Menasalvas Ruiz E, Rodríguez-González A. Disease networks and their contribution to disease understanding: A review of their evolution, techniques and data sources. Journal of Biomedical Informatics. 2019;94:103206.

2. McKusick VA. Mendelian Inheritance in Man and Its Online Version, OMIM. Am J Hum Genet. 2007;80:588–604.

3. Aymé S, Schmidtke J. Networking for rare diseases: a necessity for Europe. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2007;50:1477–83.

4. National Cancer Institute ActText of the Act of August 5, 1937, creating the National Cancer Institute and authorizing an appropriation therefor. J Natl Cancer Inst. 1957;19:133–7.

5. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp. 1998;:815–9.

6. Pereira S, Névéol A, Massari P, Joubert M, Darmoni S. Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. Stud Health Technol Inform. 2006;124:845–50.

7. Raje S, Bodenreider O. Interoperability of Disease Concepts in Clinical and Research Ontologies: Contrasting Coverage and Structure in the Disease Ontology and SNOMED CT. Stud Health Technol Inform. 2017;245:925–9.

8. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2017;45:D712–22.

9. Rappaport N, Twik M, Nativ N, Stelzer G, Bahir I, Stein TI, et al. MalaCards: A Comprehensive Automatically-Mined Database of Human Diseases. Current Protocols in Bioinformatics. 2014;47:1.24.1-1.24.19.

10. Lou Y, Zhang Y, Qian T, Li F, Xiong S, Ji D. A transition-based joint model for disease named entity recognition and normalization. Bioinformatics. 2017;33:2363–71.

11. Searls DB. Data integration: challenges for drug discovery. Nat Rev Drug Discov. 2005;4:45–58.

12. García EP, García GL, Ruiz EM, Santamaría LP, Zanin M, Rodríguez-González A. Completing Missing MeSH Code Mappings in UMLS Through Alternative Expert-Curated Sources. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). 2019. p. 174–9.

13. Névéol A, Shooshan SE, Mork JG, Aronson AR. Fine-Grained Indexing of the Biomedical Literature: MeSH Subheading Attachment for a MEDLINE Indexing Tool. AMIA Annu Symp Proc. 2007;2007:553–7.

14. Li Y, Agarwal P. A Pathway-Based View of Human Diseases and Disease Relationships. PLoS One. 2009;4. doi:10.1371/journal.pone.0004346.

15. Hu Y, Zhao L, Liu Z, Ju H, Shi H, Xu P, et al. DisSetSim: an online system for calculating similarity between disease sets. J Biomed Semantics. 2017;8 Suppl 1. doi:10.1186/s13326-017-0140-2.

16. Yoon B-H, Kim S-K, Kim S-Y. Use of Graph Database for the Integration of Heterogeneous Biological Data. Genomics Inform. 2017;15:19–27.

17. Lotfy A, Saleh AI, El-Ghareeb HA, Ali HA. A middle layer solution to support ACID properties for NoSQL databases. 2016.

18. Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. Representing and querying disease networks using graph databases. BioData Min. 2016;9. doi:10.1186/s13040-016-0102-8.

19. Himmelstein DS, Baranzini SE. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. PLOS Computational Biology. 2015;11:e1004259.

20. Mullen J, Cockell SJ, Woollard P, Wipat A. An Integrated Data Driven Approach to Drug Repositioning Using Gene-Disease Associations. PLoS One. 2016;11. doi:10.1371/journal.pone.0155811.

21. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. Journal of The Royal Society Interface. 2015;12:20150571.

22. Zuo X, Blackburn J, Kourtellis N, Skvoretz J, Iamnitchi A. The power of indirect ties. Computer Communications. 2016;73:188–99.

23. Friedkin NE. Horizons of Observability and Limits of Informal Control in Organizations. Social Forces. 1983;62:54–77.

24. Gunawan D, Sembiring CA, Budiman MA. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. J Phys: Conf Ser. 2018;978:012120.

25. Guerrini MM, Sobacchi C, Cassani B, Abinun M, Kilic SS, Pangrazio A, et al. Human osteoclast-poor osteopetrosis with hypogammaglobulinemia due to TNFRSF11A (RANK) mutations. Am J Hum Genet. 2008;83:64–76.

26. Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. 2017;45:D877–87.

27. Lagunes-García G, Rodríguez-González A, Prieto-Santamaría L, Valle EPG del, Zanin M, Menasalvas-Ruiz E. DISNET: a framework for extracting phenotypic disease information from public sources. PeerJ. 2020;8:e8580.

28. Gleissner B, Chamberlain MC. Neoplastic meningitis. Lancet Neurol. 2006;5:443–52.

29. Sarntivijai S, Vasant D, Jupp S, Saunders G, Bento AP, Gonzalez D, et al. Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. J Biomed Semantics. 2016;7. doi:10.1186/s13326-016-0051-7.

30. Rance B, Snyder M, Lewis J, Bodenreider O. Leveraging Terminological Resources for Mapping between Rare Disease Information Sources. Stud Health Technol Inform. 2013;192:529–33.