DISNET: Disease understanding through complex networks creation and analysis

Gerardo Lagunes García[1]
gerardo.lagunes@ctb.upm.es

Alejandro Rodríguez González[1,2]
alejandro.rg@upm.es

Lucía Prieto Santamaría[1]
lucia.prieto.santamaria@alumnos.upm.es

Eduardo P. García del Valle[2]
ep.garcia@alumnos.upm.es

Massimiliano Zanin[1]
massimiliano.zanin@ctb.upm.es

Ernestina Menasalvas Ruiz[1,2]
ernestina.menasalvas@upm.es

[1]Universidad Politécnica de Madrid
Centro de Tecnología Biomédica
Campus de Montegancedo
Pozuelo de Alarcón, 28223, Madrid, Spain

[2]Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingenieros Informáticos.
Departamento de Lenguajes y Sistemas Informáticos e Ingeniería del Software
Campus de Montegancedo
Boadilla del Monte, 28660, Madrid, Spain

ABSTRACT

Within the global endeavour of improving population health, one major challenge is the increasingly high cost associated with drug development. Drug repositioning, i.e. finding new uses for existing drugs, is a promising alternative; yet, its effectiveness has hitherto been hindered by our limited knowledge about diseases and their relationships. In this paper we present DISNET (Drug repositioning and disease understanding through complex networks creation and analysis), a web-based system designed to extract knowledge from signs and symptoms retrieved from medical data bases, and to enable the creation of customisable disease networks. We here present the main functionalities of the DISNET system. We describe how information on diseases and their phenotypic manifestations is extracted from Wikipedia, PubMed and MayoClinic; specifically, texts from these sources are processed through a combination of text mining and natural language processing techniques. We further present a validation of the processing performed by the system; and describe, with some simple use cases, how a user can interact with it and extract information that could be used for subsequent analyses.

Database URL: http://disnet.ctb.upm.es

1. Introduction

In 1796, Edward Jenner found an important link between the variola virus, which affected only humans and was highly lethal, and the bovine smallpox virus, which attacked cows and was transmitted to humans by physical contact with infected animals, and which, despite its severity, rarely resulted in death. He found that people who became infected with the latter (also called cowpox) did not subsequently catch the former; and thus, that something in the bovine smallpox virus made humans immune to variola virus. This led him to thoroughly investigate the relationship between these diseases and understand their behaviour for more than twenty years; to be finally able to find a cure for the variola virus, saving thousands of humans lives worldwide.

This discovery illustrates the importance of the knowledge that we can get from diseases and, more specifically, from how they are related. Despite the fact that in the last 200 years our understanding of diseases has greatly increased, and valuable advances have been made in this area (1), the number of those without treatment or cure is still extremely high (e.g. Alzheimer's disease, small cell lung cancer, HIV, etc.). It is thus imperative to explore new approaches and tools to tackle them and, therefore, improve the health of the world's population.

It is almost a truism that the search for new drugs requires a better understanding about diseases. This includes finding new insights on the relationship between diseases (which diseases are related and how), as well as the creation of public and easy-to-access large databases of diseases knowledge. In this context, several works have attempted to understand these relationships by creating and analysing disease networks. The complexity of such endeavour was soon clear, as diseases may share not only symptoms and signs, but also genes, proteins, causes and, in many cases, cures (2–5).

One of the most important works on the subject was published in 2007 by K.-I. Goh et al. (2), in which the HDN (Human Disease Network) was developed, a network of human diseases and disorders that links diseases based on their genetic origins and biological interactions. Different diseases were then associated according to shared genes, proteins or protein interactions. The hypothesis that different diseases, with potentially different causes, may share characteristics allows the design of common strategies regarding how to deal with the diagnosis, treatment and prognosis of a disease.

Within this line of research it is worth mentioning the Human Symptoms-Disease Network (HSDN), published in the journal Nature Communications in 2014 (3): an HDN network in which similarities between diseases were estimated through common symptoms. This is an important change in perspective with respect to previous works, in which the focus was centred on the genetic and biological origin of the diseases. In (3), diseases are defined by their clinical phenotypic manifestations, i.e. signs and symptoms; this is not surprising, as these manifestations are basic medical elements, and crucial characteristics in the diagnosis, categorization and clinical treatment of the diseases. It was then proposed to use these as a starting point to understand the existing relationships between different diseases.

Building on top of these previous works, and stemming from the necessity of having exhaustive and accurate sources of disease-based information, in this paper we present the DISNET (Diseases Networks) system. DISNET aims at going one step further in improving human knowledge about diseases, not only by seeking and analysing the relations between them, but most importantly, by finding real connections between diseases and drugs, thus potentially enabling novel drug repositioning strategies.

The DISNET system allows to capture information about diseases from heterogeneous textual sources, and extract the relevant information from them as done in the HSDN work

[10] but DISNET, nevertheless goes one step further, since among other features it is not limited to a single source of information, provides an API-based access to the data and integrates more powerful extraction for the extraction of phenotypical manifestations from the textual sources, among other characteristics. The captured knowledge will allow to analyse the diseases and their relationships, being current version of the system focused on phenotypical information. Future content to be introduced includes genetic and drug information to create a complex multilayer network, where each layer represents the different type of information (phenotypical, biological, drugs).

Beyond this introduction, this paper is organised as follows: section 2 analyses the related works in the context of human disease networks, their results and the main technological drawbacks and limitations; section 3 explains the technologies used in the creation of DISNET; section 4 presents the main results obtained in the validation of the system, and describes several simple use cases; finally, section 5 draws some conclusions and discusses future work.

2. Related Works

In this section we present some essential works on the construction and analysis of disease networks, which is the concept underlying the DISNET system. While we here aim at providing a synthetic overview of this vast field, the interested reader may refer to (6) for a complete survey on the topic.

Although the idea of exploring the association between phenotypes and genotypes is an old one, its formalisation into the HDN, and thus the use of a graph formalism, was firstly introduced in 2007 by Goh et al. (2). As the authors pointed out, "*those genes associated with similar disorders show higher likelihood of physical interactions between their products and higher expression profiling similarity for their transcripts, supporting the existence of distinct disease-specific functional modules*". The proposed HDN is an extension of previous and successful studies that used the network-based approach to create genome-wide associations (7), list of disorder-gene association pairs (4, 8) or mappings between protein-protein interactions (9, 10); but rather than focusing on a single condition, (2) tried to achieve these goals with multiple diseases. The concept proposed by the authors is to explore human genetic disorders and the corresponding disease genes to see "*if they are related to each other at a higher level of cellular and organismal organization*".

In 2010 a paper with the first quantitative framework to compare diseases through an integrated analysis of disease-related mRNA expression data and the human protein interaction network was published (8). The authors claimed that "*14 of the significant disease correlations also shared common drugs*". They suggested that the knowledge resulting from a disease-similarity network based on molecular data can be used in the discovery of new uses for existing drugs, as similar diseases share common molecular phenotypes and can thus be treated by similar drugs. To test this hypothesis, the authors collected a list of drugs, along with their target genes and the diseases by them treated; this was done through different databases, such as RxNorm, DrugBank, National Drug File Reference Terminology (ND-FRT) and MicroMedex. The final result was that "*at least 17 of the 138 significant disease correlations shared at least one drug in common and 14 of them had a significant hypergeometric p-value less than 0.01*". In 2011 Barabási et al. published a review (4) paper where they proposed a key hypothesis: "*a disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex intracellular and intracellular networks that link tissue and organ systems*". Three years after the Barabási's paper, a paper about the visualization of genetic disease-phenotype similarities to assist in the disambiguation of symptoms was published (9). In this paper, the authors claimed that from a phenotypic point of view there are several diseases that are difficult to diagnose due to the combination of confounding symptoms. The overlapping of these symptoms suggests that there should be shared mechanisms between different diseases.

The main idea of DISNET rests upon the hypothesis previously introduced (3): "*the overlapping of symptoms in different diseases suggests that there should be some kind of shared mechanisms*". The authors claimed that there is a necessity for constructing and investigating the connection between clinical manifestations of diseases and their underlying molecular interactions; and that "*the elucidation of the connection between shared symptoms and shared genes or protein-protein interactions of two diseases could therefore help bridge the gap between bench-based biological discovery and bedside clinical solutions*". The result of the creation of such network shows that "*symptom-based similarity correlates strongly with the number of shared genetic associations and the extent to which their associated proteins interact*". Authors linked disease pairs based on the similarity of their respective symptoms. The main network of disease-similarity is created through a large-scale medical text-mining process over bibliographical records from PubMed and the use of Medical Subject Headings (MeSH) metadata. The approach is completed through the integration of disease-gene associations and protein-protein interactions (PPI), and is then used to investigate the correlations between the symptom similarity of diseases and the degree of shared genes or PPIs. The analysis of the network revealed that diseases with more similar symptoms are more likely to have common gene associations. The authors further proposed that high similarity scores could suggest yet unknown common genetic associations, something which was confirmed by other similar works (11). Results "*demonstrates that individual-level disease phenotypes (for example, symptoms) and molecular-level disease components (for example, genes and PPIs) show robust correlations, even though their direct associations are influenced by complicated intermediate factors*". Another consideration is that symptoms play a crucial role in drug-related research. Authors claimed that most of the drugs approved by the US Food and Drug Administration are merely palliative (12). The relationship encoded in the HSDN could allow creating hypothesis regarding the use of drugs that are specific for a disease in a different one. Such novel approach opens a very interesting line of research with an enormous power, but it has some deficiencies and drawbacks: as the authors recognized, the use of MeSH limits the number of medical concepts available and the semantic relationships between those concepts. Overcoming this limitation may entail, on one hand, the use of more powerful terminologies, such as SNOMED-CT (13), for the analysis and text- mining processes (14). On the other hand, the solution may include the use of specialized open access medical sources for the extraction of disease symptoms, such as PubMed (https://www.ncbi.nlm.nih.gov/pubmed), MedLine Plus (https://medlineplus.gov) (15), MayoClinic (https://www.mayoclinic.org), CDC (https://www.cdc.gov), or unspecialized source such as Wikipedia or Freebase among others.

A paper (11) based on the aforementioned assumption, i.e. that symptoms and signs are the essential clinical manifestation for diagnosis and treatment, was published in 2014 with an application to Traditional Chinese Medicine (TCM). The authors developed a computational approach to identify the candidate genes of symptoms. Some results show that, for example, genes like CALCA, ESR1 and MTHFR were predicted to be associated with headache symptoms, somethings that has been in recent literature. In the same context of TCM, other authors published a paper (16) in which they construct a clinical phenotype network (CPN), with phenotype entities such as symptoms and diagnosis being represented by nodes, and the correlation between these entities by links. The authors based their research assumption on the idea that the interconnections between genotypes and phenotypes to the relevant diseases (17–19) form a complex network (20). As part of the results, the same conclusions are drawn: clinical phenotypes, such as diseases and symptoms, are complex and usually co-occurring, suggesting that they have common underlying molecular mechanisms (25, 26). Another result that came up from this paper is the relationship between the diseases and the treatments in TCM, which are mostly based in the use of herbs. There is evidence of "*correlations between symptoms and herbs because herb prescription consists of herb ingredients and syndromes are differentiated based on the manifestation of symptoms*".

3. Materials and Methods

This Section discusses the technical aspects of the DISNET system, focusing on two aspects: the sources of information hitherto considered, and the DISNET workflow. More specifically, the last point describes how the system retrieves phenotypic information, in the form of raw texts, from the discussed sources; how these texts are processed to obtain diagnostic terms; and how these terms are validated to compile a final list of valid symptom-type terms.

A. Information Source

As it has previously been shown, it is customary for works aimed at unveiling relationships between diseases to focus on single source of information, in most cases just abstracts of Medline articles. On the other hand, the proposed system aims at obtaining inputs from as many sources as possible, to guarantee the recovery of as much knowledge as possible. By bringing together information from different sources, we expect them to complement each other, creating a network with a higher capacity of relating diseases. The rationale for this is that the different sources of textual knowledge, such as MayoClinic, Wikipedia, or PubMed, are written in different styles and by people with different backgrounds; the information they contain may therefore be complementary. In order to take advantage of such richness, the DISNET system allows the user to query the symptoms according to different rules: for instance, from one or multiple sources, by applying filters based on prevalence information, or on percentages of similarity among others. This clearly comes at a cost: the system should be flexible enough to be able to process sources with different structures. In the remainder of this Section we discuss the patterns used to select data sources, how they have been mined, and finally the challenges involved in such tasks.

B. Source Selection

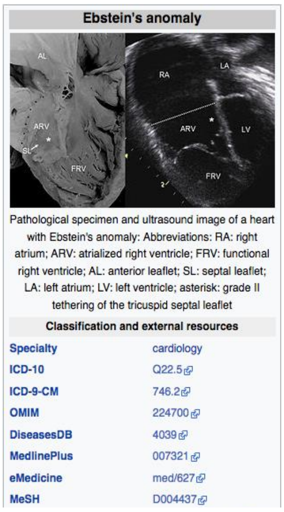
Traditionally, in order to obtain the whole body of knowledge that mankind has accumulated about a given disease, one would refer to medical books. Although books usually contain much of the information available, they also present some important limitations: they are not constantly updated; the automatic access to their content is difficult, especially when digital versions are not available; and they are usually written for study, thus the information they contain is not structured for data mining tasks. On the other hand, one has the World Wide Web, whose main characteristic is to be (mostly) free accessible to anyone with an internet connection. It mainly offers three sources of information. Firstly, the abstract, and in some cases, the full text, of medical papers, which can be accessed through platforms like PubMed. Secondly, specialized sources of information, such as MedlinePlus, MayoClinic, or CDC. Finally, good medical data can be obtained in sources of knowledge that are not specialized, such as

Wikipedia or Freebase. Note that all of them have different characteristics, in terms of comprehensiveness, degree of structure of the information, and up-to-datedness.

The criteria used for the selection of the sources of information in DISNET are: i) open access, ii) recognised quality and reliability, and iii) availability of substantial quantities of data (structured or not). This suggested to include the following three web sites in the system, which are described below: i) Wikipedia, ii) PubMed, and iii) MayoClinic. It is important to note that the system is not closed; on the contrary, thanks to its flexibility, new sources could (and will) be incorporated in the future.

C. Wikipedia

Wikipedia is an online, open and collaborative source of information. It was created by the Wikimedia Foundation and its English edition is the largest and most active one. The monumental and primary task of editing, revising and improving the quality of all articles is not performed by a core of administrators: it is instead the collaborative result of thousands of users. Consequently, this encyclopaedia is considered the greatest collective project in the history of humanity (23).



Ebstein's anomaly

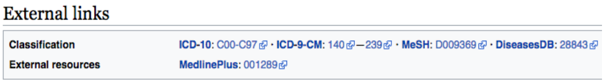
Pathological specimen and ultrasound image of a heart with Ebstein's anomaly: Abbreviations: RA: right atrium; ARV: atrialized right ventricle; FRV: functional right ventricle; AL: anterior leaflet; SL: septal leaflet; LA: left atrium; LV: left ventricle; asterisk: grade II tethering of the tricuspid septal leaflet

Classification and external resources	
Specialty	cardiology
ICD-10	Q22.5
ICD-9-CM	746.2
OMIM	224700
DiseasesDB	4039
MedlinePlus	007321
eMedicine	med/627
MeSH	D004437

Fig. 1. External vocabularies in a vertical *infobox* in Wikipedia article on Ebstein's anomaly

Wikipedia contains more than 155,000 articles in the field of medicine (24) and is one of the most widely used medical sources (25) by the general community (23) and also by medical specialists (26), the latter ones having deeply been involved in its enrichment (25, 28). One of the initiatives is the Cochrane/Wikipedia, which aims at increasing reliability in articles with medical content (28). In 2014 Wikipedia was referred to as "*the single leading source of medical information for patients and health care professionals*" by the Institute of Medical Science (IMS) (29). This stems from the fact that an increasing number of people in the medical field are becoming aware of the importance of collaborating and generating quality content in the world's largest online encyclopedia.

We have focused on Wikipedia in its English edition, and specifically on those articles categorized as diseases. In order to obtain a list of such articles we resort to DBpedia (30), an open and free Web repository that stores structured information from Wikipedia and other Wikimedia projects (http://wiki.dbpedia.org). By containing structured information, this source allows complex questions to be asked through SPARQL queries (https://www.w3.org/TR/rdfsparql-query). We developed a query (http://bit.ly/get_diseases_query_sparql) that is able to get all the articles of Wikipedia in English referring to human diseases and run it in the **Virtuous environment SPARQL Query Editor of DBpedia** (https://dbpedia.org/sparql). This first approach to detecting and extracting Wikipedia's web links can be addressed in different ways and in the **Discussions** section we will talk about them.



External links

Classification	ICD-10: C00-C97 · ICD-9-CM: 140—239 · MeSH: D009369 · DiseasesDB: 28843
External resources	MedlinePlus: 001289

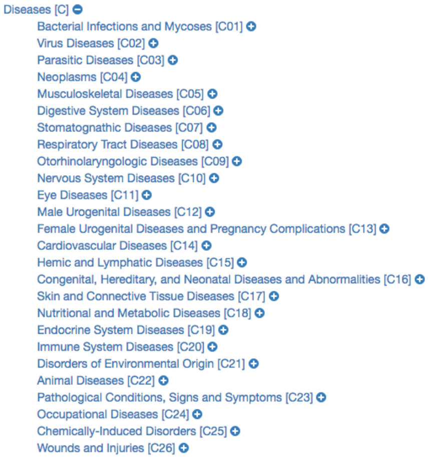
Fig. 2. External vocabularies in a horizontal *infobox* in Wikipedia article on Cancer

Even though disease articles have a standard structure, due to the very nature of Wikipedia, articles can be edited by anyone; consequently, it is possible to find articles that do not comply with the standard form that the creators of the encyclopedia propose (https://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style/Medicine-related_articles&oldid=826413087). The structure is organized in sections, of which we have selected those whose content is related to the phenotypic manifestations of the disease. The essential sections mined by DISNET are: "*Signs and symptoms*", "*Causes*", "*Diagnosis*", "*Presentation*" and "*infobox*".

The data retrieved from these sections are: i) the texts (paragraphs, lists and tables) contained in the previously described sections; ii) the links contained in these texts; and iii) the disease codes of vocabularies external to Wikipedia, which can be found in the *infoboxes* of the article. Note there are two types of *infobox*. **Fig. 1** shows an example of the external vocabulary codes retrieved in a vertical *infobox*, usually located at the beginning of the document; Fig. 2 shows an example of a horizontal *infobox*, generally located at the foot of the document. These disease codes in different vocabulary are relevant elements when searching for diseases in the system's database. The list of external vocabularies to DISNET can be found online at (http://bit.ly/wikipedia_medical_vocabularies_txt).

D. PubMed

PubMed comprises more than 28 million biomedical literature citations from MEDLINE, life science journals and online books. Quotations may include links to full text content from PubMed Central (https://www.ncbi.nlm.nih.gov/pmc) and editorial websites (31). As in other studies, we here only considered the abstracts of the articles, as, firstly, it is not always possible to access the full text, and secondly, the full text of articles does not follow a standard format. However, we are aware of the limitations of the extraction of information only for abstracts (32), and future versions of DISNET platform will focus in extracting the content from the full paper when possible. Note that in PubMed the information about one single disease is spread among multiple documents – as opposed to Wikipedia, in which there is a bijective relationship between articles and diseases.



Diseases [C]
- Bacterial Infections and Mycoses [C01]
- Virus Diseases [C02]
- Parasitic Diseases [C03]
- Neoplasms [C04]
- Musculoskeletal Diseases [C05]
- Digestive System Diseases [C06]
- Stomatognathic Diseases [C07]
- Respiratory Tract Diseases [C08]
- Otorhinolaryngologic Diseases [C09]
- Nervous System Diseases [C10]
- Eye Diseases [C11]
- Male Urogenital Diseases [C12]
- Female Urogenital Diseases and Pregnancy Complications [C13]
- Cardiovascular Diseases [C14]
- Hemic and Lymphatic Diseases [C15]
- Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16]
- Skin and Connective Tissue Diseases [C17]
- Nutritional and Metabolic Diseases [C18]
- Endocrine System Diseases [C19]
- Immune System Diseases [C20]
- Disorders of Environmental Origin [C21]
- Animal Diseases [C22]
- Pathological Conditions, Signs and Symptoms [C23]
- Occupational Diseases [C24]
- Chemically-Induced Disorders [C25]
- Wounds and Injuries [C26]

Fig. 4. Disease MeSH Term tree clasification

Obtaining the list of diseases in PubMed involves two main steps. Firstly, one should extract the list of MeSH terms (DMTL) relating to human diseases *C*, which are categorized from *C01* to *C26* (excluding those categories such as "Animal Diseases" or "Wounds and Injuries") as shown in the classification tree in Fig. 3 (https://b.nlm.nih.gov/treeView); and map each disease with Human Disease Ontology (http://www.obofoundry.org/ontology/doid.html) to obtain disease codes of the vocabulary ICD-10, OMIM, MeSH, SNOMED_CT and UMLS. Note that the use of multiple vocabularies aims at obtaining the greatest amount of means (identified codes) to identify diseases in different sources of information. As a second step, it is necessary to extract all relevant PubMed articles whose terms are associated with each of the elements of the previously extracted disease list DMTL, through PubMed's API Entrez (https://www.ncbi.nlm.nih.gov/home/develop/api) that we have configured to obtain, if they exist, the 100 most relevant articles of each MeSH term consulted. Specifically, for each article we retrieve: 1) abstract, 2) authors' names, 3) unique identifier in PubMed and PubMed Central, 4) doi (digital object identifier), 5) title, 6) associated MeSH terms and 7) keywords. The workflow for extracting texts from PubMed documents is shown in **Fig. 4**.

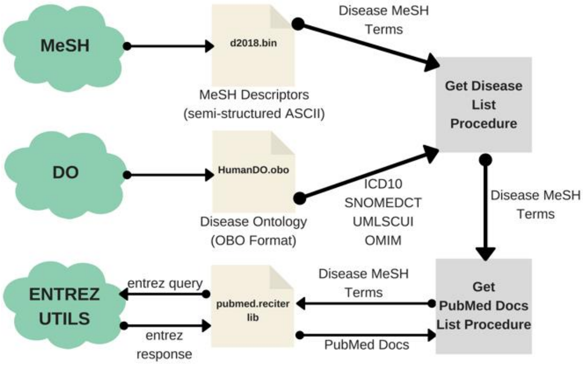


Fig. 3. PubMed Text Extraction Procedure workflow

E. MayoClinic

According to the official website, MayoClinic (https://www.mayoclinic.org) is a nonprofit organization committed to clinical practice, education and research, providing expert, whole-person care to everyone who needs it. In the USA, it is considered one of the best Hospitals and Health Systems; and beyond being a provider of health services (https://www.mayoclinic.org/es-es/about-mayo-clinic/quality/rankings, https://www.mayoclinic.org/es-es/about-mayo-clinic/office-diversity-inclusion), it also dedicates efforts to research and publication (https://www.mayo.edu/research/publications) of scientific knowledge through doctors and researchers. It thus does not come as a surprise that this online source contains relevant medical information on diseases and their phenotypic manifestations. The MayoClinic website actually contains a freely accessible list of diseases; each one of them is described in terms of an overview, the symptoms it presents, causes, diagnoses, treatments, the types of doctors who treat it and their departments or centers, among other information regarding related services provided by MayoClinic.

By mid 2018 this important source of medical knowledge had 1,170 articles on diseases (https://www.mayoclinic.org/es-es/diseases-conditions). These articles are structured by means of sections: "Symptoms", "Causes" and "Diagnostic", in which we have detected a greater concentration of phenotypic textual content (paragraphs and lists).

In contrast to Wikipedia, MayoClinic does not have disease codes in external medical databases, and its list of diseases is considerably shorter; yet, it presents the advantage of being an official and curated website, being thus easier to obtain information.

F. Challenges

Mining information from the sources previously described entails several computational challenges, which may be boiled down to one requirement for the DISNET system: the need of a high versatility in data acquisition. We here review such challenges, as these partly explain the adopted software solution.

First of all, the mapping disease-webpage may take different forms. Specifically, it is one to one for Wikipedia and MayoClinic, as all the information of a disease is included in a single page; but it becomes one to many for PubMed, in which multiple articles are available for each single concept. Consulting the latter thus requires a more complex procedure.

Secondly, in most of the cases the information we want to access is always available: a user can for instance access Wikipedia or MayoClinic at any time. There are nevertheless exceptions: Freebase (which aims to be part of DISNET project in a near future) is no longer available online, and a dump has instead to be downloaded and installed locally. The system should thus be able to access both online and offline documents.

Thirdly, and as one may expect, the specific structure of each source of information is different – i.e. a page of Wikipedia has not the same structure of a PubMed article. This requires further flexibility, in terms of the development of a modular structure with specific crawlers for each source.

Finally, it is worth noting that, while here we have only considered texts, much information is available in different medias, like plain text, HTML, PDF, Word or Excel files. While not implemented at this stage, the system should be flexible enough to accommodate such sources in the future.

G. Data Retrieval and Knowledge Extraction

This section describes the general architecture of the DISNET system, including the data extraction and the subsequent knowledge extraction. In the sake of clarity, such architecture is further depicted in Fig. 5.

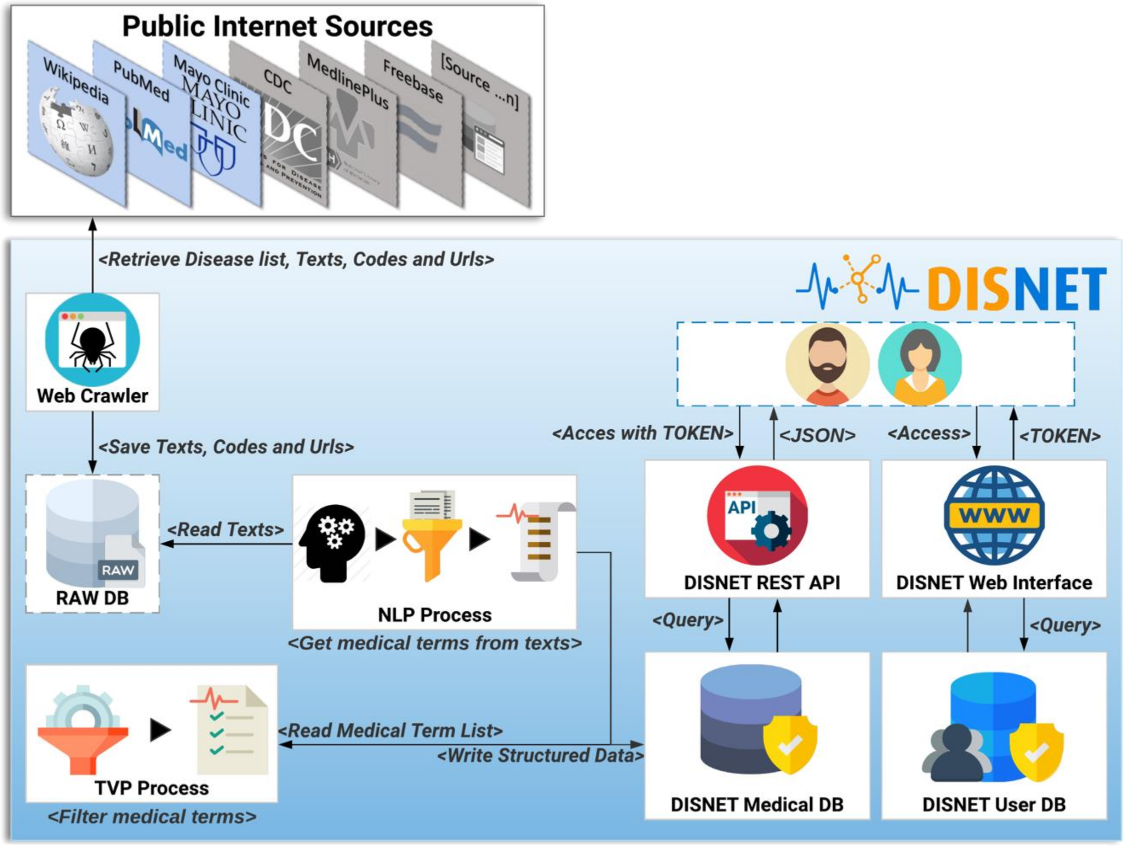


Fig. 5. DISNET Architecture/Workflow

1) The Extraction Process

The first step of the DISNET pipeline is in charge of retrieving the information from the sources previously identified and described. For each one of this, and before running the actual web crawler, the "Get Disease List Procedure" (GDLP) component is responsible for obtaining the list of diseases to be mined, thus providing links to all available disease related documents. For example, the GLDP associated to Wikipedia articles makes use of the SPARQL query (http://bit.ly/get_diseases_query_sparql); similarly, the links for the PubMed's articles are retrieved through a list of MeSH terms. However, in the case of MayoClinic, the terms are retrieved by scrapping strategies.

Once the URL list has been collected, the "Web Crawler" (WC) module is in charge of connecting to each of the hyperlinks and extracting the specific text that describes the phenotypical manifestations, as well as the links (references) contained within the texts (https://jsoup.org). In addition, and whenever possible, it attempts to extract information related to the coding of diseases, i.e. the codes used to identify the disease in different databases or existing data vocabularies. Currently it is able to retrieve information from more than 5,500 articles in Wikipedia, from 229,160 article abstracts in PubMed and from 1,176 articles in MayoClinic. The information mined by WC is stored in an intermediate database called "Raw DB", which contains the raw unprocessed text.

The next step within the pipeline is called "NLP Process" (NLPP). This component is responsible for: i) reading all the texts of a snapshot, and ii) obtaining for each text a list of relevant clinical concepts/terms, discarding any unrelated paragraphs or words. At the moment NLPP uses Metamap (33, 34) as a Natural Medical Language Processing tool to extract clinical terms of interest – see online NLP Tools and Configuration section (http://disnet.ctb.upm.es/apis/disnet#NLP_Tools_and_Configuration).

The output of the NLP process is stored in the "DISNET Medical DB" (DMDB) database. It stores, in a structured way, the medical concepts that have been obtained by the NLPP, as well as any information required to track the origin of such concepts – in order to track any error that may later be detected. Therefore, and to summarize, the information stored in a structured way in DMDB is: i) the medical concepts with their location, information and semantic types, ii) the texts from which they were extracted and the links by them contained, iii) the sections which the texts belong to, iv) the document or documents describing the disease (Web link) and v) the disease identifiers codes in different vocabulary or databases. Additional information, as the day of the extraction and the source, is further saved.

Before reaching the last step of the process, it is important to highlight the nature of the information hitherto stored. Specifically, the system has not extracted only signs or symptoms of a disease, but instead medical terms that we believe may be phenotypic manifestations of disease. It is thus necessary to filter those that are not relevant for the objective initially described.

Having clarified this, the next component of the pipeline, the "TVP Process" TVPP, reads all the concepts of a snapshot - source pair and filters them. This process is responsible for determining whether these UMLS medical terms are really phenotypic manifestations, and for storing the results back in the DMDB. TVPP is based on the Validation Terms Extraction Procedure that was developed, implemented and tested by Rodriguez-Gonzalez et al (14). The results of this component (a purification of concepts) are thus those validated terms that we will consider as true phenotypic manifestations of diseases.

The DISNET extraction process (IEPD), i.e. the process of retrieving and storing information about diseases, basically ends here. Nevertheless, for the sake of providing an accessible and user-friendly way of retrieving and manipulating this information, DISNET also offers a REST-based interface. This is described in detail in the system website (http://disnet.ctb.upm.es/apis/disnet); also refer to Sec. 4.3 for an application example.

4. Results

This section describes how the medical concepts data set is built, for then validating and analyzing its content. We finally present how the system could be used by means of the description of a basic use case.

A. Construction of the DB

The database in the DISNET system contains information recovered from three sources of information: Wikipedia, PubMed and MayoClinic. From Wikipedia we have nine snapshots, from February 1, 2018 to September 15th, 2018, for PubMed we have one snapshot, that of April 3, 2018 and for MayoClinic we have one snapshot, that of August 15th, 2018. Within the system it is possible to consult, for each snapshot and source, the total number of articles with medical terms, the total number of medical terms found, the number of processed texts, the total number of retrieved codes, and the total number of semantic types found (http://bit.ly/wikipedia_knowledge_csv, http://bit.ly/pubmed_knowledge_csv).

When summing that sources, the system counts with 5,954 diseases, 2,127 medical terms from UMLS (SNOMED-CT) and 17 semantic types, which can be consulted online (http://bit.ly/DISNET_diseases_txt, http://bit.ly/UMLS_terms_txt, http://bit.ly/semantic_types_txt).

Wikipedia snapshots are built using the configurations that are available online (http://bit.ly/snapshot_settings_txt). We have obtained a list of 9,857 articles catalogued as diseases in Wikipedia according to DBpedia (http://bit.ly/wikipedia_diseases_articles_incorrect_in_dbpedia_txt), from which we obtained 4,455 articles with at least one text referring to phenotypic knowledge of the disease, or at least one code to an external information source, 4,178 of which were found to be relevant medical concepts (http://bit.ly/wikipedia_articles_with_relevant_terms_txt).

The snapshot for PubMed has been built using the configuration described online (http://bit.ly/snapshot_settings_txt). This snapshot has been built on top of a list of 2,354 MeSH terms (http://bit.ly/mesh_terms_human_diseases_txt) referring to human diseases, but only for 2,213 MeSH terms did we obtain information (199,013 scientific articles in total, i.e. about 0.71% of the 28 million articles existing in PubMed (http://bit.ly/list_pubmed_papers_txt)) and of each of these PubMed articles obtained, only in 174,900 were abstracts found and only in 105,252 were relevant medical terms found. The snapshot for MayoClinic has been built on top a list of 1,176 diseases, but only on 1,082 did we obtain relevant medical terms. Fig. 6 and Fig. 7 presents some basic database statistics at an aggregated level as well as by source (for Wikipedia and PubMed). Some notable differences can be observed; for instance, the five most common terms for Wikipedia are *Pain*, *Lesion*, *Magnetic resonance imaging*, *Malnutrition* and *Convultions*, while for PubMed these are *Lesion*, *Magnetic resonance imaging*, *Malnutrition*, *Inflammation* and *Infection*. Similarly, the three diseases with the greatest number of concepts in Wikipedia are *Kawasaki disease*, *Hypoglycemia* and *Anorexia nervosa,* while for PubMed these are *Familial hypocalciuric hypercalcemia*, *Urinary bladder disease* and *Henoch–Schönlein purpura*.

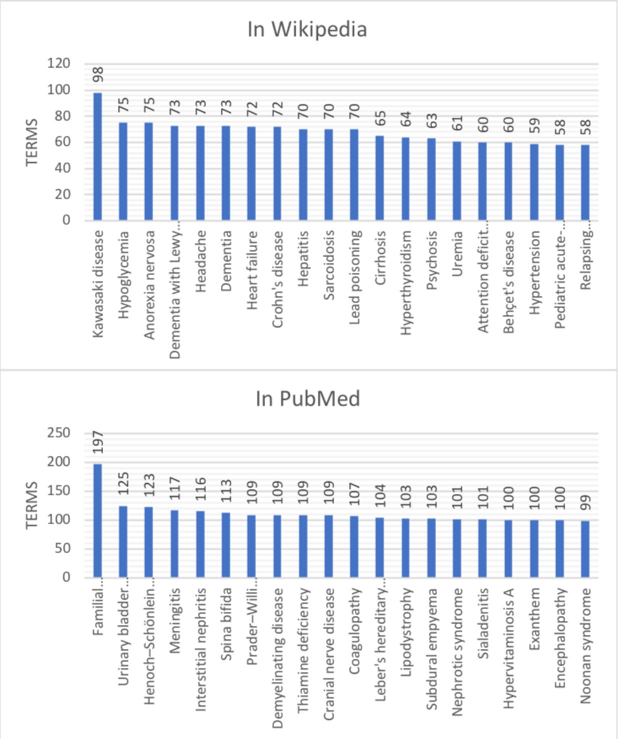


Fig. 6 Basic data base statistics (diseases with more medical terms. Comparison of PubMed and Wikipedia)

B. Data evaluation of the DB

In this section, we discuss the results of the validation process we executed on the system, to ensure the relevance of the diagnostic knowledge (valid medical diagnostic terms) generated through our NLP process (Metamap and TVP). The evaluation has been made on both Wikipedia and PubMed mined texts due the relevance of both sources.

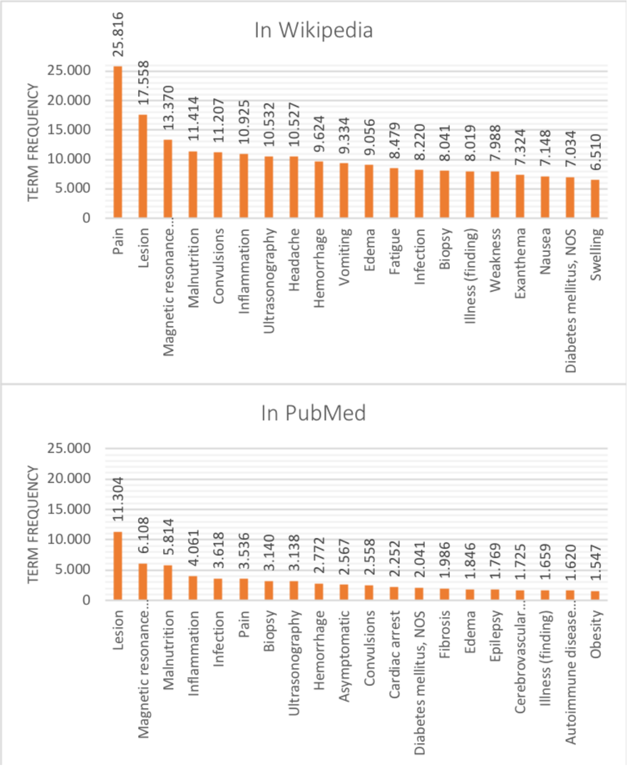
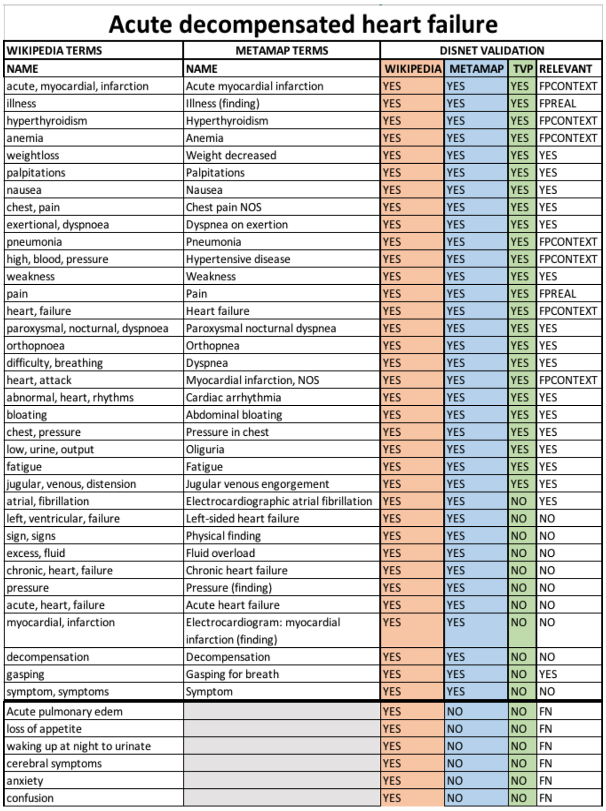


Fig. 7 Basic data base statistics (most common medical terms)

The validation for Wikipedia was carried out on the February 1, 2018 version, selecting 100 diseases at random with the only condition of having at least 20 valid medical terms. Similarly, the validation for PubMed has been done on the April 3, 2018 version, selecting a random sample of 500 article abstracts. These snapshots were performed at different times, and therefore with different configurations – the latter ones can be viewed online (http://bit.ly/snapshot_settings_txt). During the validation of Wikipedia, we detected that the initial configuration of Metamap did not find all the necessary medical concepts: for instance, Anxiety, Stress, Amnesia, Bulimia and other psychological concepts were missing. We therefore decided to update the initial list of semantic types to be detected (see online NLP Tools and Configuration section (http://disnet.ctb.upm.es/apis/disnet#NLP_Tools_and_Configuration) by adding the following elements: **Intellectual Product**, **Mental Process**, **Mental or Behavioral Dysfunction**, **Pathologic Function**, **Congenital Abnormality**.

The evaluation was conducted through a thorough manual analysis of the basic data. For each disease obtained from Wikipedia or PubMed we compared: (1) the list of medical terms extracted manually from the texts describing the disease; (2) the list of medical terms extracted by Metamap from the same texts; (3) the value (TRUE=valid or FALSE=invalid) resulting from the TVP process for each term found by Metamap; (4) the value of diagnostic relevance for a disease for each term. An example of the format of the Acute decompensated heart failure validation sheet for Wikipedia is shown in Fig. 8.

It is possible to note that an additional column was also present, called RELEVANT, and which synthesises all the information available about the relevance of a term to a disease. The possible values of this column are defined as:



Acute decompensated heart failure

WIKIPEDIA TERMS	METAMAP TERMS	DISNET VALIDATION			
NAME	NAME	WIKIPEDIA	METAMAP	TVP	RELEVANT
acute, myocardial, infarction	Acute myocardial infarction	YES	YES	YES	FPCONTEXT
illness	Illness (finding)	YES	YES	YES	FPREAL
hyperthyroidism	Hyperthyroidism	YES	YES	YES	FPCONTEXT
anemia	Anemia	YES	YES	YES	FPCONTEXT
weightloss	Weight decreased	YES	YES	YES	YES
palpitations	Palpitations	YES	YES	YES	YES
nausea	Nausea	YES	YES	YES	YES
chest, pain	Chest pain NOS	YES	YES	YES	YES
exertional, dyspnoea	Dyspnea on exertion	YES	YES	YES	YES
pneumonia	Pneumonia	YES	YES	YES	FPCONTEXT
high, blood, pressure	Hypertensive disease	YES	YES	YES	FPCONTEXT
weakness	Weakness	YES	YES	YES	YES
pain	Pain	YES	YES	YES	FPREAL
heart, failure	Heart failure	YES	YES	YES	FPCONTEXT
paroxysmal, nocturnal, dyspnoea	Paroxysmal nocturnal dyspnea	YES	YES	YES	YES
orthopnoea	Orthopnea	YES	YES	YES	YES
difficulty, breathing	Dyspnea	YES	YES	YES	YES
heart, attack	Myocardial infarction, NOS	YES	YES	YES	FPCONTEXT
abnormal, heart, rhythms	Cardiac arrhythmia	YES	YES	YES	YES
bloating	Abdominal bloating	YES	YES	YES	YES
chest, pressure	Pressure in chest	YES	YES	YES	YES
low, urine, output	Oliguria	YES	YES	YES	YES
fatigue	Fatigue	YES	YES	YES	YES
jugular, venous, distension	Jugular venous engorgement	YES	YES	YES	YES
atrial, fibrillation	Electrocardiographic atrial fibrillation	YES	YES	NO	YES
left, ventricular, failure	Left-sided heart failure	YES	YES	NO	NO
sign, signs	Physical finding	YES	YES	NO	NO
excess, fluid	Fluid overload	YES	YES	NO	NO
chronic, heart, failure	Chronic heart failure	YES	YES	NO	NO
pressure	Pressure (finding)	YES	YES	NO	NO
acute, heart, failure	Acute heart failure	YES	YES	NO	NO
myocardial, infarction	Electrocardiogram: myocardial infarction (finding)	YES	YES	NO	NO
decompensation	Decompensation	YES	YES	NO	NO
gasping	Gasping for breath	YES	YES	NO	YES
symptom, symptoms	Symptom	YES	YES	NO	NO
Acute pulmonary edem		YES	NO	NO	FN
loss of appetite		YES	NO	NO	FN
waking up at night to urinate		YES	NO	NO	FN
cerebral symptoms		YES	NO	NO	FN
anxiety		YES	NO	NO	FN
confusion		YES	NO	NO	FN

Fig. 8 Disease Acute decompensated heart failure sheet validation from the Wikipedia snapshot of February 1st, 2018

(1) RELEVANT = **YES**. If (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = (YES or NO)), that is, it is considered to be a valid medical concept for the diagnosis of a disease.

(2) RELEVANT = **NO**. If (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = NO), that is, it is considered to be a medical concept that is nonspecific, and thus too general to be helpful in the diagnosis of a disease.

(3) RELEVANT = **FPREAL**. If (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES). The term **is not relevant** because it is considered to be a nonspecific, general concept that does not make sense for diagnosis, even though Metamap has detected it and the TVP process has evaluated it as a diagnostic term. For example, in an excerpt from Acute decompensated heart disease on Wikipedia: "*Other cardiac symptoms of heart failure include chest pain/pressure and palpitations…*", Metamap has detected **Chest pain** and **Pain** from "*chest pain*", both were marked as TRUE by TVP but the concept dismissed by nonspecific and general was Pain.

(4) RELEVANT = **FPCONTEXT**. If (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES). The term **is not relevant** because it is outside the diagnostic context, even though Metamap has detected it and the TVP process has evaluated it as a diagnostic term. In other

words, this term has been obtained from texts whose content is outside the diagnostic context. For example, in an excerpt from *Acute decompensated heart failure* disease on Wikipedia: *"Other well recognized precipitating factors include anemia and hyperthyroidism…"*, Metamap has detect ***Anemia*** and ***Hyperthyroidism*** which are medical terms but in context we dismiss them because they are risk factors for that disease.

(5) RELEVANT = **FN**. If (WIKIPEDIA = YES) & (METAMAP = NO) & (TVP = NO). These terms were manually detected in the texts, but Metamap failed in recognising them.

The cases (3) and (4) above define situations in which the detected term is esteemed to be of no relevance, and as such represent cases of false positives. It is nevertheless necessary to discriminate the reason behind such error, which can be because: i) it is a very general, nonspecific concept whose definition does not represent and contributes nothing to the diagnosis (FP_REAL), or ii) because the term is a medical term that is out of place with respect to the context that is narrated in the text – in other words, it could be a valid diagnostic term but not for the disease that is under validation or in the context in which have been described and therefore should be discarded (FP_CONTEXT).

Using this information for all diseases and terms, true positive (**TP**), false positive (**FP**), true negative (**TN**) and false negative (**FN**) rates were computed in order to calculate precision, recall and F1 score values as metrics to measure the performance of DISNET system. The mean values for these parameters are depicted in Fig. 9. The **TP** is all terms with (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES) & (RELEVANT = YES). As previously explained, the **FP** is composed of two parts, being the total FP the sum of **FP_REAL** + **FP_CONTEXT**:

- **FP_REAL** = (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES) & (RELEVANT = FPREAL).
- **FP_CONTEXT** = (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = YES) & (RELEVANT = FPCONTEXT).

FN is also composed of two parts, i.e. **FN_METAMAP** + **FN_TVP**.

- **FN_METAMAP** = (WIKIPEDIA = YES) & (METAMAP = NO) & (TVP = NO) & (RELEVANT = FN). These are terms that Metamap has not found.
- **FN_TVP** = (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = NO) & (RELEVANT = YES). These are terms that TVP has validated as false while being relevant.

Finally, the **TN** measures the TVP process (WIKIPEDIA = YES) & (METAMAP = YES) & (TVP = NO) & (RELEVANT = NO). In the **Table 1** are reported the values obtained for Wikipedia and PubMed.

Detailed results for each disease are available online, for Wikipedia (http://bit.ly/wikipedia_validation_sheets) and for PubMed (http://bit.ly/pubmed_validation_sheets), including the list of terms manually extracted from the relevant texts of the articles, the matching with the list of terms provided by Metamap, the result of the TVP process for each term and the value of relevance as annotated by our researchers.

Table 1. Total values from the February 1st, 2018 snapshot of Wikipedia and the April 3, 2018 snapshot of PubMed

Parameter	Wikipedia	PubMed
TP	(32.07%) 1,867.000	(31.30%) 722.000
FP	(10.51%) 612.000	(17.30%) 399.000
FPREAL	192.000	99.000
FPCONTEXT	420.000	300.000
TN	(30.75%) 1,790.000	(32.69%) 754.000
FN	(26.64%) 1,551.000	(18,69%) 431.000
FN_METAMAP	926.000	206.000
FN_TVP	625.000	225.000
TOTAL	(100%) 5,820.000	(100%) 2,306.000
PRECISION	0.753	0.644

Results indicate that our NLP (Metamap + TVP) process is sufficiently reliable, with an accuracy of 0.753 (confidence interval of (0.730, 0.775)) for Wikipedia and of 0.644 (confidence interval of: (0.606, 0.680)) for PubMed (Fig. 9). The results of the calculations of these parameters for each disease can be viewed online for Wikipedia (http://bit.ly/wikipedia_individual_validation_results_csv) and for each abstract in PubMed (http://bit.ly/pubmed_individual_validation_results_csv).

About the results for **FP** presented in **Table 1**, we can say that they are mainly due to the configuration used for Metamap for the extraction of terms, extended in successive extractions to avoid leaving out terms that are relevant for the detection of diseases.

Thus, one of the last extensions in the search terms added the semantic types Mental or Behavioral Dysfunction and Intellectual Product; thanks to this extension, important symptoms have been detected for certain diseases, which were not detected before, such as: *Anxiety*, *Bulimia*, *Anorexy*, *Stress*, etc. We believe that it is better to discard those terms that are not relevant than to omit those that are relevant to a disease.

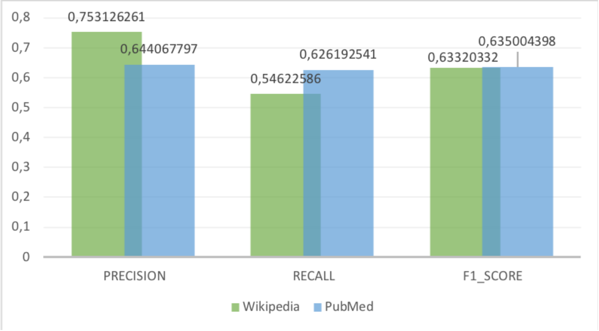


Fig. 9 Validation metrics comparative

It is further interesting to observe the large difference in the false positive rates between Wikipedia (10.51%) and PubMed (17.30%). We speculate that this is due to the concretion of articles. Accordingly, in Wikipedia, articles referring to one disease refer almost exclusively to that particular disease, and thus include no irrelevant terms – with a few exceptions related to differential diagnoses. Nevertheless, this is not the case of PubMed articles as a significant part of them are not so specific. Many are the articles describing real medical cases, where the symptoms are those displayed by a given patient, plus others referring to congenital diseases of the patient, or even diseases that he/she previously possessed. Consequently, the same PubMed article includes symptoms of many different diseases that, although being true medical terms and thus being recognized by Metamap, are not relevant to the disease under analysis.

For **TN,** we must also take into account that most of the terms extracted by Metamap as relevant have been purged by TVP, which has been in charge of determining which terms are relevant and which are not, so that the vast majority of terms extracted by Metamap that are not relevant to the disease have been classified in this way by TVP (30.75% for Wikipedia and 32.69% for PubMed).

In addition, we have observed that most of the true negative terms in both Wikipedia and PubMed are constant, and include: *indicated*, *syndrome*, *disease*, *illness*, *infected*, *sing*, *symptoms*, *used to*, etc.

Finally, **FN** are those terms that are relevant to the disease in question, but that have not been detected by Metamap; note that these have been manually extracted for the validation process. The vast majority of **FN** are formed by complex expressions of the language, so their detection is challenging for any NLP tool. We can further observe that the difference in the ratio of false negative between Wikipedia (26.64%) and PubMed (18.69%) is 7.95%. We believe that this difference is mainly due to the forms of expression used in both sources, with Wikipedia being more discursive, as opposed to the scientific style of PubMed.

In synthesis, we can conclude that a clear relationship can be observed between the performance of the system and the nature of the underlying data source. Specifically, while PubMed is an exclusively medical source, created, written and edited by specialists in the field, Wikipedia is a source of public information, written by anyone who has access to the web, so that the articles in it contained can be written by medical students or just users with some knowledge in the field, whose expressions cannot be assimilated to those of specialists who write PubMed. Considering that the tool used by DISNET for the extraction of medical terms (Metamap) is a medical tool, it is not surprising that it displays a greater capacity for the recognition of medical terms, as opposed to more colloquial terms formed by more complex phrases; thus, there are terms such as "*Swollen lymph glads under the jaw*", or "*sensation of swelling in the area of the larynx*"... that Metamap cannot recognize.

C. A use case

To illustrate the possible use of the DISNET system, we here present a simple use case, which consists of the creation of several basic DISNET queries, and the visualization of the corresponding results.

Creation of DISNET queries

For the sake of simplicity, we will here focus on two of the most important characteristics of DISNET: **i)** the ability to create relationships between diseases according to their phenotypic similarity (**C1**) and **ii)** the ability to increase/improve the phenotypic information of diseases by means of periodic extractions of knowledge (**C2**).

The scenario C1 implies obtaining data for two diseases, which we suspect may share symptoms; we will here use "Influenza" and "Gastroenteritis". The resulting DISNET queries are:

(1) disnet.ctb.upm.es/api/disnet/query/**disnetConceptList**?source=**wikipedia**&version=**2018-08-15**&diseaseName=**Influenza**&matchExactName=**true**

(2) disnet.ctb.upm.es/api/disnet/query/**disnetConceptList**?source=**pubmed**&version=**2018-04-03**&diseaseName=**Influenza**&matchExactName=**true**

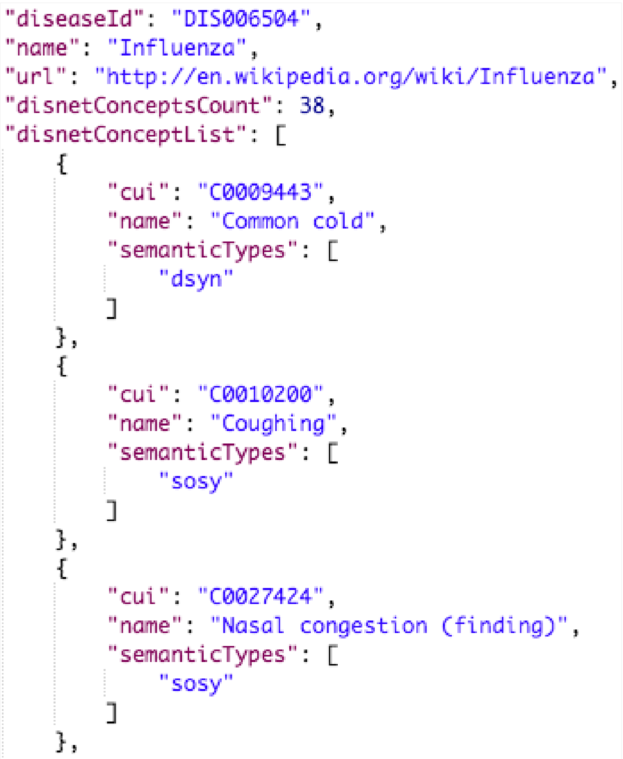
(3) disnet.ctb.upm.es/api/disnet/query/**disnetConceptList**?source=**wikipedia**&version=**2018-08-15**&diseaseName=**Gastroenteritis**&matchExactName=**true**

(4) disnet.ctb.upm.es/api/disnet/query/**disnetConceptList**?source=**pubmed**&version=**2018-04-03**&diseaseName=**Gastroenteritis**&matchExactName=**true**

We have here used the DISNET query "**disnetConcepList**", which allows retrieving the list of "**DISNET Concepts**" associated with a given disease. The parameters of this query include: "**diseaseName**", with the name of the disease; "**matchExactName**", to indicate that the search by disease name is exact; and "**source**" and "**snapshot**", to respectively indicate the source and snapshot we want to consult. In this case, we selected to consult the two sources Wikipedia and PubMed, and respectively the snapshots of August 15th, 2018 and April 3rd, 2018. Note that the result will consists of four total lists, two for each disease. To illustrate, Fig. 10 shows an extract of the response from the query (3).

As for the scenario C2, it requires retrieving data for a disease whose list of symptoms may have changed with time, i.e. either increased or decreased. As an example, we considered the disease "Acrodynia", and executed the following DISNET queries:

(1) disnet.ctb.upm.es/api/disnet/query/**disnetConceptList**?source=**wikipedia**&version=**2018-02-01**&diseaseName=**Acrodynia**&matchExactName=**true**



```
"diseaseId": "DIS006504",
"name": "Influenza",
"url": "http://en.wikipedia.org/wiki/Influenza",
"disnetConceptsCount": 38,
"disnetConceptList": [
    {
        "cui": "C0009443",
        "name": "Common cold",
        "semanticTypes": [
            "dsyn"
        ]
    },
    {
        "cui": "C0010200",
        "name": "Coughing",
        "semanticTypes": [
            "sosy"
        ]
    },
    {
        "cui": "C0027424",
        "name": "Nasal congestion (finding)",
        "semanticTypes": [
            "sosy"
        ]
    },
```

Fig. 10 Answer to the DISNET query "disnetConcepList" C1.(1)

(2) disnet.ctb.upm.es/api/disnet/query/**disnetConceptList**?source=**wikipedia**&version=**2018-02-15**&diseaseName=**Acrodynia**&matchExactName=**true**

Note that, as in C1, we have here used the query "**disnetConceptList**"; nevertheless, we have here executed it twice, on the same disease (**Acrodynia**) and two different snapshots: February 1st, 2018 and February 15th, 2018.

Visualization of the result of the DISNET queries.

Once the results of the query have been retrieved, the next natural step is their visualization; while the actual output format may vary according to the needs of each specific project, for the sake of clarity we here created a graph representation by using as external tool called Cytoscape (http://www.cytoscape.org). In both scenarios (i.e. C1 and C2) we generated relationships between diseases and their symptoms, with the aim of visualizing the value and scope of the medical data stored and processed by DISNET. In Fig. 11.b we see the relationship between the Influenza and Gastroenteritis diseases on one hand (highlighted in white rectangles), and the set of symptoms on the other. Symptoms were obtained from two different sources, specifically Wikipedia and PubMed: relationships are then respectively represented by red and blue edges. Common symptoms are merged by the layout algorithm in the center of the graph; the medical terms that are not common among the two diseases, on the contrary, form a peripheral shell. Note that "**Influenza**" has 59 DISNET Concepts and "**Gastroenteritis**" has 48, 19 of which are in common.

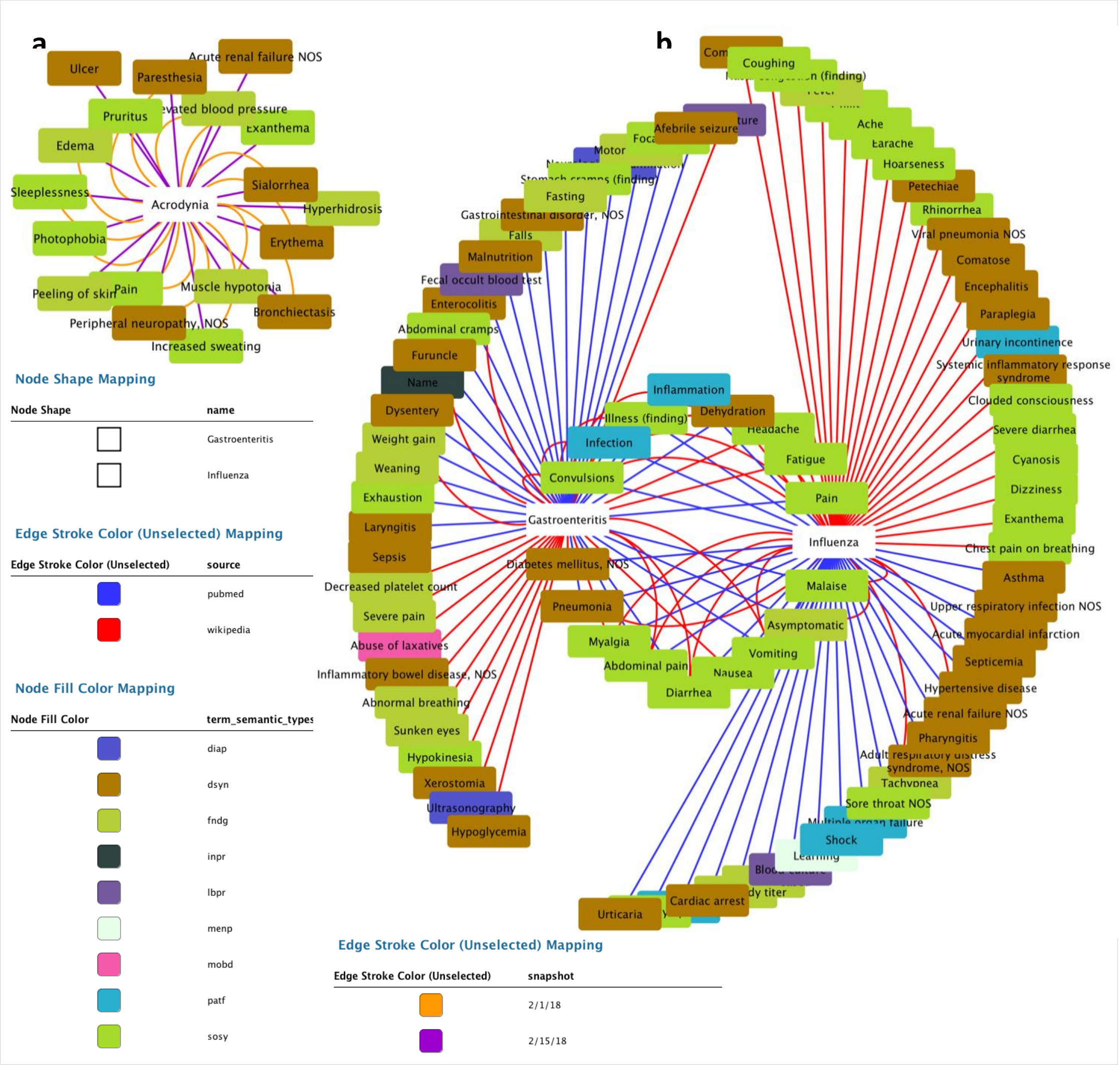


Fig. 11 a) Network of graphs representing the evolution of phenotypic knowledge in Wikipedia and b) Network of graphs representing similar medical terms between two diseases.

In Fig. 11.a we observe the network representation of the disease "**Acrodynia**" and of its 18 medical terms, 15 of which were found in the snapshot of February 1st, 2018 and three new ones in that of February 15th, 2018. This is thus an example of an increase in phenotypic knowledge.

This simple use case illustrates how the DISNET system allows generating a network of diseases and their symptoms on a large scale, and that it provides the right environment to know how diseases are related according to their phenotypic manifestations. By applying similarity algorithms, such as Cosine (3,11,35) or the Jaccard index (36), it is possible to estimate the similarity between two diseases, and thus to focus further medical analyses on those pairs showing a large overlap. These features will be also implemented as native features in next DISNET release.

Conclusions and Future Work

This work presented the DISNET system, starting from its underlying conception, up to its technical structure and data workflow. DISNET allows retrieving knowledge about the signs, symptoms and diagnostic tests associated with a disease. It is not limited to a specific category (all the categories that the selected sources of information offer us) and clinical diagnosis terms. It further allows to track the evolution of those terms through time, being thus an opportunity to analyse and observe the progress of human knowledge on diseases. We also presented the DISNET REST API, which aims at sharing the retrieved information with the wide scientific community. We further discussed the validation of the system, suggesting that it is good enough to be used to extract diseases and diagnostically-relevant terms. At the same time, the evaluation also revealed that improvements could be introduced to enhance the system's reliability.

Among the potential lines of future works, priority will be given to increasing the number of information sources, by including other websites like Medline Plus or CDC. Secondly, we are considering the possibility of extending the TVP procedure, by adding new data sources, with the aim of increasing the number of validation terms and hence of reducing the number of false negatives. Note that this could also be partly achieved by resorting to a different NLP tool to process the input texts, as for example to Apache cTakes (37). Future implementations of DISNET also aim to provide ways to automatically compute the similarity between diseases (by using already mentioned and well-known similarity metrics), extending the DISNET platform to include biological and drug information and developing new visualization strategies, among others.

Funding

This paper is supported by Consejo Nacional de Ciencia y Tecnología (CONACYT) and European Union's Horizon 2020 research and innovation programme under grant agreement No. 727658, project IASIS (Integration and analysis of heterogeneous big data for precision medicine and suggested treatments for different types of patients).

Conflict of interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

References

1. Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, *Nat. Genet.*, **33 Suppl**, 228–237.
2. Goh, K.-I., Cusick, M. E., Valle, D., et al. (2007) The human disease network, *Proc. Natl. Acad. Sci.*, **104**, 8685–8690.
3. Zhou, X., Menche, J., Barabási, A.-L., et al. (2014) Human symptoms-disease network, *Nat. Commun.*, **5**, 4212.
4. Barabási, A.-L., Gulbahce, N. and Loscalzo, J. (2010) Network medicine: a network-based approach to human disease, *Nat. Rev. Genet.*, **12**, nrg2918.
5. Chen, Y., Zhang, X., Zhang, G., et al. (2015) Comparative analysis of a novel disease phenotype network based on clinical manifestations, *J. Biomed. Inform.*, **53**, 113–120.
6. Valle, E. G. del, Garcia, G. L., Santamaria, L. P., et al. (2018) Disease networks and their contribution to disease understanding and drug repurposing. A survey of the state of the art, *bioRxiv*, 415257.
7. Hirschhorn, J. N. and Daly, M. J. (2005) Genome-wide association studies for common diseases and complex traits, *Nat. Rev. Genet.*, **6**, nrg1521.
8. Suthram, S., Dudley, J. T., Chiang, A. P., et al. (2010) Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets, *PLoS Comput. Biol.*, **6**, e1000662.
9. Xu, W., Jiang, X., Hu, X., et al. (2014) Visualization of genetic disease-phenotype similarities by multiple maps t-SNE with Laplacian regularization, *BMC Med. Genomics*, **7 Suppl 2**, S1.
10. Loscalzo, J., Kohane, I. and Barabasi, A.-L. (2007) Human disease classification in the postgenomic era: A complex systems approach to human pathobiology, *Mol. Syst. Biol.*, **3**, 124.
11. Li, X., Zhou, X., Peng, Y., et al. Network Based Integrated Analysis of Phenotype-Genotype Data for Prioritization of Candidate Symptom Genes https://www.hindawi.com/journals/bmri/2014/435853/ (accessed Nov 17, 2017).
12. Yıldırım, M. A., Goh, K.-I., Cusick, M. E., et al. (2007) Drug—target network, *Nat. Biotechnol.*, **25**, nbt1338.
13. Côté, R. A. and Robboy, S. (1980) Progress in Medical Information Management: Systematized Nomenclature of Medicine (SNOMED), *JAMA*, **243**, 756–762.
14. Rodríguez-González, A., Martínez-Romero, M., Costumero, R., et al. In *9th International Conference on Practical Applications of Computational Biology and Bioinformatics*; Advances in Intelligent Systems and Computing; Springer, Cham, 2015; pp. 79–87.
15. Miller, N., Lacroix, E.-M. and Backus, J. E. B. (2000) MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service, *Bull. Med. Libr. Assoc.*, **88**, 11–17.
16. Zhou, X., Li, Y., Peng, Y., et al. (2014) Clinical phenotype network: the underlying mechanism for personalized diagnosis and treatment of traditional Chinese medicine, *Front. Med.*, **8**, 337–346.
17. Savova, G. K., Masanz, J. J., Ogren, P. V., et al. (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, *J. Am. Med. Inform. Assoc. JAMIA*, **17**, 507–513.
18. Oellrich, A., Collier, N., Groza, T., et al. (2016) The digital revolution in phenotyping, *Brief. Bioinform.*, **17**, 819–830.
19. Leaman, R., Islamaj Dogan, R. and Lu, Z. (2013) DNorm: disease name normalization with pairwise learning to rank, *Bioinforma. Oxf. Engl.*, **29**, 2909–2917.
20. Barabási, A.-L., Gulbahce, N. and Loscalzo, J. (2010) Network medicine: a network-based approach to human disease, *Nat. Rev. Genet.*, **12**, nrg2918.
21. Hassan, M., Makkaoui, O., Coulet, A., et al. 2015; p. 184.
22. Okumura, T., Meñez, D. A. and Abayawickrama, T. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; 2015; pp. 938–939.

23. Aibar, E. La ciencia de la Wikipedia. *Rev. Mètode* **2017**.
24. Azzam, A., Bresler, D., Leon, A., et al. (2017) Why Medical Schools Should Embrace Wikipedia: Final-Year Medical Student Contributions to Wikipedia Articles for Academic Credit at One School, *Acad. Med.*, **92**, 194–200.
25. Friedlin, J. and McDonald, C. J. (2010) An evaluation of medical knowledge contained in Wikipedia and its use in the LOINC database, *J. Am. Med. Inform. Assoc. JAMIA*, **17**, 283–287.
26. Shafee, T., Masukume, G., Kipersztok, L., et al. (2017) Evolution of Wikipedia's medical content: past, present and future, *J Epidemiol Community Health*, jech-2016-208601.
27. Cohen, N. Editing Wikipedia Pages for Med School Credit. *N. Y. Times* **2013**.
28. Matheson, D. and Matheson, C. (2017) Open Medicine Journal Wikipedia as Informal Self-Education for Clinical Decision-Making in Medical Practice, *Open Med. J.*, **4**, 1–25.
29. Heilman, J. M. and West, A. G. (2015) Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language, *J. Med. Internet Res.*, **17**.
30. Auer, S., Bizer, C., Kobilarov, G., et al. In *The Semantic Web*; Lecture Notes in Computer Science; Springer, Berlin, Heidelberg, 2007; pp. 722–735.
31. pubmeddev Home - PubMed - NCBI https://www.ncbi.nlm.nih.gov/pubmed/ (accessed Feb 16, 2018).
32. Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., et al. (2018) A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts, *PLOS Comput. Biol.*, **14**, e1005962.
33. Aronson, A. R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, *Proc. AMIA Symp.*, 17–21.
34. Rodríguez González, A., Costumero Moreno, R., Martínez Romero, M., et al. (2015) Extracting diagnostic knowledge from MedLine Plus: a comparison between MetaMap and cTAKES Approaches, *Curr. Bioinforma.*, **375**, 1–7.
35. van Driel, M. A., Bruggeman, J., Vriend, G., et al. (2006) A text-mining analysis of the human phenome, *Eur. J. Hum. Genet. EJHG*, **14**, 535–542.
36. Hoehndorf, R., Schofield, P. N. and Gkoutos, G. V. (2015) Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases, *Sci. Rep.*, **5**, srep10888.
37. Savova, G. K., Masanz, J. J., Ogren, P. V., et al. (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, *J. Am. Med. Inform. Assoc. JAMIA*, **17**, 507–513.