Abstract— Nowadays there is a huge amount of medical information that can be retrieved from different sources, both structured and unstructured. Internet has plenty of textual sources with medical knowledge (books, scientific papers, specialized web pages, etc.), but not all of them are publicly available. Wikipedia is a free, open and worldwide accessible source of knowledge. It contains more than 150,000 articles of medical content in the form of texts (non-structured information) that can be mined. The aim of this work is to study whether the evolution of information contained in Wikipedia medical articles can be used in a research context. The study has been focused on extracting the elements, from Wikipedia disease articles, that can be used to guide a diagnosis process, support the creation of diagnostic systems, or analyze the similarities between diseases, among others.

Keywords— wikipedia; diseases; diagnosis; wikipedia evolution

I. INTRODUCTION

Wikipedia is an online source of information, open and collaborative. At the moment of analysis, it contains more than 43 million of pages, more than 5 million articles (in English)\(^1\) and is released in 298 languages\(^2\). It is one of the most visited web sites\(^3\), being the English version the biggest and most active.

\(^1\) https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia
\(^2\) https://en.wikipedia.org/wiki/List_of_Wikipedias
\(^3\) https://www.alex.com/topsites
especially relevant in tasks like the creation of disease-phenotype knowledge bases [5] or of automated diagnostic systems [6] [7]. In spite of this, little attention has been devoted to the use of Wikipedia as a source of diagnostic information, and to its validation.

This paper tackles this issue, by analysing the diagnostic information contained in Wikipedia medical articles, and by showing the results of applying a text-mining process over Wikipedia disease articles. The paper studies the content of Wikipedia evaluating several metrics, such as the number of disease articles and findings, and the information in them contained.

II. RELATED WORK

There are several works in the biomedical field that have used Wikipedia as their main information source. In 2013 a paper was published [7], describing a web application created to store information about diseases and their symptoms as extracted from Wikipedia, with the aim of detecting diseases suffered by specific users based on the tracking of the searches performed on the Web. With data from the same year (2013-2014), in 2015 a second study was published [8] that used Wikipedia to forecast the influenza outbreak. For that, the authors did a study of the registry access to Wikipedia articles. In the same line, in 2014 was created a system [9] for the monitoring and prognosis of diseases by, again, analysing the access to Wikipedia articles. In 2015 a study made use of the “2014 West African Ebola virus disease epidemic” article to get information regarding death counts and hospitalization counts in the narratives; it further proposed the use of Wikipedia as a community-driven open-source emerging disease, detection, monitoring and repository system, with the rationale that current surveillance systems suffer from disadvantages such as reporting lags and antiquated technology.

Other works have been focused on the improvement of Wikipedia articles by means of Natural Language Processing (NLP) techniques, and by applying automatic evaluation of Wikipedia medical articles [10]. In the line of NLP and Name Entity Recognition (NER), there are several works that implement recognized identifiers of medical concepts such as MetaMap [4] and cTAKES [5]. Works like [11] and [12] make comparisons between them and calculate their precisions obtaining good results. It is also worth noting that the above described NLP have also been used to analyse texts from various sources of medical information such as MedlinePlus [13] and Wikipedia [14] [15].

In the diagnosis context, a system was proposed to automatically infer the most probable diagnosis from clinical narratives [16]. The authors tested their system with texts from Wikipedia, Mayo clinic, Freebase and UMLS, and found that the systems based on Wikipedia and Mayo clinic content reached respectively a 60% and 70% of correct diagnoses. This suggests that those sources were very relevant for finding the correct diagnosis. In a similar approach, very aligned with our study, a recent work has studied the feasibility of using Wikipedia for extracting disease terms, aimed at disease understanding [17], with promising results.

Other studies have been focused in the creation of medical ontologies, using repositories such as Wikipedia, under the assumption that the latter “provides a valuable resource from which to mine structured information” [18]. Another study in the same direction used Wikipedia for the creation of a clinical thesaurus [19].

The knowledge contained in Wikipedia has been also used to enrich SNOMED-CT [6] to obtain medical terms synonyms. As a result of this approach, 183,100 new synonyms were retrieved with an accuracy of 85.6%, demonstrating again the powerful value of the knowledge contained in Wikipedia [20].

The analysis of the related work leads to the conclusion that Wikipedia has been used as a trustful source of knowledge. It contains online information that can be retrieved and used for many purposes, including medical research ones. Using Wikipedia to obtain medical terms regarding the diagnosis of

![Fig. 1. Pipeline extraction process](https://www.snomed.org/snomed-ct)
diseases is an unusual, different, interesting and unique approach.

III. MATERIALS AND METHODS

The aim of this paper is to perform an analysis of the “diagnostic knowledge” information contained in Wikipedia disease articles and its evolution. For this work, we consider as diagnostic knowledge any element related to a disease that allows physicians, within the diagnosis process, to determine or discard possible diseases. Stricto sensu, these include the phenotypic manifestations, i.e. findings, signs and symptoms. However, other elements such as diagnostic procedures, laboratory tests and results, i.e. blood test, cell count, liver function test, among others are also considered diagnostic knowledge, as they allow guiding the diagnosis process [11].

We have created a pipeline (Fig. 1) that is executed two times per month to extract the diagnostic knowledge contained in Wikipedia disease articles. This process allows the creation of snapshots of the Wikipedia data, thus enabling the study of their evolution. The pipeline performs the following steps: i) retrieves a list of available Wikipedia diseases articles; ii) retrieves texts from Wikipedia articles; iii) performs text-mining over such texts and iv) analyses the obtained results.

A. Information Source

The information source is composed of all Wikipedia articles categorized as diseases in DBpedia. The diseases list were extracted from DBpedia [21] using a SPARQL query. The retrieved list is composed only of Wikipedia articles categorized as “Disease”. In the real Wikipedia pages, a disease article is, in most of the cases, structured using different sections that include description, disease codification, causes, diagnosis, treatment, prognosis and clinical manifestations. For extracting diagnostic knowledge elements, the relevant sections are: “Signs and symptoms”, “Symptoms and causes”, “Signs”, “Symptoms”, “Causes”, “Cause”, “Causes of injury”, “Diagnosis”, “Diagnostic”, “Diagnostic approach”, “Symptoms of …”, “Causes of …” and “Presentation” and for extracting disease codifications the relevant section is: “infobox”. The information retrieved was: 1) the texts contained in the aforementioned sections; 2) links contained in the texts and 3) disease codifications, available in the infoboxes, which normally are located at the beginning and at the foot of the article.

B. Data retrieval and knowledge extraction

The pipeline for extracting information has been executed twice per month, being the first execution on February 1st, 2018 and the last one reflected in this study on February 1st, 2019, thus yielding a total of 25 snapshots. As explained, the list of articles categorized as diseases in Wikipedia was retrieved from DBpedia executing the DBpedia Disease List Extraction (DDLE) process: a procedure that performs a SPARQL query\(^7\) against DBpedia [22]. The second step consists in a process named Wikipedia Text Extraction (WTE) developed using Jsoup API\(^8\) (see [13] for further information), which extracts the texts from the sections and the disease codifications available in the infoboxes using web-scrapping and stores them in a MySQL database.

The next step applies the Extraction of Medical Findings (EMF) process: a NLP procedure to retrieve the relevant elements. This process is based in a two-step approach. Firstly, MetaMap [23] performs a Name Entity Recognition (NER) over the texts and retrieves the relevant ones based on the arguments provided (sources to be used to identify terms and UMLS semantic types\(^9\) to be detected). Secondly, our Term Validation Procedure (TVP) module, described in [13], is applied. TVP validates the terms retrieved by MetaMap minimizing the number of terms incorrectly detected. Finally, all results are stores in a MySQL database.

The source used to identify medical terms in the NER process was SNOMED-CT in English. The semantic types passed as argument to MetaMap were: acab, anab, comd, cgab, dsyn, emod, fnud, mobd, neop, patf, sosy, lbpr, lbtr, clna, menp, diap. Regarding semantic types, from now on all the elements found by the EMF process will be named as Diagnostic Knowledge Element (DKE), independently on the associated semantic type.

More details about Data Recovery and Knowledge Extraction pipelines (DDLE, WTE, EMF, NER and TVP) are also available [24].

IV. RESULTS

The execution of the pipeline yielded 25 snapshots from February 1st, 2018 until February 1st, 2019. Table I reports some key features describing the last snapshot; similar information for the other time points is available in Table V.

| Key properties | Count |
|----------------|-------|
| Articles categorized as disease in DBpedia (DDLE process) | 11,084 |
| Articles that contained diagnostic knowledge elements applying TVP validation | 4,692 |
| Diagnostic knowledge elements (DKE) found (no duplicates) | 14,722 |
| UMLS semantic types found (no duplicates) | 17 |
| Disease codes found (no duplicates) | 20,237 |
| External vocabularies found (no duplicates) | 59 |
| Number of texts (no duplicates) | 40,913 |
| Number of links found in the texts | 193,860 |

As can be observed there is a significant difference between the number of articles categorized in DBpedia as diseases (11,084) and the number of those that finally contained DKEs.

---

\(^7\) https://midas.ctb.upm.es/gitlab/disnet/paperWikipedia/blob/master/getDiseases.sparql

\(^8\) https://jsoup.org/

\(^9\) https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml
and hence are considered as real disease articles (4,692 – 42.33%).

There are two explanations for this phenomenon: i) the articles did not contain DKEs; or ii) the articles are incorrectly catalogued – i.e. in DBpedia were catalogued with the class "Disease", but the content of the Wikipedia article did not contain information about a disease. We consider a DBpedia listed article as "genuine" only if it has, at least, one code from an external vocabulary/classification system (ICD, OMIM, etc.) in the data retrieved from DBpedia (DDLE process) or in the data retrieved from the scrapping of Wikipedia (WTE process). Those articles that return 0 codes in both sets have been classified as non-disease articles, implying that DBpedia had catalogued them incorrectly. More information about those articles that have been retrieved as diseases in DBpedia have is available as online supplementary material.10

The number of external vocabularies/classification systems found in Wikipedia articles is 6511. This information is relevant in order to get more information about a specific disease in an external database. Results indicate that Wikipedia is not only useful because of its content; it’s also as a bridge to reach external sources.

With respect to the scope of this work, the main metric is the number of retrieved DKEs, which is 14,722. Table II reports additional information about this metric, including the three diseases with more and less DKE. The average number of DKE is 11.72.

Table III shows what are the most and less frequent DKEs and the number of appearances.

Another metric is the number of UMLS semantic types found. These semantic types correspond to a grouping of DKEs into homogeneous and conceptually-related families. The list of the 14,722 distinct DKEs that have been retrieved in all the snapshots with their corresponding semantic types is also available online12. The distribution of the DKEs grouped by their semantic type is available in Table IV and the distribution of the DKEs grouped by relevant sections is available in Fig. 2. As can be seen, the semantic types with the highest number of DKE are dsyn, sosy, fndg, diaf, patf and mobd, something that was to be expected given the type of information that we wanted to retrieve. On the other side, we have comd and lbtr as the less frequent semantic types.

### Table II. Disease articles with more and less disease DKEs (Snapshot: February 1st, 2019)

| Metric | Disease                  | URL                           | Number of unique DKEs |
|--------|--------------------------|-------------------------------|-----------------------|
| Disease with more DKE (1st) | Kawasaki disease | https://en.wikipedia.org/wiki/Kawasaki_disease | 98                    |
| Disease with more DKE (2nd) | Cerebral palsy     | https://en.wikipedia.org/wiki/Cerebral_palsy | 77                    |
| Disease with more DKE (3rd) | Hypoglycemia       | https://en.wikipedia.org/wiki/Hypoglycemia | 76                    |
| Disease with less DKE (1st) | Tuberculoma        | https://en.wikipedia.org/wiki/Tuberculoma | 1                     |
| Disease with less DKE (2nd) | True hermaphrodism | https://en.wikipedia.org/wiki/True_hermaphrodism | 1                     |
| Disease with less DKE (3rd) | Mediastinal lymphadenopathy | https://en.wikipedia.org/wiki/Mediastinal_lymphadenopathy | 1                     |

### Table III. More and less frequent DKEs (total appearances of the term, total disease articles where the term appears) (Snapshot: February 1st, 2019)

| Term                        | Number of appearances |
|-----------------------------|-----------------------|
| Pain (C0030193)             | 1,927                 |
| Lesion (C0221198)           | 1,355                 |
| Neoplasms (C0027651)        | 1,179                 |
| Musset's sign (C0277927)    | 1                     |
| Pruriitus of vagina (C0042256) | 1                     |
| Stage fright (C0395003)     | 1                     |

https://midas.ctb.upm.es/gitlab/disnet/paperWikipedia/blob/master/DBpediaDiseaseArticles

https://midas.ctb.upm.es/gitlab/disnet/paperWikipedia/blob/master/ExternalVocabularyFound.csv
Regarding information evolution, we have measured several metrics in the different snapshots: number of articles retrieved by DBpedia as diseases (DBpDis), number of Wikipedia articles that contain DKE applying TVP validation (WRDArt), of the number of DKE found by MetaMap (not applying TVP validation but removing duplicates) (WRawDF), number of texts (WTxt), number of semantic types found (again: before applying TVP) (WST), number of external codes found in Wikipedia (WExCd), number of external sources found in Wikipedia (WExtSrc) and number of links found in the texts (WLink). Table V shows the evolution of these metrics in the different snapshots.

| Snapshot   | DBpDis | WRDArt | WRawDF | WTxt   | WST | WExCd | WExtSrc | WLink |
|------------|--------|--------|--------|--------|-----|--------|---------|-------|
| 2018-02-01 | 8,161  | 3,625  | 13,332 | 30,932 | 17  | 19,203 | 61      | 148,640 |
| 2018-02-15 | 8,161  | 3,631  | 13,356 | 31,099 | 17  | 19,161 | 60      | 149,073 |
| 2018-03-01 | 8,161  | 3,636  | 13,393 | 31,199 | 17  | 19,116 | 60      | 149,799 |
| 2018-03-15 | 8,161  | 3,644  | 13,431 | 31,463 | 17  | 19,102 | 60      | 151,065 |
| 2018-04-01 | 9,857  | 3,841  | 13,598 | 32,956 | 17  | 19,302 | 62      | 157,585 |
| 2018-04-15 | 9,858  | 3,846  | 13,608 | 33,016 | 17  | 19,289 | 62      | 158,107 |
| 2018-05-01 | 9,858  | 3,860  | 13,630 | 33,153 | 17  | 19,272 | 62      | 158,910 |
| 2018-05-15 | 9,858  | 3,871  | 13,692 | 33,246 | 17  | 19,270 | 62      | 159,262 |
| 2018-06-01 | 9,858  | 4,047  | 13,853 | 34,376 | 17  | 19,246 | 58      | 163,710 |
| 2018-06-15 | 9,858  | 4,061  | 14,169 | 34,584 | 18  | 19,224 | 58      | 164,709 |
| 2018-07-01 | 9,858  | 4,062  | 13,980 | 34,953 | 17  | 19,209 | 59      | 165,732 |
| 2018-07-15 | 9,858  | 4,069  | 13,991 | 35,044 | 17  | 19,207 | 58      | 166,127 |
| 2018-08-01 | 9,858  | 4,083  | 14,013 | 35,153 | 17  | 19,202 | 58      | 166,814 |
| 2018-08-15 | 9,858  | 4,087  | 14,019 | 35,247 | 17  | 19,201 | 58      | 167,300 |
| 2018-09-01 | 9,858  | 4,090  | 14,022 | 35,296 | 17  | 19,333 | 58      | 167,584 |
| 2018-09-15 | 9,858  | 4,091  | 14,022 | 35,293 | 17  | 19,339 | 58      | 167,688 |
| 2018-10-01 | 9,858  | 4,093  | 14,032 | 35,343 | 17  | 19,351 | 58      | 167,928 |
| 2018-10-15 | 9,858  | 4,097  | 14,050 | 35,449 | 17  | 19,360 | 58      | 168,497 |
| 2018-11-01 | 9,858  | 4,101  | 14,064 | 35,522 | 17  | 19,360 | 58      | 168,972 |
| 2018-11-15 | 9,858  | 4,106  | 14,089 | 35,657 | 17  | 19,366 | 58      | 169,745 |
| 2018-12-01 | 9,858  | 4,111  | 14,140 | 35,778 | 17  | 19,355 | 59      | 170,344 |
| 2018-12-15 | 9,858  | 4,119  | 14,159 | 35,949 | 17  | 19,365 | 59      | 171,203 |
| 2019-01-01 | 9,858  | 4,128  | 14,182 | 36,111 | 17  | 19,365 | 59      | 172,082 |
| 2019-01-15 | 9,858  | 4,139  | 14,185 | 36,192 | 17  | 19,366 | 59      | 172,456 |
| 2019-02-01 | 11,084 | 4,692  | 14,722 | 40,913 | 17  | 20,237 | 59      | 193,860 |

V. DISCUSSION

Before starting a discussion about the obtained results, it is necessary to mention how this information has been validated. First, we have relied on the accuracy of MetaMap as NER tool. MetaMap has been widely tested, and is considered as a valid system to perform NLP processes, and more specifically NER over medical texts [25] [26] [27] [28]. On the other hand, the data pipeline includes the validation of the terms retrieved by MetaMap using our TVP module [13]. This latter element has been previously validated using MedLine Plus medical texts as information source, which are content and structure-wise qualitatively similar to Wikipedia’s ones. Similarly, an analysis of the performance of MetaMap in extracting DKEs from Wikipedia texts have been already done previously [24]. Due to this, it was estimated that an ex-novo validation was not necessary.

Time plays an important role because it allows to observe the evolution of the knowledge stored in Wikipedia. As can be
observed in Table V, there is a progression in most of the elements. Only Semantic Types (WST) remains with no significant changes – as is to be expected, since this number could only change by finding new DKEs of different types. All metrics (except the WExCd, which presents an irregular behavior) are monotonically increasing, thus suggesting that the modifications of the Wikipedia articles result in the inclusion of new information that is captured by our pipeline.

From a global point of view, all these metrics support the feasibility of using Wikipedia as a source of medical information. We have been able to obtain information from 5,059 articles that are catalogued as diseases and contained diagnostic information, of which 4,775 were diseases with at least one medical term (after applying TVP). While noteworthy, this number should be compared with those obtained in other medical sources: ≈4,500 for MeSH\(^\text{13}\) [29]; ≈7,000 for OMIM\(^\text{14}\)[30]; and ≈10,500 for DisGeNET\(^\text{15}\). While prima facie DisGeNET includes information about twice the diseases of Wikipedia, two additional aspects have to be taken into account. First of all, Wikipedia is an alive system, frequently updated by its users, and whose knowledge base is constantly increasing - as shown in Table V. Secondly, different websites rely on different vocabularies: a same disease can then be classified differently, split into different subtypes, merged, and so forth. Therefore, the high number found for DisGeNET does not necessarily imply a larger body of information, at least in what is referred to DKE information.

The average number of obtained disease findings is another important factor. 1,918 (40.87\%) diseases have at least 11 identified disease findings. This is positive from the point of view of creating, for instance, diagnosis systems based on the information here retrieved. The list of these diseases and their number of disease findings can be found online\(^\text{16}\).

In ![Error! No se encuentra el origen de la referencia.](https://www.omim.org/statistics/entry) we see a graph comparing the first generated snapshot against the last one. We confirm that during these last seven months we have found an increase in the information contained in the articles catalogued as diseases and, above all, that each new contribution has enough medical textual content to allow us to identify an increasing number of disease findings.

In spite of all the above, there are a few issues that are worth discussing. First of all, we have found many articles that were catalogued as diseases in DBpedia, but that were discarded after applying our filter method (no external codes in DBpedia nor Wikipedia). Some of them are related to the medical domain (for example: “famous outbreaks”, “health crisis”, etc.), but they are not "diseases". Some examples of those articles are “1924 Los Angeles pneumonic plague outbreak”, “1852–60 cholera pandemic”; “1863–75 cholera pandemic” and “2013 Swansea measles epidemic”. However, many others are not even related with medical information in any way (for example: “2008 Western Australian gas crisis”, “2010 in film”, “2003 Wimbledon Championships – Women's Singles”). Another drawback is the detection of potentially relevant Wikipedia articles that have returned no results in the EMF process. The reason lies in the fact that we have limited the text-mined sections to “Signs and symptoms”, “Causes”, “Diagnosis” and “Presentation”; but these are empty in some articles (e.g.: “Hereditary sensory and autonomic neuropathy”, “17q21.31 microdeletion syndrome”, “2,4 Dienoyl-CoA reductase deficiency”). On the other hand, we have found disease articles that have not been structured using the sections that we are using to retrieve terms (e.g.: “Bleb (medicine)”, “Oligactylid” or “Meteoropathy”), and also one should take care of articles referred to diseases that are no longer catalogued as such (e.g.:
with its predecessor and more details is available as online\textsuperscript{17}. The first column compares the disease list and indicates whether the number of diseases has been increased (+) or decreased (-). The second column indicates the overlap percentage between the common diseases of both snapshots (how many diseases are in both snapshots). And the last column shows the percentage diseases variation, in other words, get the percentage of disease have had a change in some their diagnostic elements.

We can observe that the percentage of variation of disease has very high values in the snapshots of May 1\textsuperscript{st}, 2018, June 1\textsuperscript{st}, 2018, July 1\textsuperscript{st}, 2018; these increases of variation can be due to that in those intervals of time the articles were edited to add phenotypic content. We can also observe high values in the number of diseases in the snapshots of April 1\textsuperscript{st}, 2018, June 1\textsuperscript{st}, 2018 and February 1\textsuperscript{st}, 2019; this situation can be due to several factors associated to DBpedia, as new disease articles have been identified; new articles that were not catalogued as diseases; inconsistency in the results of the HTTP/SPARQL server service; and also factors associated with Wikipedia such as articles being edited adding both phenotypic content and external disease identifiers, causing an article to stop going unnoticed. In particular, the increase in the number of diseases in the snapshot of February 1\textsuperscript{st}, 2019 is due to the inclusion of the DBPedia-Live results.

However, even taking those drawbacks into account, results are promising: if we discard the Wikipedia articles not considered as disease based on our filter (11,084) we have found medical findings in 4,775 (43.08\%), which represents a really high number. The numbers that have here been presented and analysed also provide a good benchmark for the aim established: obtain as much diagnostic information as possible from Wikipedia articles.

Finally, we can conclude that Wikipedia is an accurate and relevant source of medical information due to the collaborative quality control that Wikipedia nowadays enforces, the introduction of wrong information is really difficult, and the creation of snapshots and the analysis of the difference between them could allow us to discard irrelevant data. On the other hand, the constant updates make Wikipedia a highly up to date source of information – an essential feature for research. We can also add that a database of diseases and diagnostic medical terms can be useful in various types of research such as diagnostic systems, decision support systems, building disease networks, etc.

| Snapshot | No. Disease | % Overlapping | % Disease Variation |
|----------|-------------|---------------|---------------------|
| 2018-02-01 | -           | -             | -                   |
| 2018-02-15 | +6          | 99.83         | 2.12                |
| 2018-03-01 | +8          | 99.78         | 3.20                |
| 2018-03-15 | +15         | 99.59         | 2.98                |
| 2018-04-01 | +203        | 94.71         | 3.66                |
| 2018-04-15 | +10         | 99.74         | 2.55                |
| 2018-05-01 | +21         | 99.46         | 3.10                |
| 2018-05-15 | +13         | 99.66         | 12.34               |
| 2018-06-01 | +179        | 95.58         | 7.45                |
| 2018-06-15 | +18         | 99.56         | 44.40               |
| 2018-07-01 | +10         | 99.75         | 45.26               |
| 2018-07-15 | 9           | 99.78         | 2.44                |
| 2018-08-01 | +19         | 99.53         | 2.61                |
| 2018-08-15 | +6          | 99.85         | 2.67                |
| 2018-09-01 | +7          | 99.83         | 1.86                |
| 2018-09-15 | +5          | 99.88         | 1.84                |
| 2018-10-01 | +4          | 99.90         | 2.18                |
| 2018-10-15 | +9          | 99.78         | 1.81                |
| 2018-11-01 | +12         | 99.71         | 2.93                |
| 2018-11-15 | +9          | 99.78         | 2.29                |
| 2018-12-01 | +10         | 99.76         | 3.24                |
| 2018-12-15 | +9          | 99.78         | 2.31                |
| 2019-01-01 | +12         | 99.71         | 3.04                |
| 2019-01-15 | +11         | 99.73         | 1.89                |
| 2019-02-01 | +558        | 88.12         | 3.29                |

**TABLE VI. COMPARISON BETWEEN DISEASE SNAPSHOTS**

\textsuperscript{17} https://midas.ctb.upm.es/gitlab/disnet/paperWikipedia/blob/master/comparisonBetweenDiseaseDsnapshots.xlsx
VI. FUTURE WORK
Future work will be focused on a three-fold strategy.

First of all, we consider that the information extracted to perform this study can be a valuable resource for researchers working in fields like disease understanding, disease networks or diagnosis systems. For this reason, we are currently developing a platform that will allow accessing all the data that have been processed and summarized in this study [24]. Secondly, effort will be devoted to solve the drawbacks that were here found, including: improving and applying the filter method to discard no-disease articles; improving the extraction of texts for those articles without the predefined sections; and discarding deprecated articles.

A final future work will be to perform this analysis again in a longer time scale, trying to analyse the same data with more information, and trying also to find an explanation or correlation for some of the metrics described.

ACKNOWLEDGMENT
This paper is supported by Mexican 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).

REFERENCES
[1] T. Shafee, G. Masukume, L. Kipersztock, D. Das, M. Häggström, and J. Heilman, “Evolution of Wikipedia’s medical content: past, present and future,” J Epidemiol Community Health, p. jech-2016-208601, Aug. 2017.
[2] R. Al Tamime, R. Giordano, and W. Hall, “Humans and bots in controversial environments: A closer look at their interactions,” Qatar Found. Annu. Res. Conf. Proc., vol. 2018, no. 3, p. ICTPD425, Mar. 2018.
[3] A. Azzam et al., “Why Medical Schools Should Embrace Wikipedia: Final-Year Medical Student Contributions to Wikipedia Articles for Academic Credit at One School,” Acad. Med., vol. 92, no. 2, pp. 194–200, Feb. 2017.
[4] J. M. Heilman and A. G. West, “Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language,” J. Med. Internet Res., vol. 17, no. 3, Mar. 2015.
[5] R. Xu, L. Li, and Q. Wang, “Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature,” Bioinformatics, vol. 29, no. 17, pp. 2186–2194, Sep. 2013.
[6] A. Rodríguez-González and G. Alor-Hernández, “An approach for solving multi-level diagnosis in high sensitivity medical diagnosis systems through the application of semantic technologies,” Comput. Biomed. Res., vol. 43, no. 1, pp. 51–62, Jan. 2013.
[7] L.-W. Ku, W.-L. Li, and T.-C. Chang, “Disease Detection and Symptom Tracking by Retrieving Information from the Web,” 2013.
[8] K. S. Hickmann et al., “Forecasting the 2013–2014 Influenza Season Using Wikipedia,” PLOS Comput. Biol., vol. 11, no. 5, p. e1004239, May 2015.
[9] N. Generous, G. Fairchild, A. Deshpande, S. Y. D. Valle, and R. Friedhorsky, “Global Disease Monitoring and Forecasting with Wikipedia,” PLOS Comput. Biol., vol. 10, no. 11, p. e1003892, Nov. 2014.
[10] V. Cozza, M. Petrocchi, and A. Spognardi, “A Matter of Words: NLP for Quality Evaluation of Wikipedia Medical Articles,” in Web Engineering, 2016, pp. 448–456.
[11] A. Rodríguez Gonzalez, R. Costumero Moreno, M. Martínez Romero, M. D. Wilkinson, and E. Menasalvas Ruiz, “Extracting diagnostic knowledge from MedLine Plus: a comparison between MetaMap and cTAKES Approaches,” Curr. Bioinforma., vol. 375, no. 6, pp. 1–7, 2015.
[12] R. Reátegui and S. Ratté, “Comparison of MetaMap and cTAKES for entity extraction in clinical notes,” BMC Med. Inform. Decis. Mak., vol. 18, no. 3, p. 74, Sep. 2018.
[13] A. Rodríguez-González, M. Martínez-Romero, R. Costumero, M. D. Wilkinson, and E. Menasalvas-Ruiz, “Diagnostic Knowledge Extraction from MedlinePlus: An Application for Infectious Diseases,” in 9th International Conference on Practical Applications of Computational Biology and Bioinformatics, Springer, Cham, 2015, pp. 79–89.
[14] M. A. Mouriño-García, R. Pérez-Rodríguez, and L. E. Anido-Rifón, “A Bag of Concepts Approach for Biomedical Document Classification Using Wikipedia Knowledge,” Methods Inf. Med., vol. 56, no. 5, pp. 370–376, 2017.
[15] D. Zhang and D. He, “Enhancing Clinical Decision Support Systems with Public Knowledge Bases,” Data Inf. Manag., vol. 1, no. 1, pp. 49–60, Sep. 2017.
[16] Y. Ling, Y. An, and S. A. Hasan, “Improving Clinical Diagnosis Service by Integration of Structured and Unstructured Knowledge,” 2017.
[17] Eduardo P. García del Valle, Gerardo Lagunes García, Lucía Prieto Santamaría, Massimiliano Zanin, Alejandro Rodríguez González, and Ernestina Menasalvas Ruiz, “Evaluating Wikipedia as a source of information for disease understanding,” presented at the 31st IEEE CBMS International Symposium on Computer-Based Medical Systems, 2018.
[18] V. Pedro, R. Stefán Niculescu, and L. Lita, “Okinet: Automatic Extraction of a Medical Ontology From Wikipedia,” Jan. 2008.
[19] D. Milne, O. Medelyan, and I. H. Witten, “Mining Domain-Specific Thesauri from Wikipedia: A Case Study,” in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, 2006, pp. 442–448.
[20] D. R. Schlegel, C. Crownor, and P. L. Elkin, “Automatically Expanding the Synonym Set of SNOMED CT using Wikipedia,” Stud. Health Technol. Inform., vol. 216, pp. 619–623, 2015.
[21] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “DBpedia: A Nucleus for a Web of Open Data,” in The Semantic Web, Springer, Berlin, Heidelberg, 2007, pp. 722–735.
[22] J. Lehmann et al., “DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia,” p. 29.
[23] A. A. Ronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program,” Proc. AMIA Symp., pp. 17–21, 2001.
[24] G. L. Garcia, A. R. Gonzalez, L. P. Santamaria, E. G. P. Del Valle, M. Zanin, and E. M. Ruiz, “DISNET: Disease understanding through complex networks creation and analysis,” bioRxiv, p. 428201, Sep. 2018.
[25] J. D. Osborne, B. Gyawali, and T. Solorio, “Evaluation of YTEX and MetaMap for clinical concept recognition,” ArXiv14021668 Cs, Feb. 2014.
[26] P. Gooch and A. Roudsari, “A tool for enhancing MetaMap performance when annotating clinical guideline documents with UMLS concepts,” presented at the IDAMAP Workshop at 13th Conference on Artificial Intelligence in Medicine (AIME’11), Bled, Slovenia, 2011.
[27] L. Al-Safadi, R. Alomran, and F. Almutairi, “Evaluation of Metamap Performance in Radiographic Images Retrieval,” Res. J. Appl. Sci. Eng. Technol., vol. 6, pp. 4231–4236, Dec. 2013.
[28] S. Meyestre and P. J. Haug, “Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MXTs),” Stud. Health Technol. Inform., vol. 116, pp. 823–828, 2005.
[29] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, “Human symptoms-disease network,” Nat. Commun., vol. 5, p. 4212, Jun. 2014.
[30] J. Amberger, C. Bocchini, and A. Hamosh, “A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®),” Hum. Mutat., vol. 32, no. 5, pp. 564–567, May 2011.
[31] M. Morsey, J. Lehmann, S. Auer, C. Stadler, and S. Hellmann, “DBpedia and the live extraction of structured data from Wikipedia,” Program, vol. 46, no. 2, pp. 157–181, Apr. 2012.