Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Collecting Specialty-related Medical Terms: Development and Evaluation of a Resource for Spanish

Pilar López-Úbeda (plubeda@ujaen.es)
Universidad de Ja□en, Campus Las Lagunillas, s/n, 23071, Ja□en, Spain
Alexandra Pomares-Quimbaya
Pontificia Universidad Javeriana
Manuel Carlos Díaz-Galiano
University of Jaén
Stefan Schulz
Medical University of Graz

Research Article

Keywords: Natural Language Processing, Vocabulary, Medical sub-language, Clinical Specialty, Medical Sub-domain

Posted Date: December 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-118585/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

RESEARCH

Collecting specialty-related medical terms: Development and evaluation of a resource for Spanish

Pilar López-Úbeda[1*], Alexandra Pomares-Quimbaya[3], Manuel Carlos Díaz-Galiano[1] and Stefan Schulz[2]

*Correspondence:
plubeda@ujaen.es
[1] Universidad de Jaén, Campus Las Lagunillas, s/n, 23071, Jaén, Spain
Full list of author information is available at the end of the article

Abstract

Background: Ontologies and controlled vocabularies are fundamental resources for Information Extraction (IE) from clinical texts using Natural Language Processing (NLP). Standard language resources available in the healthcare domain such as the UMLS metathesaurus or SNOMED CT are widely used for this purpose. A known limitation is lexical ambiguity of clinical language, particularly regarding short forms. Much of them are unambiguous within documents limited to a given clinical specialty. For this and other NLP tasks, the identification of the specialty using document classification would be of great value.

Methods: This paper addresses this limitation by proposing and applying a method that automatically extracts Spanish medical terms classified and weighted per sub-domain, using Spanish MEDLINE titles and abstracts as input.

The hypothesis is biomedical NLP tasks benefit from collections of domain terms that are specific to clinical subdomains. We use Pubmed queries that generate sub-domain specific corpora from Spanish titles and abstracts, from which token n-grams are collected and metrics of relevance, discriminatory power, and broadness per sub-domain are computed.

Results: The generated term set, called SCOVACLIS (Spanish Core Vocabulary About Clinical Specialties), was made available to the scientific community and used in a text classification problem obtaining improvements of 6 percentage points in the F-measure compared to the baseline using Multilayer Perceptron, thus demonstrating the hypothesis that a specialized term set improves NLP tasks.

Conclusion: The creation and validation of SCOVACLIS support the hypothesis that specific term sets reduce the level of ambiguity when compared to a specialty-independent and broad-scope vocabulary.

Keywords: Natural Language Processing; Vocabulary; Medical sub-language; Clinical Specialty; Medical Sub-domain

1 Background and contributions

1.1 Limitations of language resources for the analysis of clinical narratives

Information extracted from clinical narratives has been used for a wide range of biomedical applications [1–4]. To this end, using Natural Language Processing (NLP) and Machine Learning (ML) techniques as part of clinical information extraction initiatives [5]. Information extraction supports a wide variety of clinical and research use cases, such as building disease-specific cohorts [6–8], processing and

analyzing mentions of signs and symptoms [9, 10], detecting and assessing adverse drug events and risks [11–13], extracting key information for reporting or quality assurance [14–17], among others.

Most of these studies rely on controlled vocabularies (CVs) in different flavors, known as dictionaries, lexicons, terminologies and ontologies. CVs are curated by experts and public bodies and used for specific purposes like disease or adverse event reporting, annotation of health records for billing, data collection for clinical research, and literature indexing. They exhibit large differences in scope, granularity and underlying formalism, ranging from term lists, over informal thesauri, e.g., Medical Subject Headings (MeSH) [18], single-hierarchy classification systems such as ICD-10 [19] and formal ontologies such as SNOMED CT [20, 21].

Biomedical CVs have known drawbacks: some of them have a broad scope (covering all medicine), but lack the granularity required by particular clinical specialties or services. Others are restricted to a specific semantic category like diseases or drugs. Even those that provide a good conceptual coverage of a domain, often lack sufficient lexical coverage, in particular for languages other than English. In fact, some studies have found out that semantic features recognized using CVs in clinical narratives are useless for some classification problems, probably because they are too broad [22]. Because of that, research projects often end up creating their own vocabulary.

The predominance of the English language in biomedical publications is reflected by the fact that many CVs are restricted to English, whereas others have only partial translations to other languages. The Unified Medical Language System UMLS [23, 24] has been addressing these problems for several decades by linking common identifiers to close to 200 international clinical terminologies, thus extending the representation of medical terms in several languages. More recently, the creation of interface terminologies, separated from, but linked to reference terminologies has been emphasized [25]. Whereas reference terminologies are defined as representing, first, a domain in terms of formally or informally defined and language-independent representational units (concepts, descriptors, or classes), interface terminologies focus on the collection of clinical terms as used in practice, found in clinical narratives, with a focus on sub-language and user aspects.

1.2 Contribution

Considering the availability of terminologies suited to cover a given clinical specialty in a certain natural language, a way to effectively apply NLP and ML technology to clinical narratives is to extend the scope of specialized language resources that are openly available for international and multilingual communities. Such resources are expected to accelerate development and use of NLP and ML technology in the clinical domain, and to reduce the complexity of NLP tasks that deal with the severe problem of semantic (especially lexical) ambiguity, particularly in tasks like concept extraction [26], co-reference resolution [27] and domain-specific text classification.

This study scrutinises the aforementioned problem from a Spanish language perspective. With its several dialects and varieties, Spanish is an official language in about 20 countries and spoken by around half a billion people. Using Spanish language content from the literature database PubMed we propose a method for automated harvesting of medical term sets that are highly specific for a clinical specialty.

Clinical specialties are subdivisions of the field of health care, such as represented by institutional divisions in health facilities, by medical research filed, and by undergraduate and post-graduate medical curricula. Instances of specialties are neurology, pathology, radiology, surgery and internal medicine, among many others, with sub-specialties like nephrology, diabetology, etc. There is no world-wide standard of clinical specialties, which explains high variations regarding subdivision of and overlaps between specialties.

An additional contribution of this paper is to customize existing CVs to clinical specialties. This is the reason why we created SCOVACLIS (Spanish Core Vocabulary About Clinical Specialties)[1].

1.3 Analysis of medical sublanguages in clinical narratives

Clinical sublanguage aspects have been addressed in many studies. Some of them identified differences in lexical and semantic patterns used within clinical specialties and categories of authors of clinical texts [28, 29], such as by applying clustering to a large set of clinical narratives using bag-of-words plus bag-of-UMLS features.

Bernhardt et al. [30] proposed a method for identifying prominent clinical specialties. In order to detect the specialties concerned with disease prevalence in the U.S. they connected mortality and morbidity information to the medical literature. Epidemiology-related terms were extracted from national reports and standardized with MeSH terms.

Zhang et al. [31] automatically identified clinically relevant new information in inpatient and outpatient notes and compared the quantity of redundant information across specialties and clinical settings. Their strategy uses semantic similarity techniques that compare the language model extracted from a note and the model extracted from previous notes of the same patient. Once they identified new and redundant information, they compared differences by specialty obtaining variations of redundancy from 60.7% and 68.3%. Pediatric notes have the most redundancy and radiology notes have the least redundancy.

Finally, some studies [30, 32] applied supervised learning-based NLP to develop a medical subdomain classifier. They produced different classifiers and tested 105 combinations of data representations of the medical notes. Their main conclusion was that a "deep learning architecture with distributed word representation yielded better performance, yet the shallow learning algorithm".

1.4 Vocabulary extraction methods in healthcare

Harvesting vocabulary from biomedical literature has been subject to many studies in biomedical NLP research. In an early review study, Krauthammer and Nenadic [33], distinguished three steps in a term identification process: term recognition, term classification, and term mapping. Meystre et al. [34] reviewed studies on clinical terminology extraction, most of which combined NLP techniques for term discovery with lexico-syntactic patterns for semantic relation discovery. The importance of terminologies to improve query expansion, information retrieval, ontology creation and data analysis was emphasised.

[1] *https://github.com/plubeda/scovaclis*

Another method for extracting terms in a molecular biology context was described by Takeuchi and Collier [35]. The extracted terms were classified into ten semantic categories (e.g. protein, virus, cell type), using a Support Vector Machine (SVM) model trained with a manually annotated MEDLINE abstract dataset. As concerns other sub-domains of medicine, there are studies that extract concepts that describe medical images [36] or medical curricula [37].

For languages other than English, Marciniak and Mykowiecka [38] obtained a list of single and multi-word terms used in hospital discharge documents written in Polish. They observed that 70% of the obtained terms were not included in the Polish MeSH.

Finally, Sandoval et al. [39] created a biomedical corpus of validated terms from Spanish, Arabic and Japanese, by using several tools for optimal exploitation of the information contained in the corpus.

In this paper, we describe the use of NLP techniques and tools for the selection of Spanish vocabulary from Pubmed for a specific goal, namely the construction of sets of terms that are maximally specific to clinical specialties. Such term sets can be useful for document classification, but also for enriching existing CVs by new terms. Our technique can be applied to any languages as long as the related PubMed records include enough titles and links to abstracts in the original language. Subsequently, we will suggest a refinement for the term identification task, for which different statistical measures will be proposed in order to improve the selection of candidate terms. These measures are based on the importance of each term in each specialty and its importance in the overall corpus. Finally, this list of terms is used as an extra feature in a multi-label classifier.

2 Methods

2.1 Overview

The method was designed to extract terms that are both *frequent in* and *specific for* a clinical specialty, thus resulting in a a maximally characteristic term set for each clinical specialty. The balance between term frequency discriminative power then would result, e.g., for clinical oncology, in the selection of the important and frequent terms "*carcinoma*" and "*tumor*", but laso of less frequent, but highly specific terms like "*leucemia mielomonocítica*" or "*dermatofibrosarcoma*". The main phases of this method are depicted in Figure 1.

Figure 1: Overview of the extraction method.

- The first phase of our method is the acquisition of domain corpora classified by clinical specialty (cf. Section 2.2.). Clinical texts might be the best source, but de-identified and therefore shareable clinical corpora, particularly for languages other than English do not exist, a well-known problem in biomedical NLP. This was the reason we decided to use Spanish content from PubMed, aware of the known terminology mismatch between scientific clinical language. For the selection of PubMed content by clinical specialty, several sources were combined. Aware that MeSH annotations would not suffice to indicate the

clinical specialty to which a PubMed record belongs, we also included authors' affiliation as a source of specialty-related information well as a group of terms directly related to the specialty (e.g. "skin disease" for dermatology).

- The second phase (cf. Section 2.3), term extraction, yields word n-grams from PubMed titles and abstracts. Not all n-grams are good term candidates, therefore this phase contains an important automatic cleansing step.
- The last phase, term consolidation (cf. Section 2.4), identifies the importance of each previously identified word n-gram for the chosen sub-language and, according to this analysis, applies a filtering algorithm that produces a final term set for each clinical specialty. The filtering algorithm detects and removes those n-grams that are common to all or almost all clinical specialties and whose relevance to those clinical specialties is similar. These "stop n-grams" are useless for differentiating between specialties.

2.2 Specialty-specific corpus acquisition

Figure 2 illustrates this phase in detail. The number of clinical specialties was 129 in the beginning (n). L refers to the number of specialties per hierarchical level. Out if 129 specialties, 45 belonged to the first level ($L1$) and 65 belonged to the second level ($L2$) of the hierarchy.

Figure 2: Clinical specialty selection process. Including the variations in the number of specialties (left) when applied to the Spanish case (right)

For defining the set of clinical specialties, the method starts extracting the branch of the MeSH thesaurus under "Disciplines and occupations" > "Health Occupations" > "Medicine" (Step 1).

Most subcategories obtained under this branch are clinical and paraclinical specialties, but others on education, research, epidemiology, or public health, were considered out of scope because they do not produce routine, non-research, textual content in health care scenarios (Step 2). Examples are "aerospace medicine", and immunochemistry (a subspecialty of allergology and immunology). Another category considered as non-specialties, such as "clinical medicine" - an umbrella term for all clinical specialties, but in fact it is orthogonal within the taxonomy, including "precision medicine" and "evidence-based-medicine" - none of them being clinical specialties that would produce a distinct kind of textual data and therefore be relevant for specific vocabulary generation. So, this branch was excluded as well.

Subcategory selection was the only step done manually, given the resources available, the inherent complexity of subdividing clinical specialties, the idiosyncratic nature of the division of the medical realm and the apparent lack of principled modelling of the specialty subtree in MeSH.

Out of this specialty selection we generated, the queries for retrieving specialty-specific PubMed content (Step 3), under the assumption that a text belongs to a clinical specialty S and is, therefore, relevant for term extraction if it fulfils one of the following conditions:

- The affiliation of the first author corresponds to an institution or department associated with S.

- The article was written for a publication in the area of S.
- The article was categorized using keywords relevant to S.
- The article was indexed with MeSH terms relevant to S.

The rationale behind these criteria is that the articles that contain specialty-relevant terms are rarely ever indexed with a MeSH term from the clinical specialty subtree. E.g., "RF-New Recombinant Vaccine for the Prevention of Herpes Zoster" is an article relevant for dermatology, but it is not annotated with the MeSH term "Dermatology". It can also be shown by the fact that less than 20,000 articles are indexed with the MeSH term "cardiology", opposed to 830,000 ones with the word "heart" in title or abstract. Following our goal, we exclusively analysed Spanish content; a mandatory condition is that the PubMed record contains at least a Spanish title.

Our query generators (Step 3) finally performed PubMed queries using Biopython [40] for extracting specialty-specific corpora. The terms used to query for clinical specialties were extracted from the MeSH hierarchy. First, the MeSH term denoting the specialty itself, which is then expanded by the terms available via "See Also" links in MeSH, suggesting other specialty-relevant MeSH terms, e.g., Figure 3 shows the specialty "Cardiology", where three semantically close MeSH terms are suggested under "See also".

In order to expand our query further, we added for each of these terms (MESH and "See also") Spanish MeSH translations retrieved from the UMLS Metathesaurus, marked as MSHSPA. E.g., the MeSH term "General Practice" is expanded by *Medicina General* for the MESH ID D058006.

Figure 3: Example of MeSH term information.

Despite the possibility to obtain more related terms like synonyms and term variants for describing a specialty, we decided to restrict ourselves to a limited set of cue terms that optimally delineate the scope of a clinical specialty. This strategy should maintain a high precision regarding the goal of our study, i.e. harvesting specific terms. It also allows easy reproduction of the query generation by using the UMLS Metathesaurus only. An example of the automatically generated query that selects dermatology content is the following:

SPA[LA] AND ("dermatology"[TA] OR "dermatology"[TIAB] OR "dermatology"[MH:noexp] OR "dermatology"[SH:noexp] OR "dermatology"[CN] OR "dermatology"[SI] OR "dermatology"[OT] OR "dermatology"[AD] OR "*dermatología*"[TA] OR "*dermatología*"[TIAB] OR "*dermatología*"[CN] OR "*dermatología*"[SI] OR "*dermatología*"[OT] OR "*dermatología*"[AD] OR "skin diseases"[MH])

The following PubMed query demonstrates our interest in retrieving Spanish content from dermatology by selecting the following fields and values in the PubMed record:

- Has Spanish(SPA) as document language (LA).

- Contains "dermatology" or "dermatología" in at least one of the following fields: translated journal title (TA), title or abstract translated to English (TIAB), corporate author (CN), secondary source (SI), as keyword set by the author (OT), in some affiliation (AD), in MeSH Terms (MH:noexp) or MeSH Subheadings (SH:noexp).
- Finally, we include the terms "See Also", in this case "Skin Diseases", in the query to search for them in the MeSH terms [MH][2].

Since many sub-specialties share the language of the specialty from which they are derived, our method proposes an integrative process, in which lower levels are integrated within the higher level in the specialty hierarchy (Step 4).

Harvesting important terms that are highly specific for a specialty requires a considerable amount of text. The number of required texts may vary according to the diversity of terms for each specialty; however, a minimum amount of texts is desirable. For this reason, the process removes all second-level specialties with less than 1000 Spanish titles (Step 5) and then maintains only one specialty when it has duplicates, which means that it belongs to more than one MeSH subtree (Step 6). In this step, we noticed that there was one repeated specialty: gynecology. The MeSH term "Gynecology" belonged to both the MeSH subtree "Reproductive Medicine" and "Specialties, Surgical". The final step 7 eliminates specialties in the first level that have not yet achieved a significant amount of content.

From the final set of titles and abstracts per specialty, those with "case reports" as publication type were retained to be used as evaluation benchmark (i.e. test data set). The reason was that case reports seem to use a language that is closer to clinical narratives. Case reports typically document interesting observations on individual patients, e.g., new syndromes, particular and unusual evolution of diseases (often orphan diseases), complications of common treatments as well as beneficial or adverse effects of common or unusual therapies [41].

2.3 Term extraction

Once we obtained a collection Spanish titles and abstracts per specialty, all texts were then submitted to a process that extracts possibly relevant terms.

Term candidates were word n-grams with n between 1 and 3. E.g., from "*Manejo práctico de inmunosupresores en dermatología*", the unigrams extracted are: *Manejo*, *práctico*, *de*, *inmunosupresores*, *en*, *dermatología*. The bigrams extracted are: *Manejo práctico*, *practico de*, *de inmunosupresores, inmunosupresores en*, *en dermatología*. The trigrams are: *Manejo práctico de*, *practico de inmunosupresores*, *de inmunosupresores en*, *inmunosupresores en dermatología*. We decided not to include n-grams for $n > 3$ due computational cost and expected low frequency of longer n-grams.

The first step was to split each text into sentences, the established sentence boundary being ".", ";", "?". "!". As expected, many n-grams were not terms in a strict sense, such as those beginning or ending with articles and prepositions (e.g. "de", "en", "manejo práctico de", "inmunosupresores en". Term cleansing removes those n-grams that match one of the following rules, which were formulated after the inspection of an n-gram sample:

[2] https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.Tn/

- Unigrams included in a Spanish stopword list, punctuation marks and digits.
- Bigrams that include at least one stopword or punctuation marks.
- Trigrams that include at least two stopwords or punctuation marks or whose last position is a stop word.
- N-grams that include a country, regions or nationality name.
- N-grams containing Roman or Arabic numerals.
- N-grams containing dates or years.

Finally, all n-grams were normalised to singular nouns, e.g. from "*hematomas*" to "*hematoma*".

2.4 Term consolidation strategy

In this last step, we applied a term weighting strategy that characterizes the n-grams that belong to a clinical specialty and then filters n-grams below a threshold, we refer to them as stop n-grams.

The term weighting and filtering strategies consider the following restrictions:

1 TF=1 challenge: The volume of titles and abstracts from Spanish articles, furthermore split by clinical specialty, was just too small for applying the TF-IDF measure [42]. In most cases the term frequency was just one, which is known as the TF=1 challenge [43].

2 Multi-class texts: A text may belong to more than only specialty, e.g., endocrinology and nephrology, which increases the overlapping of sets of specialty-related terms.

3 Multipurpose vocabulary: Most proposals for weighting terms (or characteristics) have a well-defined application, such as text classification [44] or query expansion [45, 46]. Since we want to generate a raw set of terms that can serve several purposes, the characterization must take into account different needs. Some applications may require to have only the most discriminating terms for a specialty, others may require its most representative terms, even if they are also frequent other specialties.

2.5 Term weighting

Taking into account the aforementioned restrictions, the weighting strategy considers the following three measures:

- **Term Global Measure (TGM):** a corpus measure that quantifies the concentration of a term along all specialties.
- **Local Precision Measure (LPM):** a specialty-level measure that represents the specificity of a term for a specialty. Accordingly, high LPM values characterise terms that never or rarely occur in texts belonging to other specialties.
- **Local Relevance Measure (LRM):** It represents the capacity of a term to describe the specialty. Terms with high values are those with high frequency in the specialty, compared to other terms in the specialty, and compared to the frequency in other specialties.

The three measures contribute to addressing the "TF=1 challenge" restriction, because the relevance of a term is not computed by text, but by the specialty. In addition, the ability to distinguish the importance of the same term for different

specialties and for the corpus as a whole deals with the "multi-class texts" restriction.

Similarly, the proposed local measures includes two types of terms: (i) infrequent ones with a high predictive power (e.g., "*celoteioma*" for oncology), and (ii) terms that are not only very frequent in text of the specialty, but also in other texts (e.g., "*cáncer*"or "*quimioterapia*" for oncology). Although unspecific they are relevant for the specialty. The differences between these measures contribute to deal with the "multipurpose vocabulary" restriction (see Section 2.4). The notation used to define these measures is explained in Table 1.

Notation	
t	The term under scrutiny
N	The total number of texts in the corpus
N_{e_i}	The number of texts belonging to the specialty i
N^t	The number of texts that contain the term t
$N^t_{e_i}$	The number of texts belonging to a specialty i that contain the term t
E^t	The number of specialties that contain the term t
fe^t_i	The number of occurrences in texts of the specialty i of the term t
F^t	The number of occurrences of the term t in all texts (Spanish abstracts and titles harvested from PubMed records) of all specialties

Table 1: Method notation.

$$TGM(t) = 1 + \sum_{i=1}^{k} \frac{\frac{fe^t_i}{F^t} \times log\frac{fe^t_i}{F^t}}{log(N)} \tag{1}$$

The global measure defined in Equation 1 is a derivative of the entropy measure that evaluates the level of disorder or unpredictability, given a set of classes and a set of features [47]. A value of 1 would correspond to terms that exclusively occur in text from the specialty, whereas 0 would correspond to terms that are spread equally across texts belonging to all specialties. Whenever a term occurs in more than one specialty, a higher value is assigned to those terms whose difference in distribution among specialties is high (e.g. $fe^t_1 = 20$, $fe^t_2 = 2$, $fe^t_3 = 1$,$fe^t_4 = 3$,$fe^t_5 = 1$) and lower values when the difference is low (e.g. $fe^t_1 = 3$, $fe^t_2 = 5$, $fe^t_3 = 2$,$fe^t_4 = 1$,$fe^t_5 = 0$).

$$LPM(t,i) = \frac{N^t_{e_i}}{N^t} + \frac{N^t_{e_i}}{N_{e_i}} \tag{2}$$

The first part of Equation 2 measures the distribution of the term among the specialties and the second part evaluates the local importance using the negative texts within the specialty, i.e., the texts that do not contain the term.

If the term occurs in the texts belonging to a single specialty only and appears in all texts of this specialty the value of this measure will be the greater. The global part range from 0 to 1 and the local from 0 to $1/(N_{e_i} - N^t_{e_i})$. Higher values are assigned to those terms that are highly specific for the specialty, but can be rare.

$$\text{Given } L = P_{90}\left(\{\forall F^t; t \in e_i\}\right) \qquad LRM(t,i) = \begin{cases} \frac{N^t_{e_i}}{N^t}, & \text{if } F^t \geq L \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Typically, the measures used for assessing the importance of a term in a class penalize frequent terms that belong to different document classes because of their inability to accurately discriminate one specific class. However, potentially important terms may be revealed considering their frequency in conjunction with their probability in the specialty and their probability in the corpus. The last measure defined in Equation 3 evaluates this importance. L corresponds to the value of the 90% percentile of the global frequency value of terms that belong to the specialty i.

2.5.1 Filtering of stop n-grams

The last step filters the stop n-grams using the $LRM(t, i)$ measure and the algorithm presented in 1.

Listing 1: Filtering algorithm for determining stop n-grams.

```
def stop_ngram_filtering(frequentNgrams):
    stopNgram = []
    for ngram in frequentNgrams:
        s = std(ngram.LRM)
        m = max(ngram.LRM)
        thresholdSD = min(quantile(ngram.LRM) , .75), 0.25)
        thresholdMAX= max(quantile(ngram.LRM) , .75), 0.1)
        if (s < thresholdSD and m < thresholdMAX )
            stopNgram.append(ngram)
    frequentNgrams = frequentNgrams - stopNgram
    return stopNgram
```

The hypothesis behind the algorithm is that an n-gram is irrelevant if its importance in all specialties where it appears is similar and if there is no specialty where its importance is considerably high. The input of the algorithm is the collection of frequent n-grams, i.e., n-grams appearing in more than 47 specialties. Using this collection, the algorithm detects the stop-n-gram candidates by evaluating the standard deviation of its LRM and the maximum value of its LRM across all specialties where the N-gram appears. When an n-gram has a standard deviation lower than a $thresholdSD$ and a maximum value lower than a $thresholdMAX$, it is considered a stop n-gram.

3 Results

The proposed method was applied to create a set of Spanish terms called SCOVACLIS. For the validation of our initial hypothesis, this resource was used to improve a text classification problem.

The following sections present the analysis of the method and the results of a text classifier using SCOVACLIS. The text classification experiments were made using the complete SCOVACLIS resource and a reduced version containing only the terms that are also found in SNOMED CT.

Figure 4: Terms per specialty and number of Spanish PubMed titles and abstracts. The value inside the point indicates how many standard deviations a specialty is away from the mean (a.k.a. z-score). The average number of tokens in the titles is 13.3, in the abstracts is 249.72.

3.1 Extraction of indicative terms for SCOVACLIS

3.1.1 Source acquisition

The first phase consisted in the acquisition of a corpus, consisting of Spanish titles and abstracts belonging to 51 clinical specialties. Figure 4 describes the number of available titles and abstracts for each specialty.

3.1.2 Term consolidation

The consolidation step yielded 635,699 n-grams. The most frequent ones across the specialties are presented in Table 2. Considering their provenance, these terms give an impression of the dominating research areas in the Spanish-speaking community, under the assumption that the rate of publishing in Spanish does not differ between communities.

N-gram	Total frequency	Specialties in which the n-gram mainly occurs
Cáncer	21,545	medical oncology, preventive medicine, geriatrics, pathology, general surgery
Riesgo	15,796	preventive medicine, epidemiology, cardiology, geriatrics, general surgery
Salud	14,608	preventive medicine, epidemiology, community psychiatry, geriatrics, family practice
Renal	13,884	urology, nephrology, preventive medicine, general surgery, geriatrics
Evaluación	10,655	preventive medicine, geriatrics, epidemiology, general surgery, cardiology
Población	9,592	preventive medicine, epidemiology, geriatrics, cardiology, endocrinology
Tumor	8,598	medical oncology, pathology, preventive medicine, geriatrics, general surgery
Virus	7,997	preventive medicine, epidemiology, venereology, immunochemistry, medical oncology
Carcinoma	7,973	medical oncology, pathology, geriatrics, preventive medicine, general surgery
Atención primaria	7,526	family practice, preventive medicine, geriatrics, epidemiology, cardiology
Factor de riesgo	6,747	preventive medicine, epidemiology, cardiology, geriatrics, endocrinology
Mortalidad	6,572	preventive medicine, epidemiology, geriatrics, neonatology, cardiology
Arterial	6,425	cardiology, preventive medicine, geriatrics, epidemiology, nephrology
Programa	6,216	preventive medicine, community psychiatry, epidemiology, geriatrics, family practice
Trasplante	6,103	general surgery, preventive medicine, urology, nephrology, thoracic surgery
Pronóstico	6,040	preventive medicine, geriatrics, medical oncology, cardiology, pathology
Insuficiencia	5,995	cardiology, preventive medicine, nephrology, urology, geriatrics

Table 2: Most frequent n-grams and clinical specialties in which they appears. Riesgo: risk, población: population, atención primaria: primary care.

As expected, some specialties overlap. Figure 5 visualizes the similarity between clinical specialties. That the three most similar pairs of specialties are obstetrics – perinatology, nephrology – urology and preventive medicine – epidemiology is evident from the closeness and partial overlapping of these disciplines.

Figure 5: Similarity between specialties according to their n-grams.

3.2 Medical text classification enriched using SCOVACLIS

In order to test the hypothesis that NLP tasks benefit from more specific terminologies, such as SCOVACLIS, we performed an assessment of how much a text classification task can be improved by using the new vocabulary.

We propose a multi-label classification. In this classification, the goal was to classify case reports, extracted from PubMed, by one or more clinical specialties to

which they belong. To create the gold standard, a PubMed query with publication type "Case report" and language "Spanish" retrieved 54,881 PubMed records with Spanish titles, of which 714 also had a Spanish abstract. Our classification uses both titles and abstracts, counting on average 11.28 and 163.23 tokens, respectively.

We used 75% of the case reports for training (41,159 texts) and the rest for blind testing (13,720 texts). Our available repository contains more information about the distribution of labels for each partition (train and test) created for this evaluation[3]. For the primary evaluation standard metrics from the ML and NLP community were used: micro-averaged precision (P), recall (R), and balanced F1-score (F1).

In addition, taking into account SNOMED CT as a reference terminology in Spanish, we have performed two experiments, (i) using the complete SCOVACLIS and, (ii) filtering SCOVACLIS with terms covered by SNOMED CT.

3.2.1 Multi-label classification

Because a PubMed record can be related to one or more clinical specialties, the experiment involved multiple labels per report (title and abstract). Labels are binary variables that indicate class (i.e., clinical specialty) membership. This scenario implies greater difficulty due to the computational cost of model generation and querying, as well as the presence of unbalanced labels. In ML, the first step towards training a classifier is text pre-processing, where feature extraction and vectorization take place.

The entire implementation was done using Python on a single Tesla-V100 32 GB GPU with 192 GB of RAM. We performed experiments with different classifiers that would allow multi-label classification such as:

- Random Forest [48]: In our experiment the number of trees taken into account is 100, as a function of measuring quality we use entropy.
- K-nearest Neighbors [49]: The number of neighbors used in our experiment is 5, the weight function used in the prediction is uniform (all points in each neighborhood are weighted equally), and the other default parameters.
- Decision Tree [50]: Similar to the parameters used in Random Forest, here we also use entropy as a function to measure the quality of a split and the other default parameters.
- Multilayer Perceptron (MLP) [51]: MLP network consisting of three layers: (i) one input layer, (ii) one hidden layer, and (iii) one output layer. The number of nodes in the hidden layer is 100, use reLU activation, 0.001 in learning rate and use early stopping for controlling overfitting.

We tested different parameters to adjust the classifiers to the problem (for detail cf. our repository[4]). We also performed tests with different word representation vectors such as: TF-IDF, one-hot encoding and bag-of-words. The best results were obtained using TF-IDF with the following parameters: lowercase = True, stopwords = Spanish stop words and ngram range = (1,3).

[3] https://github.com/plubeda/scovaclis/blob/master/Distribution-of-labels.md
[4] https://github.com/plubeda/scovaclis/blob/master/Detailed-classifiers.md

To validate the usefulness of SCOVACLIS, we added features as extra information to each text, using a vector with size 51 according to the number of clinical specialties. The value of each feature is a score calculated following the Equation 4.

$$SCOVACLIS_s(d) = \sum_{i=1}^{n} TGM(t), \forall t \in s, \tag{4}$$

where:

d is the document (in our case a title or an abstract)

n is the number of ngrams

s is the specialty

The results obtained by the different combinations of classifiers and word representation are shown in Table 3.

Classifier	Word Representation	P (%)	R (%)	F1 (%)
Random Forest	TF-IDF	71.7	25.1	38.4
Decision Tree	TF-IDF	47.9	38.1	42.4
KNeighbors	TF-IDF	63.3	39.0	48.2
MLP	TF-IDF	75.1	53.3	59.3
Random Forest	TF-IDF + $SCOVACLIS_s$	70.0	17.5	28.7
Decision Tree	TF-IDF + $SCOVACLIS_s$	46.2	43.5	44.8
KNeighbors	TF-IDF + $SCOVACLIS_s$	69.3	42.6	52.7
MLP	TF-IDF + $SCOVACLIS_s$	74.7	57.4	64.9
Random Forest	$SCOVACLIS_s$	76.0	32.6	45.6
Decision Tree	$SCOVACLIS_s$	42.6	43.1	42.8
KNeighbors	$SCOVACLIS_s$	69.4	42.1	52.4
MLP	$SCOVACLIS_s$	75.8	43.5	55.3
Random Forest	TF-IDF + $SCOVACLIS_s$ - stop ngrams	70.3	18.9	30.6
Decision Tree	TF-IDF + $SCOVACLIS_s$ - stop ngrams	46.4	43.9	45.1
KNeighbors	TF-IDF + $SCOVACLIS_s$ - stop ngrams	68.9	42.7	52.7
MLP	TF-IDF + $SCOVACLIS_s$ - stop ngrams	**77.5**	**57.7**	**65.2**
Random Forest	$SCOVACLIS_s$ - stop ngrams	76.8	32.6	45.8
Decision Tree	$SCOVACLIS_s$ - stop ngrams	43.1	43.1	43.1
KNeighbors	$SCOVACLIS_s$ - stop ngrams	68.9	42.7	52.7
MLP	$SCOVACLIS_s$ - stop ngrams	75.6	43.5	55.4

Table 3: Multi-label classification. Annotated data results with SCOVACLIS Score ($SCOVACLIS_s$) and removing stop-ngrams ($SCOVACLIS_s$ - stop ngrams).

In the KNN, Decision Tree and MLP classifiers the use of the term set as a feature improved the baseline (TF-IDF). The 10% increase using the MLP method stands out, achieving 64.9% using TF-IDF with the created collection (TF-IDF + $SCOVACLIS_s$).

In the second scenario, in which we used only the term set ($SCOVACLIS_s$), we observed a small increase in almost all cases except the MLP classifier.

In the third scenario (TF-IDF + $SCOVACLIS_s$ - stop ngrams) we removed stop-ngrams. We observed the same as in the first scenario (TF-IDF + $SCOVACLIS_s$), so we improved the base case in most cases. However the difference with the first scenario was small, which led us to conclude that the stop-ngrams did not add too

much noise to the classification. Here we achieved the highest precision and recall score, with an F1 measure of almost 65.2% using MLP.

Finally, the scenario in which we removed stop-ngrams ($SCOVACLIS_s$ - stop ngrams) obtained values similar to the second scenario ($SCOVACLIS_s$).

We concluded the study by observing that the use of the term set improved the baseline in both cases, using it alone or including it as a set of features in the classifier (see more details in GitHub[5]).

SCOVACLIS has been generated with the aim of improving vocabulary resources taking into account clinical specialties. Currently, there is no standard or reference terminology that classifies terms per clinical speciality, so no direct comparison can be made. However, to analyze the relationship of our resource with a curated terminology, we perform experiments using SNOMED CT.

The goal is to quantify the level of overlapping between SCOVACLIS and SNOMED CT, and analyze whether the non-overlapped terms in SCOVACLIS are useful to perform a task that requires discrimination per clinical specialty. To this end we have downloaded the international SNOMED CT release in Spanish (2020-04-30 release) and performed the same pre-processing as for our term set (Section 2.3).

For this experiment we have reduced SCOVACLIS for each clinical speciality, taking into account only those terms that were included in SNOMED CT. The average number of terms covered by SNOMED CT in all specialities was 28.16%. E.g., dermatology has been reduced from 46,000 terms to around 12,000 and cardiology from 110,000 to around 20,000 terms. With this resource we repeat the classification experiments and applied them to the test data set.

We analyze the characteristics of the terms that were not covered by SNOMED CT to recognize possible noise terms in our term set (cf. detailed results[6]). We could demonstrate that many SCOVACLIS terms with high E^t values (cf. Section 2.5) were found in SNOMED CT. A high E^t means that the term was found in several specialties. 90% of the terms found between 38 and 52 specialties were included in SNOMED CT, in contrast to only 15% of the terms found between 1 and 13 specialties. This corroborates that the Spanish SNOMED CT gives preference to terms of wide-spread use but lacks coverage of very specific terms only used in one specialty. This seems interesting, because SCOVACLIS could be used for adding more content when an NLP task requires these specific and complementary terms. The manual analysis of the terms yields that some of the missing terms are related terms to formal ones (e.g. “filled breast” instead of “breast implant” in “Surgery plastic”), and others are more specialized than the ones contained in SNOMED CT (e.g. “thoracoabdominal wall” in “Thoracic Specialty” and “kayakista” in “Sports Medicine”). Lastly, we found that specialties with a broad lexical nature, such as preventive medicine and general practice, apparently include terms of minor relevance. However, they are useful for improving classification performance, as illustrated in the classification experiment.

Finally, Table 4 shows the comparison between the classification results with the original term set (*original* columns) and with the reduced term set restricted to

[5] https://github.com/plubeda/scovaclis/blob/master/Detailed-results.md

[6] https://github.com/plubeda/scovaclis/blob/master/SNOMED-CT-analysis.md

		Original			SNOMED CT filter		
Classifier	Word Representation	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Random Forest	TF-IDF + $SCOVACLIS_s$	70.0	17.5	28.7	69.1	14.0	22.6
Decision Tree	TF-IDF + $SCOVACLIS_s$	46.2	43.5	44.8	37.3	34.0	35.8
KNeighbors	TF-IDF + $SCOVACLIS_s$	69.3	42.6	52.7	62.8	31.1	41.6
MLP	TF-IDF + $SCOVACLIS_s$	74.7	57.4	64.9	70.1	56.1	60.5
Random Forest	$SCOVACLIS_s$	76.0	32.6	45.6	65.4	14.0	23.1
Decision Tree	$SCOVACLIS_s$	42.6	43.1	42.8	31.5	26.8	28.9
KNeighbors	$SCOVACLIS_s$	69.4	42.1	52.4	57.9	29.0	38.7
MLP	$SCOVACLIS_s$	75.8	43.5	55.3	73.4	24.3	36.6

Table 4: Multi-label classification. Annotated data results with SCOVACLIS Score ($SCOVACLIS_s$) and filtered terms with SNOMED CT.

terms occurring in SNOMED CT (*SNOMED CT filter* columns). The latter, reduced termset does not improve the original classification. Particularly, when taking into account the F1 measure, we obtained a reduction between 4 and 18 points. This means that the terms included in SCOVACLIS improve the recognition of the specialties, even though they are not canonical or curated terms.

3.2.2 Label distribution analysis

Considering that the datasets used for training and testing were not balanced regarding the specialty, this section analyzes the classification error versus the distribution of the class by dividing the F1 results into ranges: 0 to 25%, 25% to 50%, 50% to 75% and 75% and above. 18 specialties with an average of 356 texts for training, obtained an F1 value of less than 25%. 9 specialties (including, e.g. geriatrics and psychiatry), obtained an F1 between 25 and 50%, with an average of 2,049 texts for training and 688 for testing. Specialties such as pathology and obstetrics and 13 others had an F1 between 50 and 75%, with on average, 2,635 training texts. Finally, the specialties that reported the best results (i.e., over 75%) were 9, with oncology and cardiology among them, having on average, 5,915 titles/abstracts to train the system. In the first range (i.e., 0 to 25%), there were 6 specialties with F1 = 0 measure; they contained less than 30 training titles/abstracts and less than 7 titles/abstracts for testing, among them were forensic medicine, general practice and allergy, and immunology.

This analysis allows the conclusion that, as expected, the more texts the system has for learning, the better it classifies. Our multi-label approach, makes the task more problematic as demonstrated by the following misclassifications examples:

```
    Example #1.
Text: Infiltración pleural en la recaída de un mieloma múltiple.
Translation: Pleural infiltration in multiple myeloma relapse.
True specialties: cardiology, hematology, medical oncology,
    pathology and preventive medicine.
Predicted specialties: cardiology, hematology, medical oncology and
    pathology.

    Example #2.
Text: Osteoartropatía hipertrófica en adenocarcinoma de pulmón.
Translation: Hypertrophic osteoarthropathy in lung adenocarcinoma.
True specialties: medical oncology, pulmonary medicine and rheumatology.
```

```
Predicted specialties: medical oncology.

    Example #3.
Text: Derivación gástrica en Y de Roux como procedimiento de urgencia
para resolver la fuga en un SADI-S.
Translation: Roux-en-Y gastric bypass as an emergency procedure for
resolving SADI-S leak.
True specialties: bariatric medicine and general surgery.
Predicted specialties: bariatric medicine, general surgery and
gastroenterology.
```

We highlight the difficulty that our classifier has to detect specialties considered transversal (i.e., that do not have a very specific vocabulary) such as pathology, internal medicine or general surgery (cf. detailed results[7]).

4 Discussion

Term identification is crucial for the automated processing of biomedical texts, with CVs are fundamental resources for text classification. We have proposed and validated a method not only to harvest terminology from texts but to classify them into clinical specialties.

The application of this method to a Spanish Core Vocabulary About Clinical Specialties (SCOVACLIS) is the first accomplishment of this research. We can recommend this method for obtaining specialized clinical term sets in any language sufficiently represented in PubMed titles and abstracts. It proved useful for recognizing both frequent, infrequent, and equally relevant terms characteristic for given clinical specialties. The completeness of the term sets by our method is directly related to the volume and richness of the texts obtained for each specialty. Broad-ranging specialties contained less specific terms, such as general practice or preventive medicine. Therefore, they obtained lower results when classifying texts.

Our goal is not to replace existing controlled terminologies, but to support the production of specialized term sets tailored to tasks that require terms with high predictive value, as they occur in clinical or research texts, regardless of naming conventions used for the building of CVs. In contradistinction to these intensively curated resources, the emphasis of our method is on producing term sets in a fully automatic manner, which may include non-standard terms, e.g. with frequent spelling variants. This, however, does not exclude at all the potential of our method to provide useful input to terminology developers who maintain and extend reference terminologies or interface terminologies [25]. For the purpose optimising such resources to clinical documentation, clinical texts would be preferable as input.

A resource named SCOVACLIS is the second accomplishment of our research. It allows us to enrich the set of resources available for NLP in Spanish. Creation and validation of SCOVACLIS support the hypothesis that a term set classified by clinical specialty might reduce the level of ambiguity when compared to a specialty-independent, broad-scope vocabulary. Disambiguation, particularly of short forms, is a known bottleneck in clinical text processing.

Regarding the validation of SCOVACLIS, its use for improving the features in a multi-class classification approach using a Multilayer Perceptron classifier achieved

[7] https://github.com/plubeda/scovaclis/blob/master/Binary-confusion-matrix.md

an increase in 6 percentage points in the F1-measure compared to the baseline. This shows its usefulness to improve contextual knowledge about medical texts and thus better solve NLP problems such as named entity recognition and classification. Also, SCOVACLIS proved to increase the performance of text classification compared to the use of curated terms such as the ones included in SNOMED CT.

Implementing this method and producing the language resource was not straightforward; there were several challenges, mainly caused by not having a preexisting corpus of clinical texts in Spanish. We, therefore, needed to find criteria to decide which clinical specialties to use; our solution, mainly based on the MeSH occupation hierarchy enriched by other features extracted from PubMed metadata, is more complex than it would have been with clinical texts, whose provenance (clinical department in the document header) would have been trivial to ascertain.

Most clinical document collections that are publicly accessible, e.g. MIMIC III, are in English, whereas no sufficient amount of publicly available Spanish clinical text could be obtained. In contradistinction, an easily accessible source of biomedical texts is the literature database MEDLINE, with PubMed as a free search engine. From more than 26 million records approx. 2.2 million are about non-English publications, including about 330,000 Spanish entries for which a Spanish title and sometimes a Spanish abstract is available. By using case studies for evaluation, we have extracted a publication genre that is supposed to be closer to clinical language than other, purely scientific papers.

Thanks to the existence of freely available query interfaces to PubMed and MeSH, the process of obtaining relevant texts for each specialty was executed. The fact that the universe of publications linked to a medical specialty in MEDLINE is much larger than the set of articles indexed by an occupation-specific MeSH term, led us to enrich the search method by incorporating new elements that allow us to increase the recall of the query.

In contrast to related works [34], the extraction of texts as input for term extraction is done automatically, including classification by specialty. Our manual effort was limited to the crafting of the search strategy using MeSH terms and text words in the authors' affiliations. Therefore, our method can be applied in any language that has available scientific publications in PubMed. [35, 37]. Likewise, it can be used to create term sets for different domains, depending on the initial descriptors [38, 39].

An important limitation of our study is the restriction of the language to scientific language, which is known to be much different compared to the jargon found in clinical documents, which are known to hastily written and marred with typos and cryptic short forms. To what extent SCOVACLIS is a useful resource for handling clinical documents still has to be investigated. For future work, we plan to expand our term set and use comprehensive reports.

Even though we used the Spanish titles and abstracts to create the term set, we found a limitation because only 4,592 PubMed records were linked to an automatically accessible abstract in Spanish. Thus, most of the texts under scrutiny consisted of titles only. The use of Spanish full texts could be a solution, with about 64,000 Spanish full text sources being freely accessible from PubMed. This, however, would require considerable manual effort.

For future work, we plan to explore word embeddings and train some to use them in traditional ML [52] or neural network approaches. In addition, there are available pre-trained models for the biomedical domain, such as BioBERT [53], we could consider. Although BioBERT is in English, an ideal scenario would be the generation of a new model for Spanish trained by a large biomedical corpus. It is also important to evaluate the value of using specific terminologies in NLP tasks involving specialties with a broad lexical nature, such as preventive medicine and general practice.

5 Conclusions

The following objectives were pursued with this study: first, to elaborate a method for extracting non-English content from PubMed records; second, to tag these extracts by clinical specialty; and third, to generate characteristic term lists for each clinical specialty.

The method for the automatic extraction of medical terms involves the following steps:

1. Selection of clinical specialties to be evaluated.
2. Generation of a PubMed extract (titles and abstracts) labeled by clinical specialties.
3. Extraction of word n-grams.
4. Normalization of word n-grams.
5. Generation of metrics that support the selections of the relevant terms in each clinical specialty.
6. Identification of stop n-grams.

The resource obtained by applying this method to Spanish titles and abstracts, named SCOVACLIS, was evaluated in a multi-label classification task. The results have shown that our resource improved the baseline (without SCOVACLIS). We obtained an F-measure of 65.2% using a MLP network, 77.5% in precision and 57.7% in recall using TFIDF representation, and SCOVACLIS without stop n-grams as features.

Acknowledgements
Not applicable.

Funding
This work has been partially supported by the LIVING-LANG project (RTI2018-094653-B-C21) of the Spanish Government, the Fondo Europeo de Desarrollo Regional (FEDER) and Alianza CAOBA (671-2019) Colombia.

Abbreviations
Not applicable.

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the GitHub repository, https://github.com/plubeda/scovaclis

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions

P.L.U: Software, Formal analysis, Resources, Data Curation, Writing - Original Draft, Visualization. A.P.Q: Conceptualization, Validation, Writing - Review and Editing, Supervision, Project administration, Funding acquisition. M.C.D.G: Methodology, Software, Formal analysis, Investigation, Writing - Original Draft. S.S: Conceptualization, Validation, Writing - Review and Editing, Supervision, Project administration.

Author details

[1] Universidad de Jaén, Campus Las Lagunillas, s/n, 23071, Jaén, Spain. [2] Medical University of Graz, Auenbruggerpl No 2, 8036, Graz, Austria. [3] Pontificia Universidad Javeriana, Cra. 7 No 40-62, 110231 Bogotá, Colombia.

References

1. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., Liu, H.: Clinical information extraction applications: A literature review. Journal of Biomedical Informatics **77**, (2017)
2. Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., F Jones, S., Forshee, R., Walderhaug, M., Botsis, T.: Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. Journal of Biomedical Informatics **73** (2017). doi:10.1016/j.jbi.2017.07.012
3. Hahn, U., Oleynik, M.: Medical information extraction in the age of deep learning. Yearbook of Medical Informatics **29**(1), 208 (2020)
4. López-Úbeda, P., Díaz-Galiano, M.C., Montejo-Ráez, A., Martín-Valdivia, M.-T., Ureña-López, L.A.: An integrated approach to biomedical term identification systems. Applied Sciences **10**(5), 1726 (2020)
5. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., Liu, H.: Clinical information extraction applications: A literature review. Journal of Biomedical Informatics **77**, 34–49 (2018)
6. Miotto, R., Weng, C.: Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. Journal of the American Medical Informatics Association : JAMIA **22** (2015). doi:10.1093/jamia/ocu050
7. Ford, E., A Carroll, J., Smith, H., Scott, D., Cassell, J.: Extracting information from the text of electronic medical records to improve case detection: A systematic review. Journal of the American Medical Informatics Association **23**, 180 (2016). doi:10.1093/jamia/ocv180
8. Sheikhalishahi, S., Miotto, R., T Dudley, J., Lavelli, A., Rinaldi, F., Osmani, V.: Natural language processing of clinical notes on chronic diseases: Systematic review. JMIR medical informatics **7**, 12239 (2019). doi:10.2196/12239
9. A Koleck, T., Dreisbach, C., E Bourne, P., Bakken, S.: Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. Journal of the American Medical Informatics Association : JAMIA **26** (2019)
10. Dreisbach, C., A. Koleck, T., E. Bourne, P., Bakken, S.: A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. International Journal of Medical Informatics **125** (2019). doi:10.1016/j.ijmedinf.2019.02.008
11. Luo, Y., K Thompson, W., M Herr, T., Zeng, Z., A Berendsen, M., Jonnalagadda, S., Carson, M., Starren, J.: Natural language processing for ehr-based pharmacovigilance: A structured review. Drug safety **40** (2017). doi:10.1007/s40264-017-0558-6
12. Feng, C., Le, D., McCoy, A.: Using electronic health records to identify adverse drug events in ambulatory care: A systematic review. Applied Clinical Informatics **10**, 123–128 (2019). doi:10.1055/s-0039-1677738
13. Úbeda, P.L., Galiano, M.C.D., Lopez, L.A.U., Martín-Valdivia, M.T.: Using snomed to recognize and index chemical and drug mentions. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, pp. 115–120 (2019)
14. Imler, T., Morea, J., Kahi, C., Cardwell, J., S Johnson, C., Xu, H., Ahnen, D., Antaki, F., Ashley, C., Baffy, G., Cho, I., Dominitz, J., Hou, J., Korsten, M., Nagar, A., Promrat, K., Robertson, D., Saini, S., Shergill, A., Imperiale, T.: Multi-center colonoscopy quality measurement utilizing natural language processing. The American journal of gastroenterology **110** (2015). doi:10.1038/ajg.2015.51
15. Hsu, W., Han, S.X., Arnold, C.W., Bui, A.A.T., Enzmann, D.R.: A data-driven approach for quality assessment of radiologic interpretations. **23**(e1), 152–156 (2016)
16. Leyh-Bannurah, S.-R., Tian, Z., Karakiewicz, P., Wolffgang, U., Sauter, G., Fisch, M., Pehrke, D., Huland, H., Graefen, M., Budäus, L.: Deep learning for natural language processing in urology: State-of-the-art automated extraction of detailed pathologic prostate cancer data from narratively written electronic health records. JCO Clinical Cancer Informatics, 1–9 (2018). doi:10.1200/CCI.18.00080
17. López-Úbedaa, P., Díaz-Galianoa, M.C., Martín-Valdiviaa, M.T., Ureña-Lópeza, L.A.: Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings. Proceedings of IberLEF 2020 (2020)
18. National Library of Medicine: Medical Subject Headings - MeSH. http://www.nlm.nih.gov/mesh/ (2019)
19. Organization, W.H.: International Statistical Classification of Diseases and Related Health Problems - ICD-10. https://icd.who.int/browse10/2010/en (2019)
20. SNOMED International: SNOMED. http://www.snomed.org/ (2019)
21. Schulz, S., Daumke, P., Romacker, M., López-García, P.: Representing oncology in datasets: Standard or custom biomedical terminology? Informatics in Medicine Unlocked **15**, 100186 (2019)
22. Mowery, D.L., Wiebe, J., Visweswaran, S., Harkema, H., Chapman, W.W.: Building an automated SOAP classifier for emergency department reports. Journal of Biomedical Informatics **45**(1), 71–81 (2012)
23. National Library of Medicine: UMLS - Unified Medical Language System. http://uts.nlm.nih.gov (2019)
24. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), 267–270 (2004)

25. Schulz, S., Rodrigues, J.M., Rector, A., Chute, C.G.: Interface terminologies, reference terminologies and aggregation terminologies: A strategy for better integration. Studies in health technology and informatics **245**, 940–944 (2017)
26. Kim, Y., Riloff, E., Hurdle, J.F.: A Study of Concept Extraction Across Different Types of Clinical Notes. AMIA Annual Symposium Proceedings **2015**, 737–746 (2015). Accessed 2019-06-14
27. Jindal, P., Roth, D.: Using domain knowledge and domain-inspired discourse model for coreference resolution for clinical narratives. Journal of the American Medical Informatics Association : JAMIA **20**(2), 356–362 (2013). doi:10.1136/amiajnl-2011-000767. Accessed 2019-06-14
28. Patterson, O., Hurdle, J.F.: Document clustering of clinical narratives: a systematic study of clinical sublanguages. AMIA ... Annual Symposium proceedings. AMIA Symposium **2011**, 1099–1107 (2011)
29. Doing-Harris, K., Patterson, O., Igo, S., Hurdle, J.: Document Sublanguage Clustering to Detect Medical Specialty in Cross-institutional Clinical Texts. Proceedings of the ACM ... International Workshop on Data and Text Mining in Biomedical Informatics. ACM International Workshop on Data and Text Mining in Biomedical Informatics **2013**, 9–12 (2013). doi:10.1145/2512089.2512101
30. Bernhardt, P.J., Humphrey, S.M., Rindflesch, T.C.: Determining prominent subdomains in medicine. AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium **2005**, 46–50 (2005). American Medical Informatics Association
31. Zhang, R., Pakhomov, S.V.S., Arsoniadis, E.G., Lee, J.T., Wang, Y., Melton, G.B.: Detecting clinically relevant new information in clinical notes across specialties and settings. BMC Medical Informatics and Decision Making **17**(2), 68 (2017). doi:10.1186/s12911-017-0464-y
32. Weng, W.-H., Wagholikar, K.B., McCray, A.T., Szolovits, P., Chueh, H.C.: Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Medical Informatics and Decision Making **17**(1), 155 (2017). doi:10.1186/s12911-017-0556-8
33. Krauthammer, M., Nenadic, G.: Term identification in the biomedical literature. Journal of biomedical informatics **37**(6), 512–526 (2004)
34. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics **17**(01), 128–144 (2008)
35. Takeuchi, K., Collier, N.: Bio-medical entity extraction using support vector machines. Artificial Intelligence in Medicine **33**(2), 125–137 (2005)
36. Ayadi, M.G., Bouslimi, R., Akaichi, J.: A model for multilingual terminology extraction via a medical social network. Procedia Computer Science **112**, 21–30 (2017)
37. Komenda, M., Karolyi, M., Pokorná, A., Víta, M., Kríž, V.: Automatic keyword extraction from medical and healthcare curriculum. In: 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 287–290 (2016). IEEE
38. Marciniak, M., Mykowiecka, A.: Terminology extraction from medical texts in polish. Journal of biomedical semantics **5**(1), 24 (2014)
39. Sandoval, A.M., Díaz, J., Llanos, L.C., Redondo, T.: Biomedical term extraction: Nlp techniques in computational medicine. IJIMAI **5**(4), 51–59 (2019)
40. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L.: Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11), 1422–1423 (2009)
41. Gagnier, J.J., Kienle, G., Altman, D.G., Moher, D., Sox, H., Riley, D.: The care guidelines: consensus-based clinical case reporting guideline development. Journal of medical case reports **7**(1), 223 (2013)
42. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA (1986)
43. Timonen, M.: Categorization of very short documents. In: KDIR 2012 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, pp. 5–16. SCITEPRESS, ??? (2012)
44. Yi, K., Beheshti, J.: A hidden markov model-based text classification of medical documents. Journal of Information Science **35**(1), 67–81 (2009)
45. Aronson, A.R., Rindflesch, T.C.: Query expansion using the umls metathesaurus. In: Proceedings of the AMIA Annual Fall Symposium, p. 485 (1997). American Medical Informatics Association
46. Díaz-Galiano, M.C., Martín-Valdivia, M.T., Ureña-López, L.: Query expansion with a medical ontology to improve a multimodal information retrieval system. Computers in biology and medicine **39**(4), 396–403 (2009)
47. Shannon, C.E.: Prediction and entropy of printed english. The Bell System Technical Journal **30**(1), 50–64 (1951). doi:10.1002/j.1538-7305.1951.tb01366.x
48. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
49. Yang, Y.: An evaluation of statistical approaches to text categorization. Information retrieval **1**(1-2), 69–90 (1999)
50. Song, Y.-Y., Ying, L.: Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry **27**(2), 130 (2015)
51. Yan, H., Jiang, Y., Zheng, J., Peng, C., Li, Q.: A multilayer perceptron-based medical decision support system for heart disease diagnosis. Expert Systems with Applications **30**(2), 272–281 (2006)
52. Ye, C., Fabbri, D.: Extracting similar terms from multiple emr-based semantic embeddings to support chart reviews. Journal of biomedical informatics **83**, 63–72 (2018)
53. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

Figures

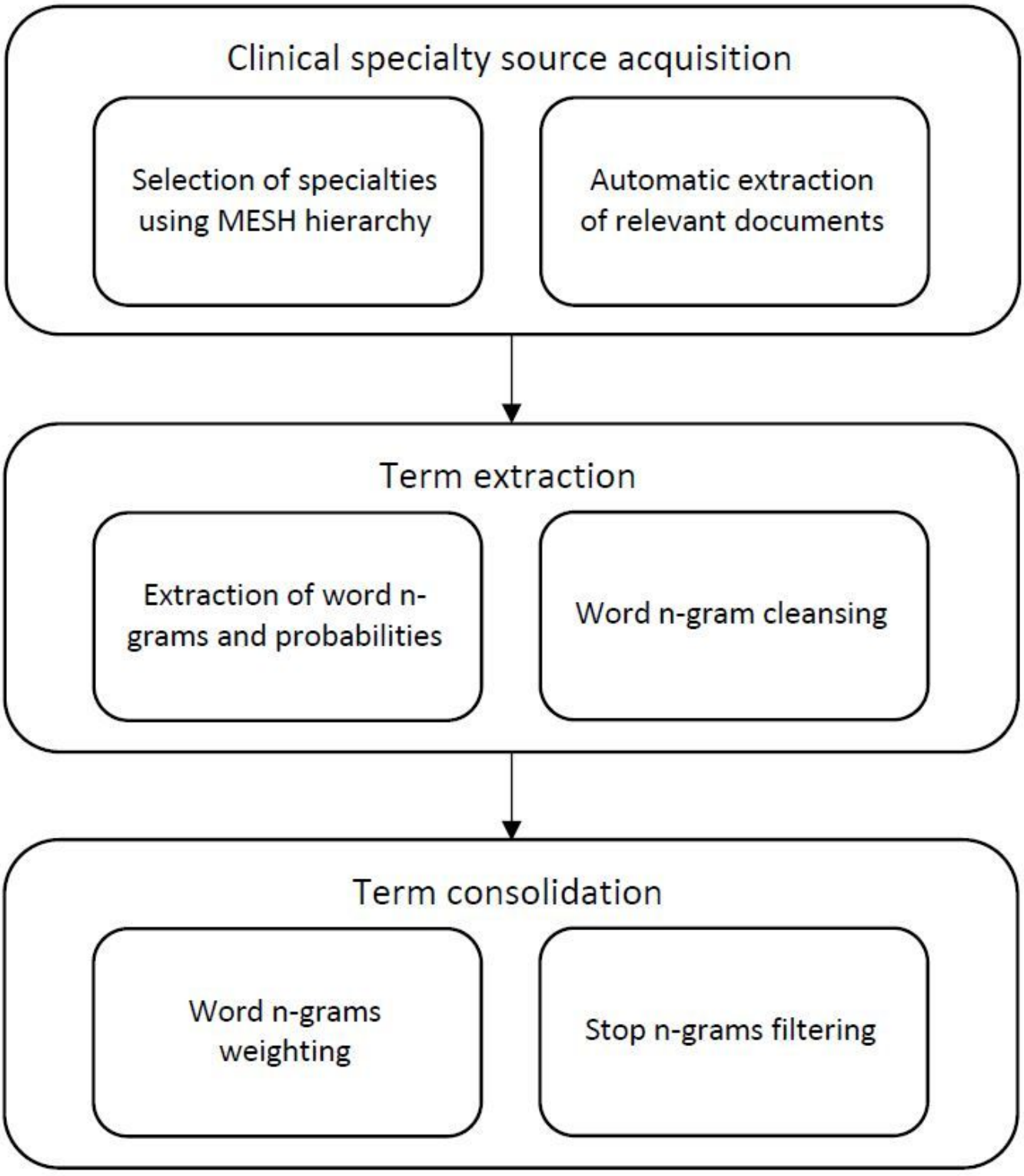

Figure 1

Overview of the extraction method.

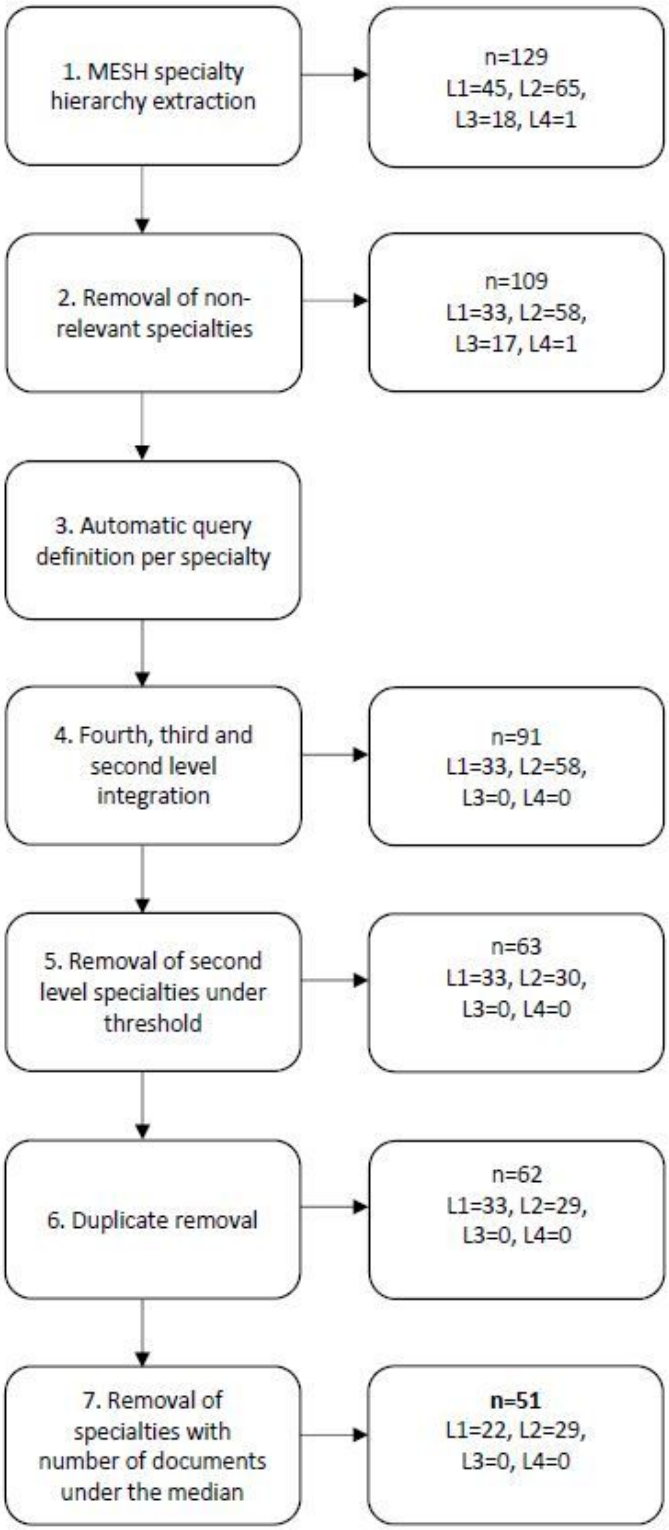

Figure 2

Clinical specialty selection process. Including the variations in the number of specialties (left) when applied to the Spanish case (right)

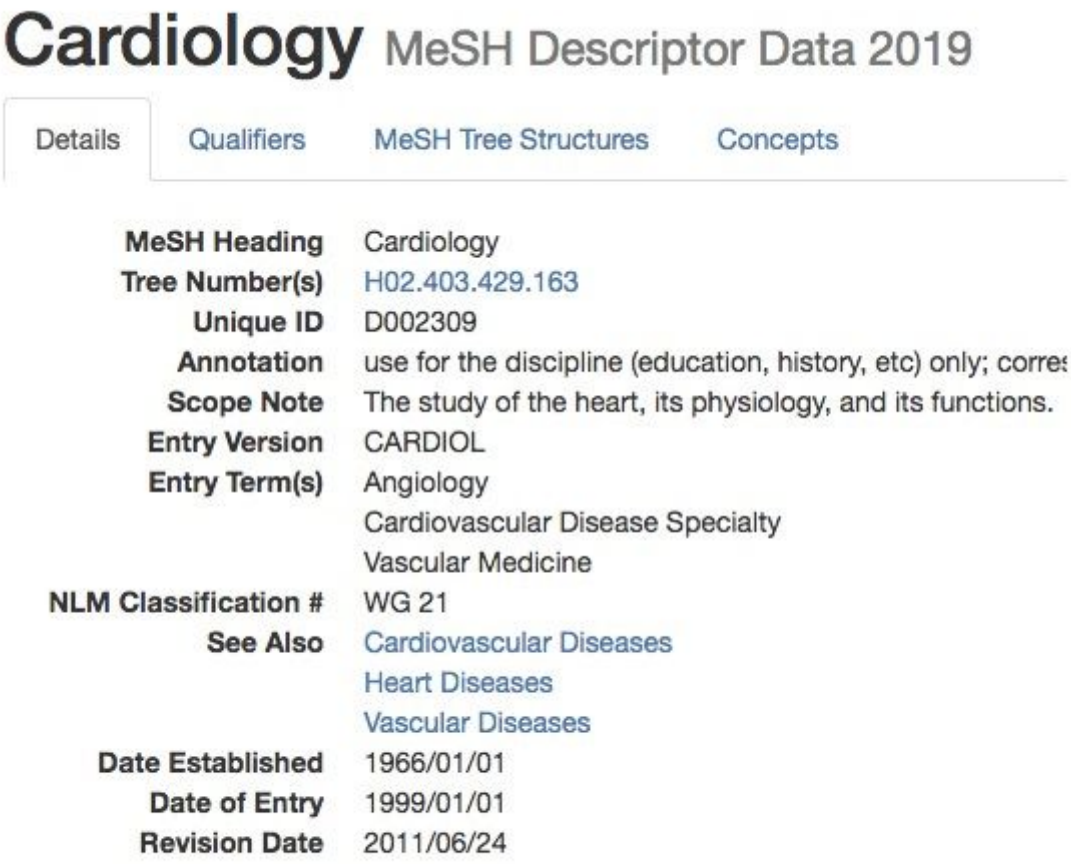

Figure 3

Example of MeSH term information.

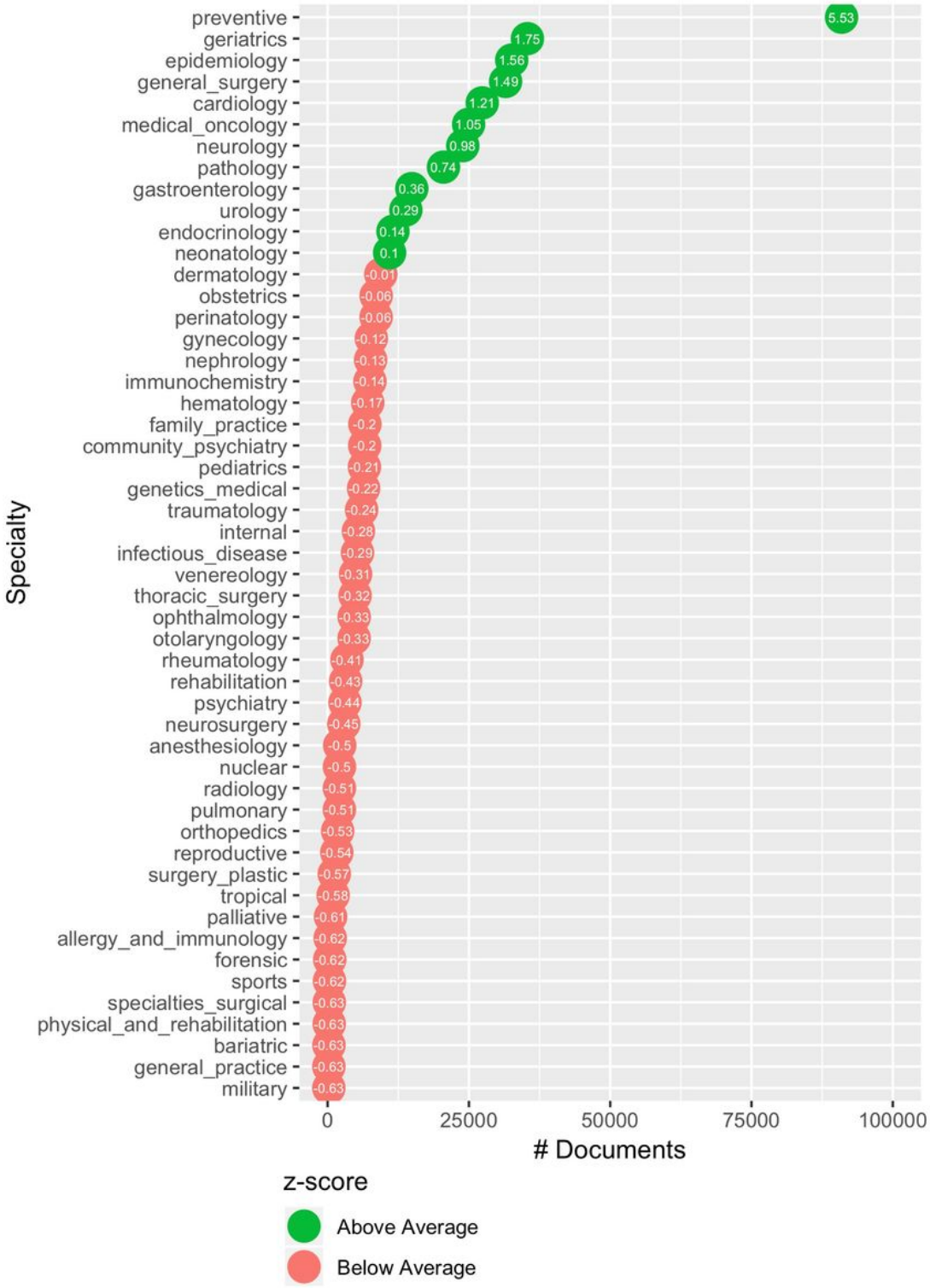

Figure 4

Terms per specialty and number of Spanish PubMed titles and ab- stracts. The value inside the point indicates how many standard deviations a specialty is away from the mean (a.k.a. z-score). The average number of tokens in the titles is 13.3, in the abstracts is 249.72.

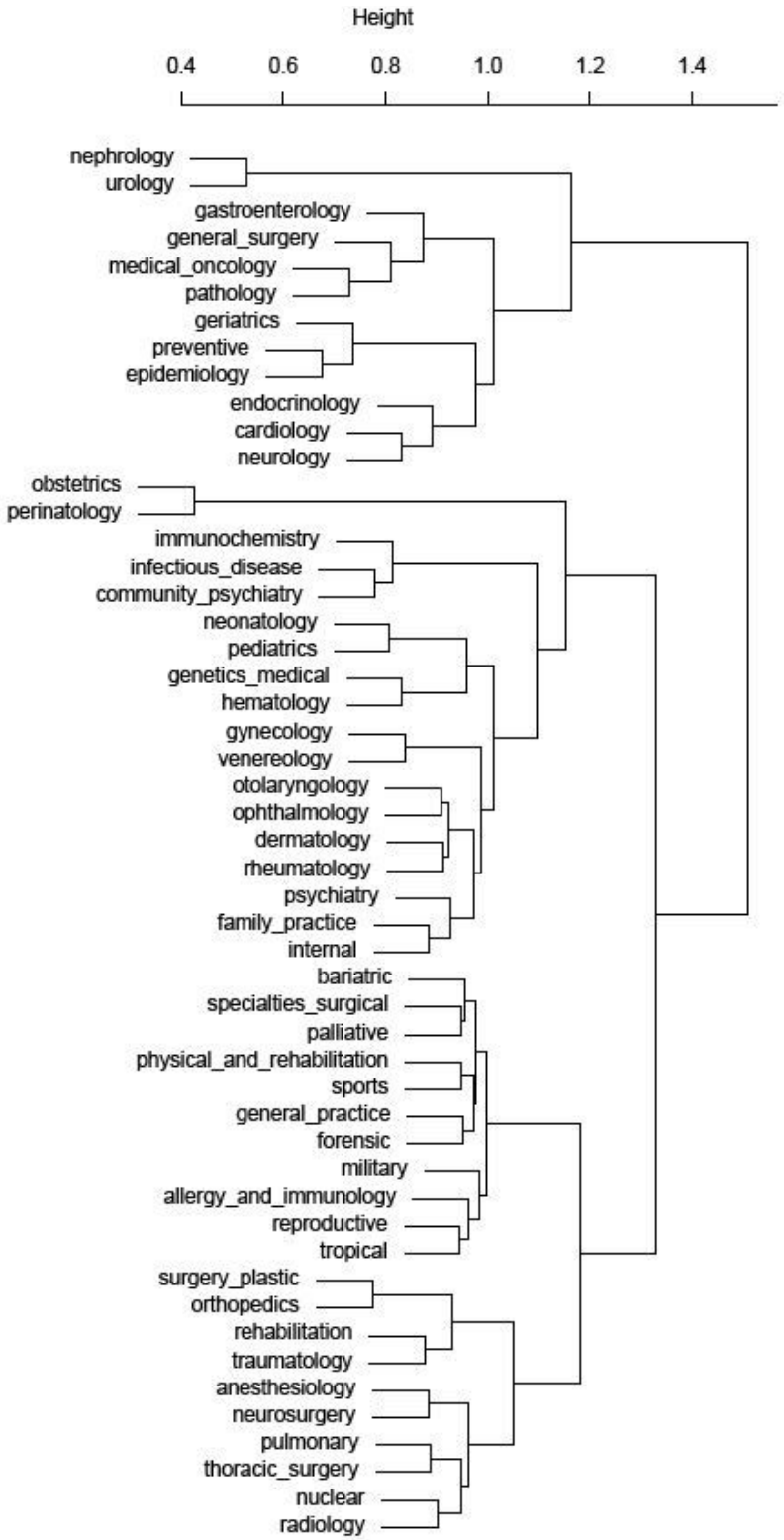

Figure 5

Similarity between specialties according to their n-grams.