Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09755-7

ORIGINAL PAPER

Entity normalization in a Spanish medical corpus using a UMLS-based lexicon: findings and limitations

Pablo Báez[1,2] · Leonardo Campillos-Llanos[3] · Fredy Núñez[4,5] · Jocelyn Dunstan[5,6,7]

Accepted: 7 June 2024

Abstract

Entity normalization is a common strategy to resolve ambiguities by mapping all the synonym mentions to a single concept identifier in standard terminology. Normalizing medical entities is challenging, especially for languages other than English, where lexical variation is considerably under-represented. Here, we report a new linguistic resource for medical entity normalization in Spanish. We applied a UMLS-based medical lexicon (MedLexSp) to automatically normalize mentions from 2000 medical referrals of the Chilean Waiting List Corpus. Three medical students manually revised the automatic normalization. The inter-coder agreement was computed, and the distribution of concepts, errors, and linguistic sources of variation was analyzed. The automatic method normalized 52% of the mentions, compared to 91% after manual revision. The lowest agreement between automatic and automatic-manual normalization was observed for Finding, Disease, and Procedure entities. Errors in normalization were associated with ortho-typographic, semantic, and grammatical linguistic inadequacies, mainly of the hyponymy/hyperonymy, polysemy/metonymy, and acronym-abbreviation types. This new resource can enrich dictionaries and lexicons with new mentions to improve the functioning of modern entity normalization methods. The linguistic analysis offers insight into the sources of lexical variety in the Spanish clinical environment related to error generation using lexicon-based normalization methods. This article also introduces a workflow that can serve as a benchmark for comparison in studies replicating our analysis in Romance languages.

Keywords Clinical text · Entity linking · Lexical variation · Linguistic resources · Medical lexicon · Normalization

Extended author information available on the last page of the article

Published online: 02 July 2024

1 Introduction

The landscape of modern healthcare delivery and medical research is markedly shaped by the digitization of patient records, particularly through the advent of electronic health records (EHRs). However, the predominantly free-text format of EHRs presents significant challenges for information retrieval and exchange (Dalianis, 2018). One of the main difficulties is the inherent lexical variability in clinical language, where multiple mentions may refer to the same medical concept (Campillos-Llanos et al., 2016; Newman-Griffis et al., 2021). This variability stems from morphological aspects (different forms of a word by inflection or derivation) and ortho-typographic factors (McCray et al., 1994), leading to a plethora of synonyms, non-standard abbreviations, misspellings, and regional preferences (Leaman et al., 2015; Dziadek et al., 2017).

Additionally, lexical ambiguity compounds the complexity, as homograph words or mentions may have diverse semantic meanings (Marrone et al., 2022). Addressing these challenges is crucial for enhancing the interoperability of healthcare data, improving the accuracy and efficiency of clinical practice, and facilitating robust research and public health surveillance efforts. Medical entity normalization—also called *entity linking* or *entity disambiguation*—emerges as a pivotal strategy to tackle these issues, involving the mapping of terms or phrases from clinical text to standard concepts or terminologies (Wajsbürt et al., 2021; Noh & Kavuluru, 2021; French & McInnes, 2023). Generally, medical entity normalization is preceded by the entity recognition task, in which the mentions or sequences of tokens in the text that belong to a defined entity class are identified (e.g. the mention 'kidney failure' belongs to the entity class Disease). After the entity recognition step, the normalization task is performed, in which each of the textual mentions of the entities is assigned to their corresponding concept within a knowledge base or standard terminology. Those mentions that cannot be matched to a concept in the knowledge base are called NIL, out-of-KB, or CUI-less entities (Shen et al., 2015) and are addressed using different approaches in a separate task (NIL entity linking task) (Ruas & Couto, 2022).

One of the main problems with the normalization task is that it predominantly relies on manual efforts by expert staff, rendering it both costly and time-consuming (Pérez et al., 2018). Moreover, developing computational systems to automate this process remains challenging. French and McInnes (2023) shed light on several critical aspects for future biomedical entity normalization efforts, underscoring the lower performance of normalization systems on non-English languages, inadequacies in datasets capturing mention ambiguity, and the need to explore alternative performance metrics such as ontological similarity.

Furthermore, the dearth of gold standard corpora for evaluation impedes the development and assessment of automatic normalization methods (Ferré & Langlais, 2023). In this context, it becomes imperative to address the unique challenges faced by low-resource languages in entity normalization tasks. These languages often lack comprehensive linguistic resources and face difficulties in accessing or adapting existing tools and methodologies (Névéol et al., 2018; Magueresse et al., 2020; García-Durán et al., 2022).

Therefore, the present study focuses on entity normalization in Spanish, a language with scarce resources for Natural Language Processing (NLP) tasks in the medical field. Our main contribution is an updated resource for entity normalization in the clinical domain, which is characterized by high lexical or terminological variability. We leveraged the Chilean Waiting List Corpus (CWLC) (Báez et al., 2020) and mapped its six medical entity classes, employing the Unified Medical Language System (UMLS) terminology (Bodenreider, 2004) and the Medical Lexicon for Spanish (MedLexSp) (Campillos-Llanos, 2023). The automatic normalization was subsequently manually reviewed by medical professionals. We meticulously evaluated the MedLexSp performance and manual coding quality to ensure a high-quality final resource. We are contributing a new version of the CWLC with a final size of 10,000 manually annotated clinical notes, containing 20% of the notes with entities normalized to the UMLS concepts. Detailed error analysis and linguistic analysis of the sources of lexical variation in the corpus accompany the resource evaluation.

2 Background

Strategies to normalize entities in the biomedical domain include rule-based methods (Kang et al., 2013; Ghiasvand & Kate, 2014; Afzal et al., 2016; D'Souza & Ng, 2015) and machine learning-based or deep learning-based approaches (Xu et al., 2017; Li et al., 2017; Luo et al., 2018; Fakhraei et al., 2020; Ji et al., 2020; Wajsbürt et al., 2021), most of which have been developed for the English language. For details on the technical components of entity normalization systems, we refer the reader to the work of French and McInnes (2023), who also provide a comprehensive list of datasets available for training the systems. Since we are interested in generating a new resource for medical entity normalization in Spanish, we list in Table 1 the few corpora available for this task and briefly describe below the annotated entities and the terminology used in the normalization process for each corpus.

The Mantra gold-standard corpus (Kors et al., 2015) is a multilingual resource for biomedical concept recognition with text from different parallel corpora in English, French, German, Spanish, and Dutch. The corpus was annotated and normalized to UMLS terminology for the following entities: Anatomy, Chemicals and drugs, Devices, Disorders, Geographic areas, Living beings, Objects, Phenomena, Physiology, and Procedures.

The PharmaCoNER (Gonzalez-Agirre et al., 2019), CodiEsp (Miranda-Escalada et al., 2020), DisTEMIST (Miranda-Escalada et al., 2022), MedProcNER (Lima-López et al., 2023) and SympTEMIST (Lima-López et al., 2023) corpora correspond to the same set of clinical cases from multiple medical specialties. These texts were manually annotated with chemicals and drugs, diagnosis and procedures, diseases, clinical procedures, and clinical symptoms, signs and findings, respectively. These mentions were manually normalized to concept unique identifiers mainly from the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) terminology (Donnelly, 2006).

Table 1 Corpora available for medical entity normalization in Spanish

Corpus	Documents	Tokens	Unique tokens	Lexical diversity[a] (&)	References
Mantra	200	3507	1360	38.8	Kors et al. (2015)
PharmaCoNER	1000	403,788	24,975	6.2	Gonzalez-Agirre et al. (2019)
Cantemist	1301	1,095,013	29,872	2.7	Miranda-Escalada et al. (2020)
CodiEsp	1000	403,788	24,975	6.2	Miranda-Escalada et al. (2020)
CT-EBM-SP	1200	341,596	19,301	5.65	Campillos-Llanos et al. (2021)
FALP	2402	1,267,396	31,592	2.5	Villena et al. (2021)
E3C[b]	1134	403,487	8362	2.1	Magnini et al. (2021)
LivingNER	1985	1,225,727	42,940	3.5	Miranda-Escalada et al. (2022)
DisTEMIST[c]	1000	403,788	24,975	6.2	Miranda-Escalada et al. (2022)

[a]Ratio of different unique tokens to the total number of tokens. [b]Only for Layer 2 in Spanish: medical entities automatically recognized by dictionary matching and annotated with their corresponding concepts in UMLS. [c]The same corpus was used in the MedProcNER task, reported in Lima-López et al. (2023); and in the SympTEMIST challenge, described in Lima-López et al. (2023)

The CANTEMIST corpus (Miranda-Escalada et al., 2020) contains mentions of tumor morphology, manually mapped to the latest Spanish version of the International Classification of Diseases - Oncology (ICD-O). The Clinical Trials for Evidence-Based Medicine in Spanish (CT-EBM-SP) corpus (Campillos-Llanos et al., 2021) contains annotations of the following entities: anatomy, pharmacological and chemical substances, pathologies, and procedures, with a small fraction normalized to the UMLS. The FALP Corpus (Villena et al., 2021) is a resource composed of anonymized pathology reports annotated with tumor morphology and topography mentions also mapped to the ICD-O. The European Clinical Case Corpus (E3C) is a multilingual resource that contains texts from different medical areas, annotated with several medical entities and with 50,000 tokens of disorders automatically mapped (without manual assessment) to the UMLS (Magnini et al., 2021). The LivingNER corpus is a collection of clinical cases from 20 medical specialties annotated with species and infectious diseases mentions and normalized to the NCBI Taxonomy terminology (Miranda-Escalada et al., 2022).

In under-explored languages, such as Spanish, there is an urgent need to develop resources that allow, for example, terminology expansion by mapping individual clinical mentions to existing codes or applying deep learning to automatically identify relevant mentions in clinical texts and link them to corresponding codes (Kugic et al., 2023). In this paper, we aim to contribute to the available Spanish entity normalization resources by publishing a new version of the CWLC with UMLS-mapped entities and analyzing the ambiguities identified during the normalization process.

Table 2 Corpus and sub-corpus statistics

Metric	Total	Medical	Dental
Documents	10,000	5000	5000
Tokens	348,660	208,125	140,535
Mean (SD) tokens per document	34.9 (±27.1)	41.6 (±34.9)	28.1 (±12.6)
Entities	68,046	38,201	29,845
Mean (SD) entities per document	6.8 (±4.7)	7.6 (±5.9)	6.0 (±2.7)
Annotated tokens	205,213	108,656	96,557
Unique tokens	31,323	25,398	10,693
Lexical diversity[a]	9.0%	12.2%	7.6%

[a]Ratio of different unique tokens to the total number of tokens

SD: Standard deviation

Table 3 Distribution of annotations per entity class

Entity	Total	Medical	Dental
Disease	21,459	12,987	8472
Finding	22,421	12,547	9874
Body part	14,798	6846	7952
Procedure	6934	3855	3079
Medication	1931	1503	428
Family member	503	463	40

3 Materials and methods

3.1 Materials

3.1.1 The Chilean waiting list corpus

The CWLC version published here is a 10,000 medical and dental referrals collection written in Spanish. A referral is a clinical note written by a general practitioner who refers a patient to see a specialist, describing the patient's conditions that trigger the need for evaluation by the specialist. The CWLC was manually annotated with six medical entities: Finding, Procedure, Disease, Family Member, Body Part, and Medication. The methods used and the quality metrics estimated in the corpus development were previously published on a 3000 sample referrals (Báez et al., 2022). The corpus is freely accessible,[1] and comprises 68,046 entities; 38,201 in medical referrals and 29,845 in dental referrals (Table 2). Note that half of the CWLC corresponds to dental texts, which are not common among the linguistic resources available for clinical NLP tasks. The distribution of annotations by

[1] https://zenodo.org/record/7555181

entity class can be seen in Table 3. Along with this extended version of the CWLC, we present a sample of 2000 medical referrals whose entities were normalized to UMLS Concept Unique Identifiers (CUIs).[2]

3.1.2 The UMLS metathesaurus

The UMLS Metathesaurus (Bodenreider, 2004), developed by the U.S. National Library of Medicine, is a resource created primarily to address two major barriers to the ability of computers to extract information: the variety of mentions referring to the same concept and the absence of an established format for distributing terminologies. It contains a compilation of names, relationships, and associated information from more than 150 biomedical resources such as the International Classification of Diseases, Tenth Revision (ICD-10) (World Health Organization, 2004), Medical Subject Headings (MeSH) (Lipscomb, 2000), or SNOMED-CT.

The 2022AB version of the UMLS integrates more than 17 million names, 4.6 million concepts, and 8.9 million codes for 26 languages. The UMLS uses Concept Unique Identifiers (CUIs) to index synonym mentions from different sources. For example, the code `C0020538` corresponds to the concept *hipertensión* (‘hypertension’) from MeSH, but also to the concept *presión arterial alta* (‘high blood pressure’) from SNOMED-CT (and also to the acronym *HTA*). Mapping variant m̊ entions indexed with CUIs enhances the interoperability across different medical ontologies and thesauri. To avoid restricting the normalization of the CWLC to one or two thesauri, we used the UMLS as the reference.

3.1.3 The MedLexSp lexicon

MedLexSp is a unified medical lexicon for Medical NLP in Spanish. It was used for normalization because it includes UMLS CUIs, gathers 100,887 term entries, and 302,543 form variants with PoS information and UMLS semantic types. This lexicon was created with heterogeneous sources ranging from corpora (e.g., MedlinePlus), thesauri, and terminologies (e.g., MeSH or the ICD-10) and other resources such as OrphaData for rare diseases.

MedLexSp includes variant forms such as gender/number variants or conjugated verb forms for each term entry. This makes it possible to map a plural adjective (e.g., *gestantes*, ‘pregnant women’), verb forms (e.g., *gestando*, ‘gestating’) or lexically-different entity mentions (e.g. *embarazadas*) to the same concept (which has CUI C0032961 in the UMLS). The current lexicon version normalizes with strict matching; therefore, misspellings or tokenization errors would not be matched (e.g. **embrazadas*). Moreover, no word sense disambiguation is supported to date, which caused several false positive errors: e.g., *pecho* can be normalized to ‘breast’ (C0006141) or ‘chest’ (C0817096).

[2] https://zenodo.org/record/7757971

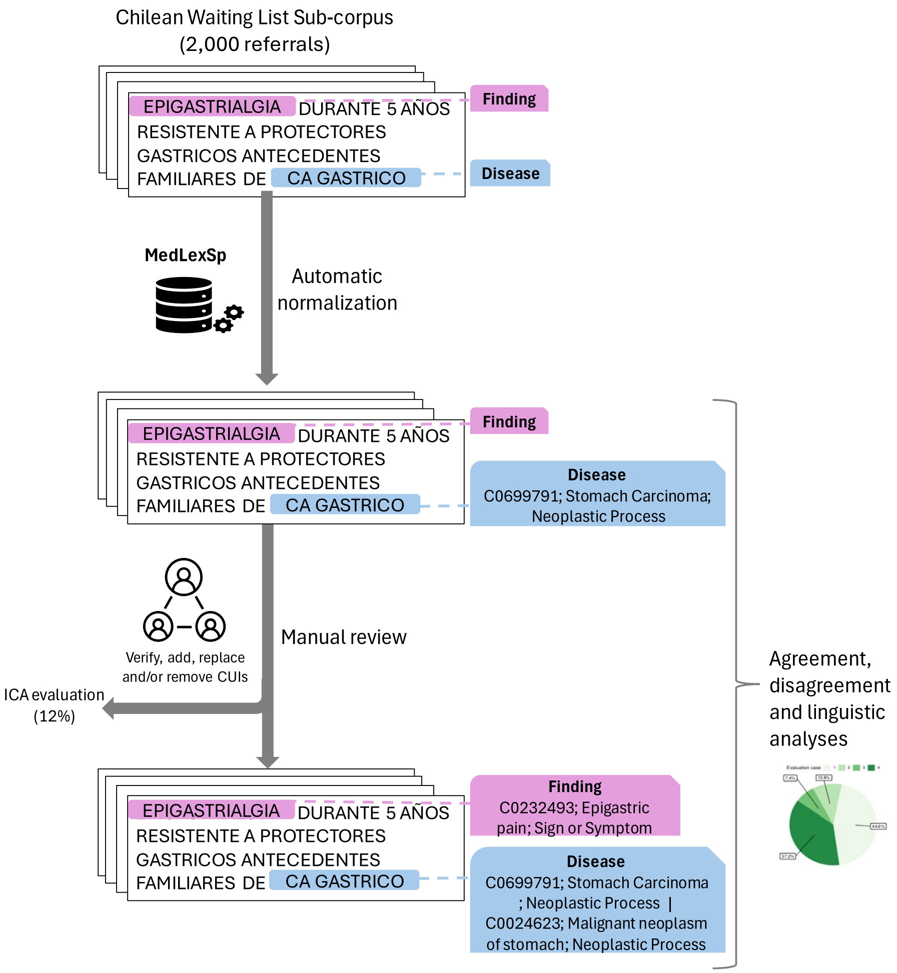

Fig. 1 Overview of the normalization process of the Chilean Waiting List sub-corpus. ICA: Inter-coder agreement

3.2 Methods

Figure 1 shows the overview of the process used to normalize the mentions of the 2000 medical referrals. A detailed description of the methods employed is provided below.

3.2.1 Automatic normalization

One of the advantages of using MedLexSp for automatic normalization is that its usefulness has been demonstrated in use cases in clinical NLP tasks, such as pre-annotation of clinical trial texts and PoS tagging/lemmatization of clinical cases, showing improved performance compared to the default Spacy and Stanza python libraries (Campillos-Llanos, 2023).

The BRAT Standoff format (Stenetorp et al., 2012) annotation files (.ann) from the CWLC were used in the automatic normalization process. These files were processed in batches using MedLexSp alongside a Python-based tool. The output generated was a file in identical format, where each annotated mention includes, as a comment, the assigned CUIs along with the concept description, separated by a semicolon. This format allows for manual modifications in BRAT.

3.2.2 Automatic-manual normalizacion

Revision of the automatic normalization

Three fourth-year medical students with experience in the manual annotation of the corpus were trained in the manual review process of the codes assigned automatically. The review was carried out using the BRAT web-based annotation tool and was based on guidelines[3] with decision and normalization criteria in case of ambiguities. The process consisted of manually verifying whether the UMLS code(s) assigned in the automatic normalization was appropriate for the annotated mention and, if necessary, removing, adding, or replacing codes.

A project manager with expertise in controlled biomedical vocabulary permanently supported the normalization process. Periodic meetings were held to resolve any doubts during the review stage. Once the normalization criteria were established and unified, they were documented in the normalization guidelines.

Inter-coder agreement

We assessed the normalization consistency after manual revision on a 12% parallel normalization of the corpus. We followed Hripcsak and Rothschild (2005) and the evaluation was based on precision, recall and F1 score, which were calculated with the following equations:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

We evaluated it in *strict* and *relaxed* conditions. In the relaxed condition, when more than one CUI was provided for one entity, an agreement occurred if at least one of the assigned CUIs matched any of the reference CUIs list. With the strict condition,

[3] https://zenodo.org/doi/10.5281/zenodo.10995018.

we only counted as an agreement case where the normalized CUIs perfectly matched (either a single CUI or a list of CUIs).

3.2.3 Evaluation of the normalization methods

Automatic and automatic-manual normalization performance Normalization performance using the lexicon MedLexSp was evaluated by measuring the F1 score between automatic and automatic-manual normalization. We employed the equality of proportions hypothesis test to determine whether the proportions of agreement b̊ etween normalization methods across all entities were equal or differed significantly between at least two entity classes. In addition, we performed pairwise comparisons of proportions between entities to identify the pairs with a difference in proportions. All analyses and data visualization were performed in R version 4.2.1 (R Core Team, 2022) using the tidyverse, easyalluvial, and arcdiagram libraries (Wickham et al., 2019; Koneswarakantha, 2022; Sanchez, 2018).

Agreement and disagreement distributions

After a manual review of the normalization, four possible scenarios of agreement and disagreement were considered:

- Case 1: The normalization is correct.
- Case 2: The entity was not normalized, and no applicable codes were identified manually. These cases correspond to CUI-less entities, which were not further analyzed since they are out of the scope of the present study.
- Case 3: The normalization was not correct. There are two subtypes:
  - Case 3.1: The normalization was partially correct: one or more codes that were also relevant and were omitted must be added, or one or more codes that the lexicon added but were not appropriate must be removed.
  - Case 3.2: The normalization was incorrect and was changed entirely.
- Case 4: The automatic normalization did not identify applicable codes for the mentions, but the manual method did.

To better understand the differences in the normalization methods, we studied in more detail the distribution of agreements and disagreements. Agreements were considered for Cases 1 and 2 (the normalization is correct; or the entity was not normalized and no applicable codes were manually identified). Disagreements were considered for Cases 3 and 4 (the normalization was not correct, or the automatic normalization did not identify applicable codes for the mentions, but the manual method did). The distribution of agreements and disagreements under these four scenarios was analyzed.

Linguistic analysis Since cases 3 and 4 are considered normalization errors, we selected a random sample of 100 errors from each case to manually determine the origin of the error in the automatic normalization. The analysis was performed by a native Spanish-speaking linguist with experience in biomedical text analysis. From this sample, we observed three categories of linguistic inadequacies that were subclassified as follows:

- Semantic: related to ambiguity due to polysemy, among others.
  - Hyponymy/hyperonymy: Lexical relation that describes the inclusion of the meaning of a lexical unit into another (like in an IS-A taxonomy). For example, the term *biological sex* (hypernym) contains the meaning of the lexical units *female*, *male*, and *intersex* (hyponyms).
  - Polysemy/metonymy: Phenomenon in which a lexical unit expresses a different meaning (polysemy) or when two lexical units are related by a whole-part classification (like in a PART-OF taxonomy).
  - Acronym-abbreviation: abbreviated forms of writing, typically not standardized in the medical sub-corpus.
- Grammatical: related to the variation in the internal structure of lexical units.
  - Inflection: Property of the internal structure of lexical units that allows the expression of information related to the gender or number variation in nouns.
  - Category: Meanings realized through different grammatical categories and their differences between English and Spanish.
  - Morphological: Differences between the selection of derivational morphemes in English or Spanish.
- Ortho-typographic: related to misspellings or typos.

Each inadequacy was further subclassified into inflection (gender or number variation), acronym-abbreviation, hyponymy/hyperonymy, polysemy/metonymy, category, and morphological.

3.2.4 Final normalized sub-corpus exploration

In recent years, several studies have explored different visualization methods and their potential to improve the understanding of medical phenomena, synthesize information, search for patterns, and ultimately support clinical decisions and policy management in public health (West et al., 2015; Roham et al., 2019; Chen et al., 2022).

Here we tested an arc diagram to visualize the co-mentions of pairs of codes for Disease and Medication entities and discuss our observations contrasted with the epidemiological data reported in the 2016-17 Chile National Health Survey (Margozzini & Passi, 2018) and the MAUCO cohort study (Oyarzún-González et al., 2020). MAUCO is a Chilean cohort study designed to study the natural history of chronic diseases in Molina, a semi-rural city. The arc diagram is a particular network graph form that allows the visualization of complex repetitions in string data such as DNA sequences (Wattenberg, 2002). In the arc diagram, the nodes represent the entities, and the links show the relationships between the entities (Byrne et al., 2007).

Table 4 Frequency of normalized entities and distribution of unique codes per entity class using the automatic and automatic-manual methods

		Normalization							
		Automatic			Automatic-manual			Difference[1]	
Entity	Frequency	NE	%	UC	NE	%	UC	NE	%
Disease	6094	3152	51.7	932	5925	97.2	2005	2773	45.5
Finding	4351	1416	32.5	513	3306	76.0	1496	1890	43.4
Body part	2680	2068	77.2	651	2584	96.4	1033	516	19.3
Procedure	1743	954	54.7	277	1631	93.6	607	677	38.8
Medication	609	407	66.8	150	592	97.2	226	185	30.4
Family member	204	163	79.9	11	202	99.0	26	39	19.1
Total	15,681	8160	52.0	2534	14,240	90.8	5393	6080	38.8

[1] Difference between automatic and automatic-manual normalization figures

NE: Normalized entities

UC: Unique codes

Table 5 Statistics on the sub-corpus of normalized medical referrals

Documents	2000
Tokens	84,295
Mean (SD) tokens per document	42.1 (±35.3)
Entities	15,681
Mean (SD) entities per document	2.7 (±2.5)
Annotated tokens	41,938
Unique tokens	14,115
Lexical diversity[1]	16.7%

[1]Ratio of different unique tokens to the total number of tokens

SD: Standard deviation

4 Results

The collection of 2000 medical referrals contained a total of 15,681 entities. The distribution by entity class is shown in Table 4, and the overall statistics of the sub-corpus are shown in Table 5.

4.1 Automatic normalization

The distribution of normalized entities and unique codes per entity class, in both automatic and automatic-manual normalization, are shown in Table 4. Of the total number of entities in the sub-corpus, 8160 (52%) were automatically normalized with 2534 unique codes. The most frequent automatically normalized entity was Family Member (79.9%), followed by Body Part (77.2%), while the least frequently

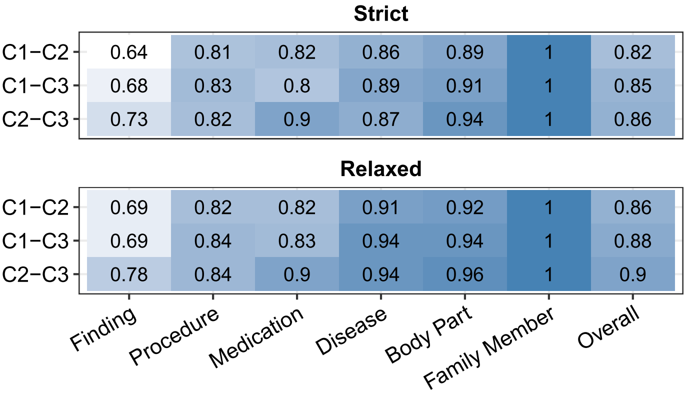

Fig. 2 Inter-coder agreement (strict and relaxed F1 score) per pair of coders on manual review of entity normalization

normalized was Finding (32.5%). The highest proportion of unique codes was assigned to Disease (932), followed by Body Part (651), while Family Member had the smallest (11).

4.2 Manual review and normalization

The inter-coder agreement ensured the quality and consistency in the manual review of the normalization. The overall average F1 score was 0.84±0.2 (strict metric) and 0.88±0.2 (relaxed metric). Figure 2 shows the inter-coder agreement values by entity class and coder pair. The highest disagreement between coders was observed for the Finding entity. As will be seen from the examples in the disagreement analysis, some differences could be mainly due to the annotation length or span, where multi-word modifiers are included or excluded.

Figure 3 shows the change in the statistical distribution of the Body Part and Disease entities after being normalized. The effect of the normalization is reflected in a modification of the general pattern of mention/string occurrence represented in Zipf's law. This is attributable to the reduction of lexical variability, where equivalent mentions such as *hta* and *hipertensión arterial* are mapped to the same UMLS concept (Hypertensive disease in this case). Table 6 shows that in the top 10 of the most common mentions are diverse diseases such as *hipertensión arterial* ('high blood pressure'), particularly, its abbreviation *hta* ('hbp'); findings such as *dolor* ('pain'); and body parts such as *adenoides* ('adenoids'). Regarding the Family member entities, the word *madre* ('mother') was the most frequent one.

After manual review and normalization, the number of normalized entities increased to 14,240, equivalent to 38.8% more entities, with more than twice as many unique codes (5393) compared to automatic normalization (Table 4). The most frequently normalized entity was Family Member (99%), followed by Disease and Medication (97.2% each), while the least frequently normalized was again Finding (76%). As can be seen, the most notable difference in the percentage of normalized

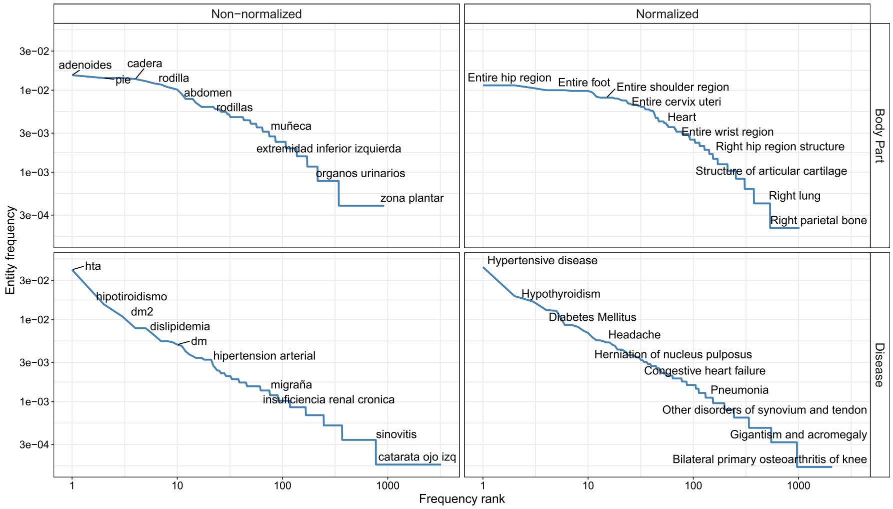

Fig. 3 Comparison of Zipf plots for normalized and non-normalized Body part and Disease entities (log-log scales). Only the text for ten entities is shown in each plot. The non-normalized entities panel (left) shows the annotated text, while the normalized entities panel (right) shows the string corresponding to the UMLS concept to which each entity was manually normalized

Table 6 Most frequent normalized mentions and top ranking by entity class

Rank	Frequency	Mention (English translation)	Entity class
1	238	*hta* (‘hbp’)[1]	Disease
2	90	*hipotiroidismo* (‘hypothyroidism’)	Disease
1	88	*madre* (‘mother’)	Family Member
3	64	*dm2* (‘dm2’)[2]	Disease
1	52	*dolor* (‘pain’)	Finding
4	46	*colelitiasis* (‘gallstone disease’)	Disease
5	46	*dislipidemia* (‘dyslipidemia’)	Disease
2	42	*embarazo* (‘pregnancy’)	Finding
1	39	*adenoides* (‘adenoids’)	Body Part
6	39	*trastorno de la refracción, no especificado (‘disorder of refraction, unspecified’)*	Disease

[1] *hbp*: high blood pressure. [2] *dm2*: type 2 diabetes mellitus

entities was in Disease, Finding, and Procedure, with an increase of almost 40% in the number of normalized entities manually reviewed. The slightest difference was in Family Member, with an increase of 19.1%.

The highest proportion of unique codes was assigned to Disease (2005), followed by Finding (1496), while Family Member had the smallest number (26). Regarding the cases of entities that can be normalized with multiple codes, the entity Body Part presented the highest average of multiple codes per mention,

Table 7 Mean number of codes per entity class assigned to annotations after manual review

Entity	Mean	SD[1]	Min	Max
Body part	1.84	0.61	1	4
Procedure	1.09	0.29	1	3
Finding	1.08	0.29	1	4
Disease	1.07	0.27	1	4
Medication	1.04	0.22	1	3
Family member	1.03	0.18	1	2
Total	1.20	0.46	1	4

[1] SD: Standard deviation

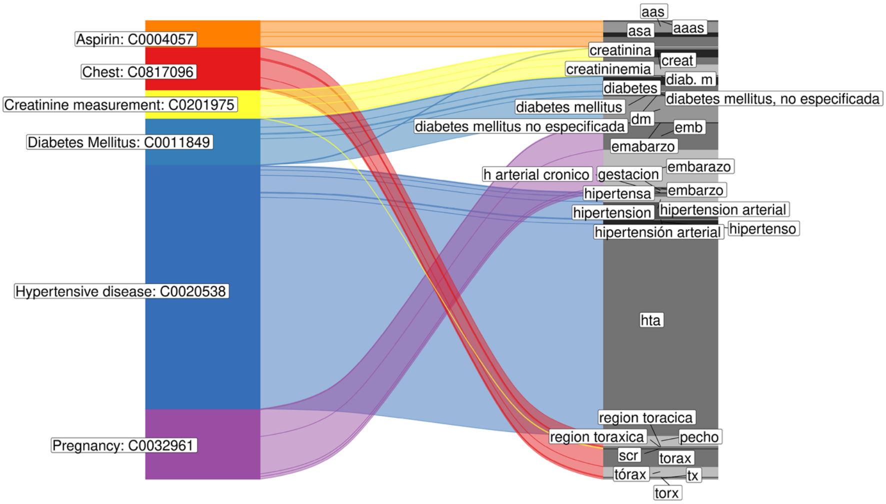

Fig. 4 Alluvial plot of the lexical variability captured using standard UMLS concepts. The left side shows the UMLS concepts with their respective CUIs, while the right side shows some of the normalized annotation strings

with an average of 1.8± 0.61 codes per mention, while the rest of the entities had averages far below that figure (Table 7).

The number of unique mentions contained in the annotations was 8710, whereas those contained in the CUIs strings were 4924, which is equivalent to a decrease in concept variability of 56.5% after normalization. The alluvial plot in Fig. 4 shows how normalization captures the lexical variability using the standard UMLS concepts, which facilitates and guarantees a correct subsequent analysis of the information.

Table 8 Agreement between automatic and automatic-manual normalization. The metrics were calculated using automatic-manual normalization as the gold standard

Entity	Precision	Recall	F1 score
Medication	0.96	0.66	0.78
Body part	0.92	0.66	0.77
Family member	0.94	0.53	0.68
Procedure	0.89	0.54	0.67
Overall	0.86	0.54	0.66
Disease	0.94	0.49	0.64
Finding	0.71	0.51	0.60

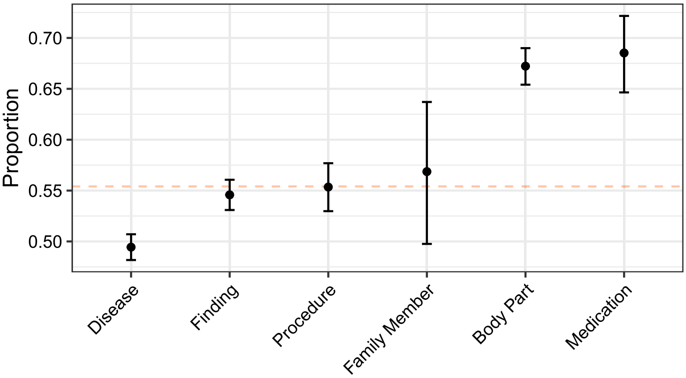

Fig. 5 Proportions and 95% confidence intervals of automatic and automatic-manual agreement per entity

4.3 Evaluation of the normalization methods

4.3.1 Automatic and automatic-manual normalization performance

Similar to the agreement between coders, the lowest agreement between automatic and automatic-manual normalization was observed for Finding, Disease, and Procedure entities with F1 scores of 0.60, 0.64, and 0.67, respectively (Table 8).

Figure 5 shows the difference in the proportion of agreements and their uncertainty for each entity class. The proportions of agreements are significantly different between entities (*p* value <0.001), suggesting that the MedLexSp lexicon sometimes fails to correctly normalize more complex entities such as Disease, Finding, and Procedure in these types of texts. A possible explanation of the drop in normalization performance for these entity classes may be the lack of some entries in the lexicon, especially acronyms/abbreviations that are common in Chilean Spanish or waiting list notes but not recorded in MedLexSp. In addition, these entity mentions may have a wider variability or blurred categories in the UMLS, whereas mentions of Medication, Body Part, or Family Member are more invariable and standardized.

Table 9 Automatic and automatic-manual agreement pairwise comparisons using pairwise comparison of proportions

	Body part	Disease	Family member	Finding	Medication
Disease	< 2e-16	–	–	–	–
Family Member	0.02	0.21	–	–	–
Finding	< 2e-16	1.7e−06	0.99	–	–
Medication	0.99	< 2e-16	0.02	1.2e−09	–
Procedure	3.3e−14	0.0001	0.99	0.99	1.6e−07

It is important to note the uncertainty—reflected in the 95% confidence interval—in the agreement proportions of the entities with the lowest representation in the corpus, such as Family Member and Medication. Because of this uncertainty, the proportions are not significantly different between the entities Disease, Finding, and Procedure and Family member, nor between Medication and Body part, as reflected in the pairwise comparisons of automatic and automatic-manual agreement in Table 9.

4.3.2 Agreement and disagreement distributions

Regarding agreements and disagreements between automatic and automatic-manual normalization, agreements were distributed in 44.6% of entities in Case 1 (correct normalization) and 10.8% in Case 2 (no applicable codes by both methods or CUI-less entities). Disagreements were distributed in 7.4% (1161 entities) in Case 3 (incorrect normalization) and 37.2% (5831 entities) in Case 4 (manual normalization only) (Fig. 6a). Case 3 can be subdivided into 5.7% (896 entities) in Case 3.1 (partially correct normalization) and 1.7% (265 entities) in Case 3.2 (completely incorrect normalization). The most significant number of completely incorrect normalized entities (Case 3.2) was found in the Finding and Disease classes (Fig. 6b).

4.3.3 Linguistic analysis

Table 10 shows the distribution of linguistic inadequacies among Cases 3.1, 3.2, and 4. The 100 errors in the Case 3 sample were distributed into 77 in Case 3.1 and 23 in Case 3.2. The most common linguistic inadequacy subclass in Case 3.1 was Polysemy/Metonymy (52%), followed by Hyponymy/Hyperonymy (29.9%), while in Case 3.2, the most common subclass was Polysemy/Metonymy (47.8%), followed by Acronym-abbreviation (30.4%). In Case 4, the most common subclass was Hyponymy/Hyperonymy (52%), followed by Polysemy/Metonymy (26%). The ortho-typographical inadequacy was only present in Case 4 normalization errors (4%).

The most frequent subclasses of linguistic inadequacies among the three error cases were Hyponymy/Hyperonymy, Polysemy/Metonymy, and Acronym-abbreviation, with 39%, 38.5%, and 17%, respectively, of the total inadequacies (Table 10).

In the following, we describe each linguistic inadequacy and its corresponding subclasses with some examples.

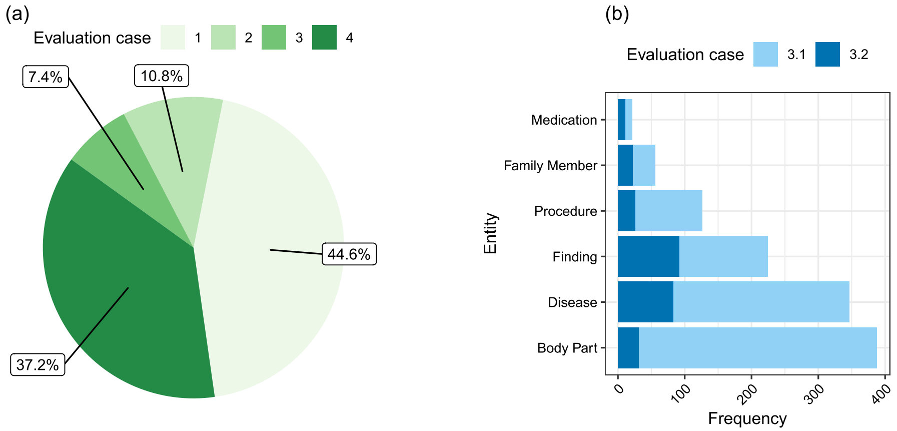

Fig. 6 **a** Distribution of entities according to the possible four scenarios in the agreement evaluation between automatic and automatic-manual normalization. *Case 1:* correct normalization. *Case 2:* no applicable codes by both methods. *Case 3:* incorrect normalization. *Case 4:* manual normalization only. **b** Frequency of entities in the two subtypes of scenario 3. *Case 3.1:* partially correct normalization. *Case 3.2:* completely incorrect normalization

Table 10 Distribution of normalization errors (Cases 3 and 4) in classes and subclasses of linguistic inadequacies

Inadequacy		Case 3.1		Case 3.2		Case 4		Total	
Class	Subclass	n	%	n	%	n	%	n	%
Grammatical	Category	0	0	1	4.4	0	0	1	0.5
	Inflection (gen)	0	0	1	4.4	0	0	1	0.5
	Inflection (num)	3	3.9	0	0	1	1	4	2.0
	Morphological	0	0	0	0	1	1	1	0.5
Semantic	Acronym-abbreviation	11	14.3	7	30.4	16	16	34	17.0
	Hyponymy/Hyperonymy	23	29.9	3	13	52	52	78	39.0
	Polysemy/Metonymy	40	52	11	47.8	26	26	77	38.5
Ortho-typographical	–	0	0	0	0	4	4	4	2.0
All	–	77	100	23	100	100	100	200	100

- Semantic inadequacies The first semantic inadequacy is the Hyponymy/Hyperonymy error, in which a lexical or phraseological unit is not precise concerning those available in the corpus because the meaning of a lexical unit is included into another. In example 4.3.3, the item *columna* ('column') can refer to either a specific part of the body or the entire anatomical structure, which would correspond to two different codes. Furthermore, example 4.3.3 shows an annotation in Spanish (in this case, a disease), which was not recognized among the terms in the lexicon, as shown in examples 4.3.3 and 4.3.3:

1. *columna* (‘column’) normalized to ‘vertebral column’ (`C0037949`) but should also be normalized to ‘entire vertebral column’ (`C1267072`).
2. *crisis de ausencias* should be normalized to ‘absence seizures’ (`C4316903`).

The second semantic inadequacy is Polysemy/Metonymy. The automatic coding was inadequate due to a specificity problem regarding the annotation and the full description made by the practitioner, as shown in example 4.3.3. In example 4.3.3, on the other hand, a case of metonymy is shown, in which the linguistic label *lentes* (‘glasses’) intends to refer to the fact that the patient wears glasses. This problem is related to the extended meaning of that lexical item or annotation.

3. *pecho* normalized to ‘breast’ (`C0006141`) and ‘chest’ (`C0817096`) should correspond only to ‘chest’ (`C0817096`).
4. *lentes* (‘glasses’) should correspond to ‘eyeglasses wearer’ (`C0920139`).

The third semantic inadequacy is the acronym-abbreviation error, in which abbreviations are used locally but standardized in the lexicon as specific entities, as in example 4.3.3, or acronyms suggested or created by the practitioner but not available in the vocabulary, as in example 4.3.3:

5. *HCT* normalized to ‘human calcitonin’ (`C0770558`) should be normalized to ‘hydrochlorothiazide’ (`C0020261`).
6. *VIF*, an acronym for *violencia intrafamiliar* (‘intrafamily violence’), which is not available in the lexicon, should be normalized to `C0206072`.

- Grammatical inadequacies The grammatical inadequacies refer to aspects related to the internal structure of word forms. We have identified three types: category, inflection (gender or number variation) and morphological derivation. Example 4.3.3 shows the subcategory of inadequacy related to a grammatical category, in which the Spanish adjective *diabéticas* (‘diabetic women/girls’) refers to a disease lexicalized as a noun in English:

7. *diabética* (‘diabetic’) should correspond to *diabetes* (`C0011847`).
8. *abuela* (‘grandmother’), normalized to *abuelo* (‘grandfather’; although it may correspond to an identification of the masculine noun as an unmarked gender for Spanish) should be normalized to ‘grandmother’ (`C0337474`).[4]

The problem of grammatical inflection corresponds to examples 4.3.3 for gender and 4.3.3 for number. In these cases, it is not possible to automatically identify the string in English due to the designation of the inflectional paradigm in Spanish:

[4] Although the masculine noun may correspond to the unmarked gender in Spanish.

9. *inhaladores* (‘inhalers’) should be normalized to ‘inhaler’ (C0021461) (although it may correspond to a problem generated by the plural allomorphy in Spanish, in which the suffix *-s* is used to construct the plural in words ending in a vowel, while the suffix *-es* is used for words ending in a consonant).

Regarding the morphological derivation error, we observed a tendency in the lexicalization that practitioners use to refer to Findings and Procedures as attributes of patients. This is realized linguistically in Spanish as a participle that functions as an adjective, with the ending *-ado/-ada*. This label is complex because it is not consistent with nominalization in nouns, as shown in example 4.3.3:

10. *operada* (as in ‘operated’) corresponds to the Procedure category if the patient had ‘cancer surgery’ (C0920424).

- Ortho-typografical inadequacies The final type of inadequacy is ortho-typographical. There is significant variability in the spellings for specific specialized mentions. This inadequacy leads to spelling errors that are difficult to anticipate and occur predominantly for entities Disease and Procedure. We observed ortho-typographical error in examples 4.3.3 and 4.3.3:

11. *intuvacion orotraquial* (sic) instead of *intubación orotraqueal* was not automatically normalized to ‘Orotracheal intubation’ (C0396621) because the word *intuvation* was misspelled with *v* instead of *b*, and *orotraquial* was misspelled with *i* instead of *e*.
12. *gondrosis intervertebral* (sic) instead of *condrosis intervertebral* should be normalized to ‘intervertebral chondrosis’ (C1262204).

- Other automatic normalization errors

In Case 4 errors, 408 entities of the class Disease were identified whose mention corresponded to a UMLS concept of type “unspecified”, e.g., *Hipotiroidismo, no especificado* (‘Unspecified hypothyroidism’, C0020676), as well as 57 cases of type “not elsewhere classified”, e.g., *Lipomatosis, no clasificada en otra parte* (‘Lipomatosis, not elsewhere classified’, C0869206), and in less proportion the “Not Otherwise Specified (NOS)” concepts. Also, 210 entities of the type “Other... ” were not normalized, e.g., *Otras inmunodeficiencias* (‘Other immunodeficiencies’, C0494264) or *Otras psoriasis* (‘Other psoriasis,’ C0477485). A specific Case 4 error concerns geographical variation between Chilean and Peninsular Spanish. For example, the term *pellet* refers to a tablet drug that is used to treat alcoholism, but this entity is not used in Peninsular Spanish. Because the lexicon was not prepared from sources in Chilean Spanish, MedLexSp did not normalize it and annotators added the CUI manually. However, we did not find many instances of this error as expected, since written texts are more standardized than oral transcripts.

A particularity of the CWLC is that the mentions of codes from other terminologies, especially ICD-10, were annotated as entities in their respective categories. In this sense, 240 entities corresponded to codes that were not automatically normalized to UMLS, e.g., `C71.0` (*Neoplasia maligna de cerebro, excepto lóbulos y ventrículos*; 'Malignant neoplasm of cerebrum, except lobes and ventricles', `C0153634`).

It is also worth mentioning that approximately 70% of the mentions (1181) with problems due CUI-less mentions (Case 2) occurred in the Finding entity, where the concepts can be vague, and the limits of the mentions are not so clear. For example, the mention *miedos y temores a nivel familiar* ('fears and apprehensions at family level'), could be thought of as normalizing to 'family tension' (`C0577730`), but the mention does not necessarily reflect that concept. Other examples with very general mentions are *nervios y medula espinal a nivel del cuello* ('nerves and spinal cord at neck level') and *tendones duros a la palpitacion* ('hard tendons at palpitation'), the latter case with a severe spelling error ('palpation' was changed to 'palpitation'). Finally, in *gastropatia erosiva antral leve* (mild antral erosive gastropathy), a modifier was added to the entity which is not considered in the concept ('Erosive gastropathy,' `C2243090`).

4.4 Final normalized sub-corpus exploration

As seen in the arc diagram in Fig. 7, normalization allows for a meaningful visualization of the top 60 co-mentions for medications and diseases that appear in CWLC referrals. The most common disease is 'hypertensive disease', frequently co-mentioned with other diseases such as 'diabetes mellitus, non-insulin-dependent', 'hypothyroidism' and 'dyslipidemias', whereas the most frequent medications are 'aspirin', 'losartan', 'metformin', and 'atorvastatin' which are frequently co-mentioned with 'hypertensive disease' and other less frequent diseases.

In this sample of only 60 top co-mentions, the majority corresponds to Disease, the most frequent entity class in the corpus. The second least frequent entity, Medication, is also present in a smaller proportion of these co-mentions. The figure provides relevant information on several aspects:

1. The possible presence of multimorbidity is suggested, given the high frequency of diseases co-mentioned with other diseases. Notably, 'hypertensive disease' is the one that presents the most associations with other diseases.
2. The relationships between medications and diseases reveal some treatment patterns. A significant set of medications co-mentioned with a single disease, in this case, 'hypertensive disease', is observed. This does not necessarily suggest that the medications are being used to treat the main pathology, but rather that they may be related to the multimorbidity of the patients.
3. A few medications co-mentioned with several diseases are identified, such as 'losartan' and 'furosemide', which are associated with a couple of diseases.

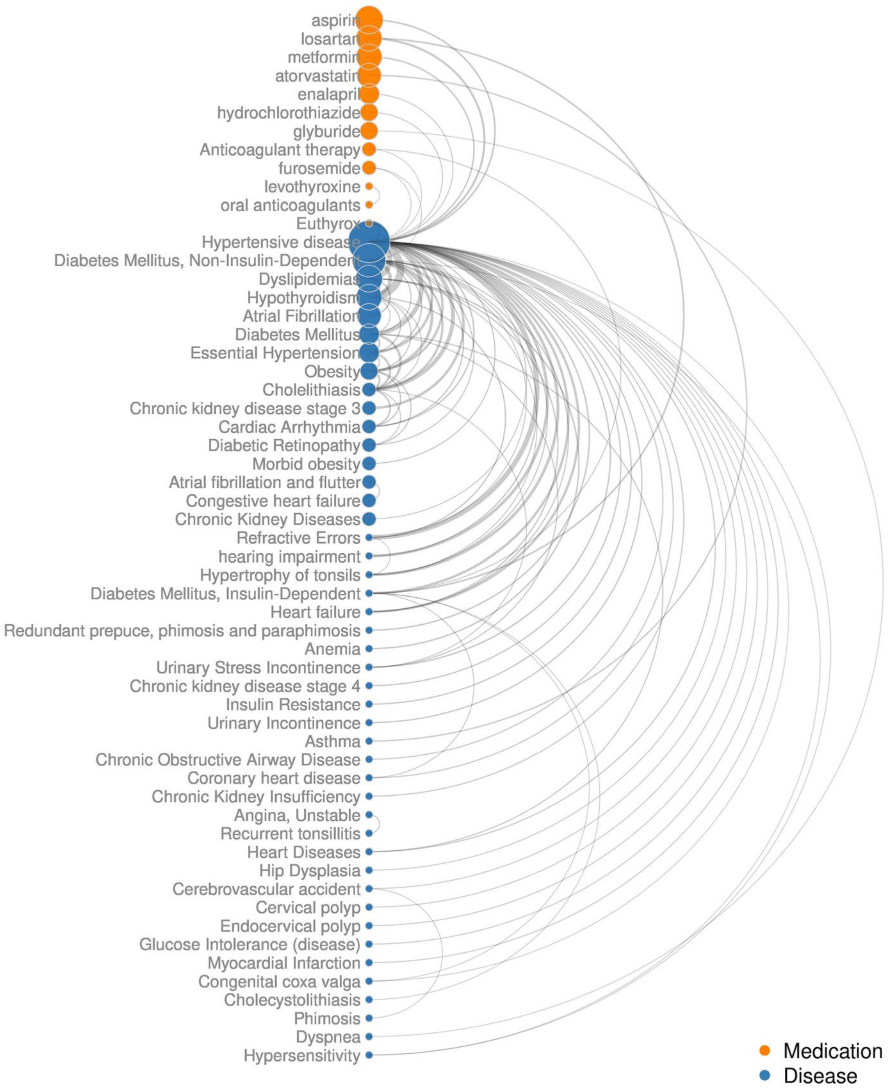

Fig. 7 Arc diagram of the top 60 UMLS CUIs co-mentions between Medication and Disease entities from the normalized Chilean Waiting List sub-corpus. The mentions are shown as strings instead of CUIs to facilitate the understanding of the figure

It should be considered that these observations could change by including a larger sample of co-mentions. Nevertheless, the observed pattern was expected due to the high prevalence of multimorbidities and their common registration in primary care medical texts, in contrast to the smaller registration of medications given to or consumed by the patient. This profile could vary considerably when analyzing other types of clinical texts, for example from hospitalized patients, where a greater number of medications and doses administered to the patient are usually registered.

5 Discussion

Normalization of medical entities remains challenging, especially for languages other than English, where the lexical variation found in standard terminologies such as UMLS is considerably under-represented (?). The state-of-the-art entity normalization in English is the BioPRO model (Zhu et al., 2024), which recently outperformed the SapBERT model (Liu et al., 2021) in 4 biomedical corpora.

External available resources are limited and there is a need to generate new resources for entity normalization that capture the linguistic variety of clinical text in languages other than English. (Liu et al., 2021) presented a novel multilingual entity normalization task (XL-BEL), in which they tested different ways of transferring domain-specific knowledge from English to other languages. Even if there was not much specific biomedical data available in those languages, the approach achieved quite promising results, with a Precision@1 close to 60 for Spanish.

Compared to the other available resources for Spanish listed in Table 1, our study shows that the variability of CWLC is higher (except for MANTRA) with a 16.7% lexical diversity. Therefore, normalizing the CWLC corpus automatically may be more difficult. This is unsurprising since the corpus comprises referrals from 23 health services across Chile and collects information from multiple healthcare centers and medical specialties (Báez et al., 2022). The terminological variability is also reflected in the normalization performance of the MedLexSp. Lexical and dictionary-based methods are limited precisely when there is high lexical diversity in the domain and when semantic meaning and sentence context need to be considered. The performance of MedLexSp was particularly good concerning its accuracy, while its recall was somewhat lower, which is related to the tendency to generate false negatives, a common disadvantage in systems of its type. Nevertheless, MedLexSp works as an automatic pre-coding method with acceptable performance, normalizing 52% of the total number of medical entities in the CWLC. These normalization figures are considerable given that the referrals contained in this corpus include relatively informal and noisy language with abundant abbreviations and spelling errors. In contrast, the performance of using MedLexSp to normalize entities in a corpus of clinical trials texts (CT-EBM-SP) reached an average of 70.68% (Campillos-Llanos et al., 2021)—note that the rest of the non-normalized entities are currently being revised manually. The different percentage values of automatic normalization seem to be aligned with the divergent linguistic properties of each corpus.

Our work has important implications for medical entity normalization, as we are contributing a resource for normalizing six entity classes in clinical texts. The sub-corpus construction presented robust inter-coder agreement metrics. The normalized sub-corpus can be employed in practical applications, especially for developing automatic entity and document normalization methods. Knowledge visualization methods can also improve understanding of medical and patient needs or look for patterns to provide better healthcare strategies. Entity normalization could favor a more effective visualization of the repeating pattern and relationship between the CWLC entities. Although the sub-corpus is a small sample of data, Fig. 7, for instance, closely depicts the epidemiological data reported in the National Survey, where the diseases with high prevalence in Chile correspond to hypertension (27.6%), HDL dyslipidemia (46%), obesity (34.4%), diabetes (12.3%) and hypothyroidism (18.6%) (Margozzini & Passi, 2018). It is also interesting to note the top disease, cholelithiasis, a disease of high prevalence in Chile whose risk factors are related to alterations in lipid metabolism, obesity, and type 2 diabetes (Cortés et al., 2020; Yuan et al., 2022), which are frequently co-mentioned in the referrals. Likewise, the list of most frequent medications is consistent with the most frequent diseases in the referrals and with the profile of medication use reported in both the National Survey (Margozzini & Passi, 2018) and the MAUCO cohort study (Oyarzún-González et al., 2020). Overall, the most commonly used medications in the MAUCO cohort were acetylsalicylic acid (aspirin) for general and preventive use, losartan, enalapril, furosemide, and hydrochlorothiazide, which are used in the treatment of hypertension, metformin and glibenclamide for diabetes, atorvastatin for dyslipidemia, and levothyroxine for hypothyroidism.

Despite the interesting nature of these observations, it should be noted that it is not possible to draw associations or conclusions from them. To validate an epidemiological study using electronic health records data, representativeness, availability, interpretability, and missing data and visits must be addressed, as described by Gianfrancesco and Goldstein (2021). In our analysis, for example, we did not take into account the year in which the referral was issued, which is very relevant considering that the prevalence of consumption of medications such as metformin, atorvastatin and levothyroxine among Chileans changed significantly between the National Surveys 2009–10 and 2016–17 (Ministerio de Salud, 2019). Nevertheless, our results show the great potential of data from national waiting lists such as the CWLC if used appropriately to test hypotheses by controlling possible sources of bias and confounding.

The linguistic analysis shows the most common linguistic inadequacies that explain the variation of mentions and normalization errors in the sub-corpus. Except for the inadequacies related to spelling, which compromise the graphic representation of the linguistic system, the most relevant cases for the error detection analysis were the semantic and grammatical inadequacies. Regarding the semantic inadequacies, to analyze the variations in the normalization, we have to consider the associations of terms that co-occur but are not equivalent and correspond, for example, to metonyms or hyponymy/hyperonymy. On the other hand, grammatical inadequacies, like word category of inflections, offer a relevant description of the morphological peculiarities that practitioners employ to refer to domain-specific uses of Chilean Spanish, which does not indicate that these grammatical choices are stabilized in the clinical vocabulary. In this paper, we have introduced a workflow that can serve as a model for researchers interested in extrapolating the results to other languages, especially Romance languages, such as French, Portuguese or Italian. The novelty of our work lies not only in creating a new linguistic resource for entity normalization but also as a benchmark for comparison in studies that replicate our analysis.

Our experiments also show the limitations of using only a lexicon for normalization. Firstly, the current version of MedLexSp would need to include more term variants and concepts. 39% of other entities were only normalized during the manual evaluation. Secondly, although the lexicon would need to be updated, the text genre and register make it difficult to normalize all entities automatically. Some reasons are the intrinsic variability of medical jargon, which varies even across hospitals, or semantic ambiguities such as polysemy. Combining a lexicon with complementary methods that model the linguistic context would achieve higher scores (e.g. word sense disambiguation or BERT-based normalization). Additionally, 7.4% of the entities presented disagreement in the normalization, with the most critical errors in the unsuitable codes assigned to the entities Finding and Disease. Despite the limitations, semi-automatic normalization using MedLexSp speeds up the task, avoiding the need to normalize manually from scratch and accelerating the mapping process of medical entities to concepts.

Having expert medical staff is critical in this type of study but costly. One point to highlight from our study is that fourth-year medical students performed the manual review of the normalization and were constantly supported by a project manager. They were well-trained in both corpus annotation and entity normalization and were already familiar with specific mentions and concepts of use in daily clinical practice. Follow-up through inter-coder agreement allowed us to control for possible inconsistencies.

One of the potential limitations of our study is that we only used a lexicon-based normalization method and did not analyze the performance of modern methods. The low agreement in the Finding entity can be seen as another limitation of our work. However, it is not surprising since it was previously reported as a problematic entity to annotate during corpus construction (Báez et al., 2022). Another limitation is that because the variety of concepts in UMLS in English is much wider than in Spanish (6:1 ratio), 1409 mentions were mapped to CUIs whose concepts appear only in English. Of these mentions, 575 corresponded to Disease, and 490 to Finding. The entities that can be normalized with multiple codes should be studied in depth in future work.

6 Conclusions

This paper presented a sub-corpus of the CWLC (a corpus developed for medical entity recognition in Spanish), consisting of 2000 medical referrals with 15681 entities, of which 14,240 were manually mapped to the UMLS CUIs. This corpus adds to the growing body of resources for NER and entity normalization tasks in Spanish that seek to contribute to improving healthcare outcomes.

The error analysis is another important contribution of this paper since it shows the performance of the MedLexSp lexicon for normalizing medical entities to UMLS terminology and some of its critical points for improvement; for example, it is important to expand the MedLexSp lexicon. In the same direction, the linguistic analysis offers insight into the sources of lexical variety in the Spanish clinical environment. Enriching dictionaries and lexicons with new terms and resources is fundamental to the functioning of modern entity normalization methods, with information retrieval-based architectures and state-of-the-art neural models such as SapBERT.

Indeed, future work may focus on using the sub-corpus of the CWLC to test state-of-the-art methods for medical entity normalization in Spanish, such as cross-lingual SapBERT (Liu et al., 2021) and ClinLinker (Gallego et al., 2024), which are based on the SapBERT model.

Acknowledgements We thank Antonia Arancibia, Matías Chaparro, and Tomás Bucarey for manually revising entity normalization on this corpus. In addition, we sincerely thank the reviewers for their valuable comments that helped improve the article.

Author Contributions Pablo Báez: Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Leonardo Campillos-Llanos: Conceptualization, Methodology, Software, Resources, Writing - Review & Editing. Fredy Núñez: Investigation, Writing - Review & Editing. Jocelyn Dunstan: Resources, Writing - Review & Editing.

Funding This study was supported by ANID Postdoctoral FONDECYT 3210395 and FONDECYT 11201250, Basal Funds for Center of Excellence FB210005 (CMM), Millennium Science Initiative Program ICN2021_004 (iHealth) and ICN17_002 (IMFD), and partly supported by the CLARA-MeD project (PID2020-116001RA-C33, funded by MICIU/AEI/10.13039/501100011033/ in call “Retos de investigación”).

Data availibility The datasets and normalization guidelines generated and/or analysed during the current study are available in the Zenodo repository: Báez P, Villena F, Rojas M, Bravo-Marquez F, Dunstan J. The Chilean Waiting List Corpus. 2020. https://zenodo.org/record/7555181 Báez P, Campillos-Llanos L, Dunstan J. The Chilean Waiting List sub-Corpus with medical entities normalized to UMLS terminology. 2023. https://zenodo.org/record/7757971. Arancibia, A. P., Chaparro, M. I., Bucarey, T. I., Báez P. (2024). Normalization Guidelines for Medical Entities of the Chilean Waiting List Corpus (1.2). https://doi.org/10.5281/zenodo.10995018

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethical approval The data analyzed in this study were requested through the Chilean Transparency Law, a nationwide initiative to improve access to data. The referrals are de-identified and constitute public information. Therefore no ethics committee approval was required.

References

Afzal, Z., Akhondi, S.A., van Haagen, H.H., Van Mulligen, E.M., & Kors, J.A. (2016). Concept recognition in french biomedical text using automatic translation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings 7*, pp. 162–173. Springer.

Báez, P., Bravo-Marquez, F., Dunstan, J., Rojas, M., & Villena, F. (2022). Automatic extraction of nested entities in clinical referrals in Spanish. *ACM Transactions on Computing for Healthcare (HEALTH), 3*(3), 1–22. https://doi.org/10.1145/3498324

Báez, P., Villena, F., Rojas, M., Durán, M., & Dunstan, J. (2020, November). The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, Online, pp. 291–300. Association for Computational Linguistics.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research 32*(suppl_1): D267–D270. https://doi.org/10.1093/nar/gkh061 .

Byrne, D., Lavelle, B., Jones, G.J., Smeaton, A.F. (2007). Visualising Bluetooth interactions: Combining the Arc Diagram and DocuBurst techniques.

Campillos-Llanos, L. (2023). Medlexsp - a medical lexicon for Spanish medical natural language processing. *Journal of Biomedical Semantics*. https://doi.org/10.1186/s13326-022-00281-5

Campillos-Llanos, L., Bouamor, D., Zweigenbaum, P., & Rosset, S. (2016). Managing linguistic and terminological variation in a medical dialogue system. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3167–3173.

Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A., & Moreno-Sandoval, A. (2021). A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making, 21*(1), 1–19. https://doi.org/10.1186/s12911-021-01395-z

Chen, K., Lin, X., Wang, H., Qiang, Y., Kong, J., Huang, R., Wang, H., & Liu, H. (2022). Visualizing the knowledge base and research hotspot of public health emergency management: A science mapping analysis-based study. *Sustainability, 14*(12), 7389. https://doi.org/10.3390/su14127389

Cortés, V. A., Barrera, F., & Nervi, F. (2020). Pathophysiological connections between gallstone disease, insulin resistance, and obesity. *Obesity Reviews, 21*(4), e12983. https://doi.org/10.1111/obr.12983

Dalianis, H. (2018). *Clinical text mining: Secondary use of electronic patient records*. Springer.

Donnelly, K., et al. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics, 121*, 279.

Dziadek, J., Henriksson, A., & Duneld, M. (2017). Improving terminology mapping in clinical text with context-sensitive spelling correction. *Informatics for Health: Connected Citizen-Led Wellness and Population Health, 235*, 241. https://doi.org/10.3233/978-1-61499-753-5-241

D'Souza, J., Ng, V. (2015). Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 297–302.

Fakhraei, S., Mathew, J., & Ambite, J.L. (2020). Nseen: Neural semantic embedding for entity normalization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 665–680. Springer.

Ferré, A., & Langlais, P. (2023). An analysis of entity normalization evaluation biases in specialized domains. *BMC bioinformatics, 24*(1), 227. https://doi.org/10.1186/s12859-023-05350-9

French, E., & McInnes, B. T. (2023). An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics, 137*, 104252. https://doi.org/10.1016/j.jbi.2022.104252

Gallego, F., López-García, G., Gasco-Sánchez, L., Krallinger, M., & Veredas, F.J. (2024). Clinlinker: Medical entity linking of clinical concept mentions in spanish. arXiv:2404.06367 .

García-Durán, A., Arora, A., & West, R. (2022). Efficient entity candidate generation for low-resource languages. arXiv:2206.15163 .

Ghiasvand, O., & Kate, R.J. (2014). UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns. In *SemEval@ COLING*, pp. 828–832.

Gianfrancesco, M. A., & Goldstein, N. D. (2021). A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Medical Research Methodology, 21*(1), 1–10. https://doi.org/10.1186/s12874-021-01416-5

Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., & Krallinger, M. (2019). Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pp. 1–10.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American medical informatics association, 12*(3), 296–98.

Ji, Z., Wei, Q., & Xu, H. (2020). Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings, 2020*, 269.

Kang, N., Singh, B., Afzal, Z., van Mulligen, E. M., & Kors, J. A. (2013). Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association, 20*(5), 876–881. https://doi.org/10.1136/amiajnl-2012-001173

Koneswarakantha, B. (2022). *easyalluvial: Generate Alluvial Plots with a Single Line of Code*. R package version 0.3.1.

Kors, J. A., Clematide, S., Akhondi, S. A., Van Mulligen, E. M., & Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition: The mantra GSC. *Journal* *of the American Medical Informatics Association, 22*(5), 948–956. https://doi.org/10.1093/jamia/ocv037

Kugic, A., Pfeifer, B., Schulz, S., & Kreuzthaler, M. (2023). Embedding-based terminology expansion via secondary use of large clinical real-world datasets. *Journal of Biomedical Informatics, 147*, 104497. https://doi.org/10.1016/j.jbi.2023.104497

Leaman, R., Khare, R., & Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics, 57*, 28–37. https://doi.org/10.1016/j.jbi.2015.07.010

Li, H., Chen, Q., Tang, B., Wang, X., Xu, H., Wang, B., & Huang, D. (2017). CNN-based ranking for biomedical entity normalization. *BMC bioinformatics, 18*(11), 79–86. https://doi.org/10.1186/s12859-017-1805-7

Lima-López, S., Farré-Maduell, E., Gascó, L., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., & Krallinger, M. (2023). Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. *Working Notes of CLEF* .

Lima-López, S., Farré-Maduell, E., Gasco-Sánchez, L., Rodríguez-Miret, J., & Krallinger, M. (2023). Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association, 88*(3), 265.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., & Collier, N. (2021). Self-Alignment Pretraining for Biomedical Entity Representations. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou (Eds.), *Proc. of the 2021 Conference of the NAACL*, pp. 4228–4238. https://aclanthology.org/2021.naacl-main.334

Liu, F., Vulić, I., Korhonen, A., Collier, N. (2021). Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. In *Proc. of the 59th ACL*, pp. 565—74. Association for Computational Linguistics. https://www.repository.cam.ac.uk/handle/1810/346234

Luo, Y., Song, G., Li, P., & Qi, Z. (2018). Multi-task medical concept normalization using multi-view convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32.

Magnini, B., Altuna, B., Lavelli, A., Speranza, M., & Zanoli, R. (2021). The E3C Project: European Clinical Case Corpus. *Language, 1*(L2), L3.

Magueresse, A., Carles, V., Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. arXiv:2006.07264 .

Margozzini, P. & Passi, Á. (2018). Encuesta Nacional de Salud, ENS 2016-2017: un aporte a la planificación sanitaria y políticas públicas en Chile. *ARS MEDICA Revista de Ciencias Médicas 43*(1): 30–34. https://doi.org/10.11565/arsmed.v43i1.1354 .

Marrone, M., Lemke, S., & Kolbe, L. M. (2022). Entity linking systems for literature reviews. *Scientometrics, 127*(7), 3857–3878. https://doi.org/10.1007/s11192-022-04423-5

McCray, A.T., Srinivasan, S., & Browne, A.C. (1994). Lexical methods for managing variation in biomedical terminologies. In *proceedings of the annual symposium on computer application in medical care*, pp. 235. American Medical Informatics Association.

Ministerio de Salud. (2019). Informe Encuesta Nacional de Salud 2016-2017: Uso de medicamentos.

Miranda-Escalada, A., Farré, E., & Krallinger, M. (2020). Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*.

Miranda-Escalada, A., Farré-Maduell, E., Lima-López, S., Estrada, D., Gascó, L., & Krallinger, M. (2022). Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of the LivingNER shared task and resources. *Procesamiento del Lenguaje Natural, 69*, 241–253.

Miranda-Escalada, A., Gascó, L., Lima-López, S., Farré-Maduell, E., Estrada, D., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., & Krallinger, M. (2022). Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., & Krallinger, M. (2020). Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020. *CLEF (Working Notes)* 2020 .

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: Opportunities and challenges. *Journal of biomedical semantics, 9*, 1–13.

Newman-Griffis, D., Divita, G., Desmet, B., Zirikly, A., Rosé, C. P., & Fosler-Lussier, E. (2021). Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets. *Journal of the American Medical Informatics Association, 28*(3), 516–532.

Noh, J. & Kavuluru, R. (2021). Joint learning for biomedical NER and entity normalization: encoding schemes, counterfactual examples, and zero-shot evaluation. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–10.

Oyarzún-González, X., Ferreccio, C., Abner, E. L., Vargas, C., Huidobro, A., & Toro, P. (2020). Polypharmacy in a semirural community in Chile: results from Maule Cohort. *Pharmacoepidemiology and drug safety, 29*(3), 306–315. https://doi.org/10.1002/pds.4941

Pérez, A., Atutxa, A., Casillas, A., Gojenola, K., & Sellart, Á. (2018). Inferred joint multigram models for medical term normalization according to ICD. *International journal of medical informatics, 110*, 111–117. https://doi.org/10.1016/j.ijmedinf.2017.12.007

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roham, M., Gabrielyan, A.R., & Archer, N. (2019). A systematic review of knowledge visualization approaches using big data methodology for clinical decision support. *Recent Advances in Digital System Diagnosis and Management of Healthcare*: 99–114 .

Ruas, P., & Couto, F. M. (2022). Nilinker: Attention-based approach to nil entity linking. *Journal of Biomedical Informatics, 132*, 104137.

Sanchez, G. (2018). *Arcdiagram: Plot pretty Arc diagrams*. R package version 0.1.12.

Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering, 27*(2), 443–460. https://doi.org/10.1109/TKDE.2014.2327028

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102–107.

Villena, F., Báez, P., Peñafiel, S., Rojas, M., Paredes, I., & Dunstan, J. (2021). Automatic support system for tumor coding in pathology reports in Spanish.

Wajsbürt, P., Sarfati, A., & Tannier, X. (2021). Medical concept normalization in French using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics, 114*, 103684. https://doi.org/10.1016/j.jbi.2021.103684

Wattenberg, M. (2002). Arc diagrams: Visualizing structure in strings. In *IEEE Symposium on Information Visualization (INFOVIS) 2002.*, pp. 110–116. IEEE.

West, V. L., Borland, D., & Hammond, W. E. (2015). Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association, 22*(2), 330–339. https://doi.org/10.1136/amiajnl-2014-002955

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software 4*(43): 1686. https://doi.org/10.21105/joss.01686 .

World Health Organization. (2004). *International Statistical Classification of Diseases and Related Health Problems vs. 10*. World Health Organization.

Xu, J., Lee, H.J., Ji, Z., Wang, J., Wei, Q., & Xu, H. (2017). UTH_CCB System for Adverse Drug Reaction Extraction from Drug Labels at TAC-ADR 2017. In *Proceedings of the Text Analysis Conference*.

Yuan, S., Gill, D., Giovannucci, E. L., & Larsson, S. C. (2022). Obesity, Type 2 Diabetes, Lifestyle Factors, and Risk of Gallstone Disease: A Mendelian Randomization Investigation. *Clinical Gastroenterology and Hepatology, 20*(3), e529–e537. https://doi.org/10.1016/j.cgh.2020.12.034

Zhu, T., Qin, Y., Feng, M., Chen, Q., Hu, B., & Xiang, Y. (2024). BioPRO: Context-Infused Prompt Learning for Biomedical Entity Linking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 374–385. https://doi.org/10.1109/TASLP.2023.3331149

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Pablo Báez[1,2] · Leonardo Campillos-Llanos[3] · Fredy Núñez[4,5] · Jocelyn Dunstan[5,6,7]

✉ Pablo Báez
pablobaez@ug.uchile.cl

Leonardo Campillos-Llanos
leonardo.campillos@csic.es

Fredy Núñez
frnunez@uc.cl

Jocelyn Dunstan
jdunstan@uc.cl

[1] Center of Medical Informatics and Telemedicine, Faculty of Medicine, University of Chile, Avda. Independencia 1027, Santiago 8380453, RM, Chile

[2] Tecnología Médica, Facultad de Medicina, Universidad del Desarrollo, Avda. Plaza 680, Las Condes 7610658, RM, Chile

[3] Instituto de Lengua, Literatura y Antropología (ILLA), CSIC (Spanish National Research Council), Albasanz 26-28, Madrid 28037, Spain

[4] Department of Language Sciences, Pontificia Universidad Católica de Chile, Santiago, RM, Chile

[5] Center for Mathematical Modeling (CNRS IRL 2807), University of Chile, Santiago, RM, Chile

[6] Department of Computer Science & Institute for Mathematical Computing, Pontificia Universidad Católica de Chile, Santiago, RM, Chile

[7] Millennium Institute for Foundational Research on Data (IMFD), Santiago, Chile