An Application of Natural Language Processing and Ontologies to Electronic Healthcare Records in the Field of Gynecology

Amanda Damasceno de Souza[1], Fernanda Farinelli[2], Eduardo Ribeiro Felipe[3], Armando Sérgio de Aguiar Filho[1] and Mauricio Barcellos Almeida [4]

[1]*FUMEC University, Graduate Program in Information and Communication Technology and Knowledge Management (PPGTICGC), Belo Horizonte, MG, Brazil.*
[2] *University of Brasília (UnB), Brasília, DF, Brazil*
[3] *Federal University of Itajubá (UNIFEI) Campus Itabira, MG, Brazil*
[4] *Federal University of Minas Gerais, Belo Horizonte, MG, Brazil*

Abstract

Electronic Health Records (EHR) usually comprise medical data sources containing unstructured data. EHRs contain various terms and idiosyncrasies, which prevent reasonable matches to standardized clinical terminologies. That, in turn, impedes information retrieval and the integration of systems of healthcare units, even systems within the same unit. The present article evaluates the application of Natural Language Processing (NLP) to EHR. The research presents a case study examining the connections among the EHR's terms for signs and symptoms, here called the *interface terminology*; a biomedical ontology, here called the *reference terminology*; and the Tenth International Classification of Diseases (ICD-10), here called the *aggregation terminology*. We collected a sample of terms for signs and symptoms in gynecology to test correlations between reference and aggregation terminologies. We report and analyze the main difficulties we encountered during the correlation process regarding the semantics of the terms and the lack of related terms.

Keywords [1]

Electronic health records, clinical terminology, natural language processing, biomedical ontologies.

1. Introduction

Electronic Health Records (EHR) are an essential source of real-world health information for several purposes. Information in EHRs is often recorded in an unstructured format, which poses challenges to using it for computational purposes. Indeed, advances in health information technologies have followed an increasing need for standardized clinical text and terminologies to facilitate information retrieval (IR) and interoperability. Usually, unstructured EHR data have a terminological variety that does not match standardized clinical terminologies, which poses a significant obstacle to achieving IR's objectives [1]. Therefore, an effective means of connecting the ordinary terms found in EHRs with standard medical terminologies could improve IR processes. One option is to map the EHR's terms onto standardized terminologies.

Health terminology standardization is a requirement for achieving effective IR. Structured and controlled data representation is essential when using a terminological system to record medical data. The terminological system consists of techniques and artifacts such as thesauri, controlled vocabularies, taxonomies, and ontologies [2]. Standardized biomedical terminologies are essential because they

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: amanda.dsouza@fumec.br; amandasd81@gmail.com (A. 1); fernanda.farinelli@unb.br (A. 2); eduardo.felipe@unifei.edu.br (A. 3), armando.filho@fumec.br (A. 4). mba@eci.ufmg.br (A. 5) ORCID: 0000-0001-6859-4333 (A. 1); 0000-0003-2338-8872 (A. 2); 0000-0003-1690-2044 (A. 3), 000-0001-5542-7165 (A. 4), 0000-0002-4711-270X (A. 5).

interface clinical data and health care systems, including the EHRs [3]. Standardized terminologies are also valuable resources for enabling interoperability in EHR by collaborating to perform audits, research, benchmarking, and management for hospitals [4].

Our investigation draws on existing literature, such as a study by Schulz et al. [5], who analyze terminology standardization and propose a methodology to connect three types of health terminologies: *interface terminologies*, namely, medical chart text or medical jargon; *reference terminologies,* which are controlled vocabularies and ontologies; and *aggregation terminologies*, which include the International Classification of Diseases (ICD), Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) and others. Our research adopts the denominations employed by Schulz et al. [5]. In this context, research by Rector [6] raises some highly relevant questions.

The gap posed by Schulz et al. [5] requires finding a way to connect the clinical data in an EHR's clinical texts to standardized clinical terminologies, including the ICD, SNOMED-CT, Medical Subject Headings (MeSH), Unified Medical Language Systems (UMLS), and biomedical ontologies such as those found on the OBO Foundry portal. Although Schulz et al.[5] connected three standardized medical terminologies, they didn’t connect any of those to the ordinary language terms found in EHRs. So, significant work remains to be done.

Interoperability among clinical terminologies promotes the generation of innovative products that helps physician better annotate EHRs, contributing to the quality of care and patient well-being. Our research examines a case study about the connections among the terms for signs and symptoms used in the patient’s EHR, a biomedical ontology, and the ICD-10. As its principal contribution, our research verified medical jargon terms that do not correspond to existing biomedical ontologies in the OBO Foundry or OntONeo. As a further contribution, we use OntONeo to connect an EHR’s textual clinical data with the standardized clinical terminologies, which Schulz et al. [5] call *reference terminology*.

2. Methodology

Our interdisciplinary study involves Librarianship and Information Science (LIS), Information Technology, and healthcare fields. We conducted applied research using qualitative, quantitative, and descriptive methods. We followed the tenets of those mentioned above, well-established researchers to standardize biomedical terminologies by adopting three designations: i) interface terminologies, which stand for ordinary language texts recorded in EHRs; ii) reference terminologies, which are ontologies and controlled vocabularies; iii) aggregation terminologies, which are ICD-10 artifacts [5]. Then, we applied natural language processing (NLP) techniques and domain ontologies, specifically OntONeo [7]. Our methodology relied on NLP to extract and analyze signs and symptoms from clinical texts, ultimately connecting them to the standards by mapping them through ontology. We performed the usual pre-processing preparation stages of the free text, including treatment of stop-words, and case-folding techniques, excluding break-lines. In the information-extraction step, we developed specific algorithms to locate signs and symptoms and compare them to a list of signs and symptoms previously prepared by domain experts seeking to improve the automatic task of term identification. The information extraction was performed in a large private hospital, which provided a sample of 32,291 real EHRs containing medical notes in free text. These groups of notes cover the evolution and medical history of patients from the gynecology department during the year 2018, and their use was authorized through the appropriate administrative and ethical processes. [8]

The medical team created a pre-list of signs and symptoms to delimit the algorithm for data processing. Other sources of information used in the pre-list of signs and symptoms were the National Library of Medicine (NLM) Classification 2020 Summer Edition [9], Wikipedia [10,11], Falcão Junior et al. [12], and ICD-10. For the pre-list of signs and symptoms, it was necessary to include data on the following systems: circulatory and respiratory; digestive and abdomen; skin and subcutaneous tissue; nervous and musculoskeletal; and urinary. The pre-list also included terms about cognition, perception, emotional state and behavior, speech and voice, and general signs and symptoms. This pre-list was validated by a gynecologist, i.e., a domain expert.

The next step was determining the most frequent signs and symptoms in the general population and their quantity in the EHRs. This list of signs and symptoms was created in a text file, which was, in turn, read by the algorithm to create a list (array) of terms found. In the database, the result of this reading was segmented according to the type of analysis ("anamnesis" and "evolution"). Therefore, in each record whose information was extracted from the hospital institution, the correspondence between those signs and symptoms (already available in the list in memory) that appeared was traced. A data structure was organized by a pair key, namely, value, called a *dictionary* in Python programming language. This model allows the storage of the ICD code (key) and the identification of its quantity (value). This data structure was later recorded in a spreadsheet format file.

The last step was to check the frequency of the interface terminology and its proper correspondence to the reference terminology. This analysis step was performed by a medical expert specializing in gynecology. After mapping the terminologies, the number of terms present in the interface terminology and reference terminology was quantified to verify the percentage of connectivity (match) between the clinical terminologies. Finally, the results were described for their respective groups.

2.1 Mappings between Terminologies

In mapping the interface terminology onto the reference and aggregation terminologies, the ABNT ISO/TR 12300 standard was taken as the base [13]. The steps for mapping were as follows:

1) Document the mapping process between clinical terminologies (Table 1).

2) Verify the semantic equivalence between terms (Table 1).

3) Utilize a source mapping for terms with multiple synonyms (Table 1).

4) Analyze risk factors and document ways to ensure consistency in mapping.

5) Clarify the meaning and fully use the form for abbreviations in the interface terminology.

6) Map the target terms of the reference terminology selected from Health Science Descriptors (DeCS)[2][14], created by The Latin American and Caribbean Center on Health Sciences Information[3]. Such terminology was developed from Medical Subject Headings (MeSH) [15], and OntONeo as the reference terminology belongs to the OBO-Foundry and aligns with principles of good practices in developing ontologies. Also, map the ICD-10 as the aggregation terminology since this is the classification used in the hospital institution whose data supported this research (Table 2).

7) Create a mapping table to demonstrate the types of interoperability verification: interoperate one term for one, interoperate one term for many terms, interoperate many terms for one term, interoperate many terms for many terms, and do not interoperate (Table 2).

It should be noted that the corpus of unstructured medical data used in the study was created in Portuguese, so the controlled vocabulary used was DeCS. It is a multilingual thesaurus that "[…] to serve as a unique language in indexing articles from scientific journals, books, congress proceedings, technical reports, and other types of materials, as well as for searching and retrieving subjects from scientific literature from information sources available on the Virtual Health Library (VHL) such as LILACS, MEDLINE, and others".[14] DeCS is a translation of MeSH [15] into Portuguese, also

[2] In Portuguese: Descritores em Ciências da Saúde. Available on the internet in: https://decs.bvsalud.org/ Access Jun. 01 2023

[3] In Portuguese: BIREME. Available on the internet in: https://www.paho.org/en/bireme. Access Jun. 01 2023.

providing terms in Spanish and French. Therefore, the research also registered the controlled vocabulary terms in English, i.e., the original version from MeSH, for publication in this language.

Table 1
Preliminary Steps for Mapping Clinical Terminologies

Terminology	Mapping	Terminology	Support (source mapping)
Interface terminology	- Check diagnostic terms, signs and symptoms - Anamnesis/Evolution of Gynecology	Anamnesis and Evolution of Gynecology	-Gynecology Anamnesis Books/ Gynecology and Obstetrics Guidelines-Wikipedia. -Domain expert
Reference terminology	- Check which are and quantity of diagnostic classes, signs and symptoms of Gynecology.	*OntONeo*	-DeCS/MeSH
Aggregation terminology	-Check which are and quantity of classifications for diagnosis, signs and symptoms of Gynecology.	International Classification of Diseases - ICD-10	-Domain expert

Fonte: [8].

Table 2
Mapping of Terms

Mapping	Relation	Final decision
Interoperate one term for one	A single source class is linked to a single target class or term	Retain
Interoperate one term for many terms	A single source class is linked to multiple target classes or terms	Define a class according to basic formal ontology (BFO) and choose term that poses no clinical risk
Interoperate many terms for one term	Multiple source classes are linked to a single target class or term	Define a class according to BFO and choose term that poses no clinical risk
Interoperate many terms for many terms	Multiple source classes are linked to multiple target classes or terms	Define a class according to BFO and choose a term that poses no clinical risk

Source: [8], [16].

3. Results

The first part of the results presents the frequency of terms found in the free-text fields of the EHR. We retrieved approximately 80 types of signs and symptoms in addition to stop-words, abbreviations, and negation expressions, which revealed the complex challenges of planning any automatic initiative. (Table 3). The principal signs and symptoms found refer to frequent complaints in gynecology: pain (n=3671); bleeding (n=2889); edema (n=800); pruritus (n=757); and discharge (n=664).

Table 3
Examples of Signs and Symptoms in Interface Terminology

Terms	Absolute Frequency (n)
Pain	3671
Bleeding	2889
Edema	800
Itching	757
Discharge	664
Dysmenorrhea	456
Vomiting	398
Nausea	336
Abdominal pain	318
Fever	308
Nausea	305
Pelvic pain	298
Tension	219
Metrorrhagia	182
Abnormal uterine bleeding	169
Heartburn	165
Atrophy	163
Headache	154
Coma	147
Depression	133
Urinary incontinence	132
Anxiety	122
Vomiting	119
Pelvic pain	110

Source: [8].

For interface terminologies, we surveyed DeCS[14] to check definitions and synonyms, following methodological step 3 (use a source mapping for terms with multiple synonyms). Then, we compared the correlated terms found with both tables of signs and symptoms of ICD-10 [17] and OntONeo [7]. By methodological steps 6 and 7, we then mapped the target terms of the reference terminology (selected from DeCS/MeSH and OntONeo as the reference terminology [...]) and created a mapping table to demonstrate the types of interoperability verification[...]) displayed in Table 1; the results are presented in Table 4.

Selected examples demonstrate the correspondence between the clinical terminologies. We verified that for signs and symptoms frequently reported in gynecological consultations, there was no correspondence between the term from the interface terminology, e.g., "itching," and that in the reference terminologies. Another example of signs and symptoms frequently reported in gynecological consultations, there was no correspondence between the term from the interface terminology, e.g., "Irregular menstrual cycle," and that in the aggregation terminologies.

The term was present only in the DeCS/MeSH-controlled vocabulary. The term "irregular menstrual cycle" did not match the clustering terminology. Only the term "dysmenorrhea" found a match in the three types of clinical terminologies, i.e., interface (EHRs); reference (OntONeo and DeCS/MeSH); and aggregation (ICD-10). Table 4 shows no correspondence between the EHRs' terms and ICD-10; similarly, the EHRs' terms did not correspond to OntONeo. The interface terminology terms that were not matched in the reference terminology, OntONeo, will be added to this ontology. Language variations will be added to the ontology's enrichment, specifically in synonyms.

Table 4
Examples of correlated terms found compared with signs and symptoms of OntoNeo, DeCS/MeSH, and ICD-10 [8].

EHRs	OntONeo	DeCS/MeSH	ICD-10
Irregular menstrual cycle	Process - biological_process - reproductive process - single organism reproductive process - ovulation cycle - menstrual cycle - Quality - Phenotypic abnormality - Abnormal genital system morphology - Abnormality of the menstrual cycle	Menstrual cycle	–
Itching	–	Pruritus	L29.0 Pruritus ani L29.2 Pruritus vulvae L29.3 Anogenital pruritus, unspecified L29.8 Other pruritus L29.9 Pruritus, unspecified Itch NOS
Dysmenorrhea	- Quality - information carrier-sintoma - nervous system symptom - sensation perception - pain	Dysmenorrhea	R10 Abdominal and pelvic pain R10.1 Pain localized to upper abdomen
Painful urination	- Quality - information carrier-sintoma - nervous system symptom - sensation perception - pain - renal colic	–	R30 Pain associated with micturition

Source: [8].
Note: The dash (–) signifies the absence of terms.

The second part of the results reports the mapping among the terms. As seen in Table 5, when applying the mapping according to the ABNT ISO/TR 12300 Standard [13], between interface terminology for reference terminology (OntONeo), 60.15% (n=80) of the signs and symptoms do not interoperate. The second most frequent mapping type was *interoperated one term for one term*.

Table 5
Mapping Interface Terminology Terms to the Reference Terminology (OntONeo)

	Signs and Symptoms	
Interoperability	**n**	**%**
Interoperate one term for one	27	20,30
Interoperate one term for many terms	5	3,76
Interoperate many terms for one term	18	13,53
Interoperate many terms for many terms	3	2,26
Non-interoperable	**80**	**60,15**
Total	133	100

Source: (8).

In Table 6, when applying the mapping according to the ABNT ISO/TR 12300 Standard [13], between interface terminology to aggregation terminology (ICD-10), it can be seen that 53.15 % (n=76) of the signs and symptoms do not interoperate.

Table 6
Mapping Interface Terminology Terms to Aggregation Terminology (ICD)

Interoperability	**Signs and Symptoms**	
	n	**%**
Interoperate one term for one	43	30,07
Interoperate one term for many terms	13	9,09
Interoperate many terms for one term	6	4,20
Interoperate many terms for many terms	5	3,50
Non-interoperable	**76**	**53,15**
Total	**143**	**100**

Source: [8].

4. Discussion

Some aspects of the results presented so far are worth stressing and discussing. For example, Table 3 indicated that the term "irregular menstrual cycle" is correlated to the OntoNeo Ontology and DeCS/MeSH terms but did not show a corresponding term in the ICD-10. The term "itching" is absent in the ontology. "Dysmenorrhea" is already included in the three terminologies. The last example, "painful urination," appears in the ontology and the ICD-10. Table 2 shows the semantic variety to represent signs and symptoms in clinical terminology and the absence of terms in these instruments. Applying the matching between interface terminology and the reference terminology (OntONeo) indicates that 60.15% of the signs and symptoms do not interoperate.

In matching terms in the interface terminology to those in the reference terminology for OntONeo classes, we mapped multiple interface terminology source classes to multiple classes or target terms in the ontology. Defining a single class according to the BFO was necessary to avoid multiple inheritances. We performed the same procedure for a single source class in the interface terminology, which we mapped to multiple classes or target terms in the reference terminology (OntONeo ontology). In the case of multiple interface terminology source classes, we mapped to a single ontology target class or term. The excess terms were used to enrich the OntONeo synonym class.

In mapping terms from the interface terminology to terms in the aggregation terminology (ICD), we found that the type "does not interoperate" stood out, and signs and symptoms were absent in 53.15% (Table 6). It is worth noting that the mapping of "interoperates many terms for many terms" obtained an equivalence of 3.50% of the signs and symptoms. A significant absence of interface terms was detected in the aggregation terminology (ICD-10), demonstrating the need to review and update this artifact for better application in the medical profession’s clinical practice.

Schulz et al. [5] note the difficulty in reconciling interface terminologies, reference terminologies (e.g., SNOMED CT), and aggregation terminologies (e.g., ICD-11), tying that difficulty to the distinct functions of each terminology. Such difficulties were demonstrated in this research through the percentages of terms that did not interoperate with each other in clinical terminologies: 60.15% of signs and symptoms between interface terminology and reference terminology (OntONeo), and 53.15% of signs and symptoms did not interoperate in the mapping step between interface terminology and aggregation terminology (ICD-10).

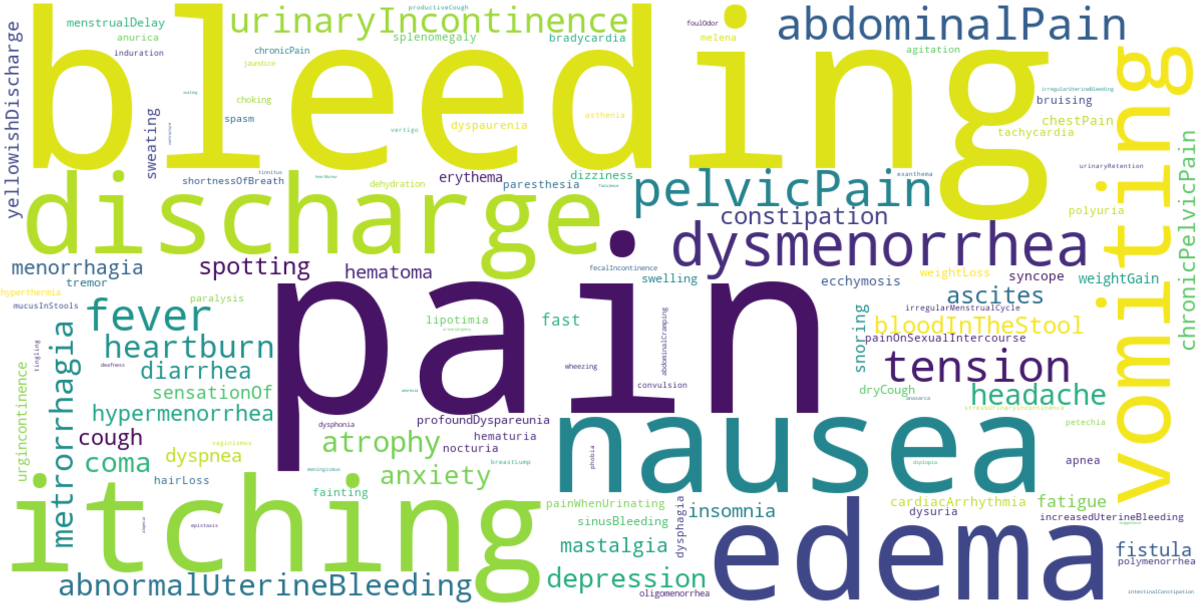

Figure 1: Word Cloud of Most Frequent Signs and Symptoms.
Source: Souza [8].

The frequencies or percentages between mappings indicate that interface terminology is more distant from reference terminology than aggregation terminology. This is explained by physicians' greater familiarity with the aggregation terminology than with the reference terminology; consequently, the terms used in reporting the open fields of the EHR resemble the terms in ICD more than those in OntONeo[7]. Terms in the interface terminology tended to be absent from the aggregation and reference terminologies, demonstrating that interface terminology has richly diverse terms. Notably, the sample used in this research was satisfactory; the richness of its terminology, as shown in Figure 1, enabled it to contribute substantially to the OntONeo ontology and other biomedical ontologies.

5. Final Considerations[4]

Having modified the second step of the proposal by Schulz et al. [5], we performed the connections (mappings) for this research in two steps: first, we mapped interface terminologies to reference terminologies, and subsequently, we mapped the interface terminologies to aggregation terminologies. Instead of the reconciliation step between reference and aggregation terminologies, we mapped interface terminologies to aggregation terminologies. This modification was necessary because we focused on analyzing the mappings between interface terminology and clinical terminologies.

The medical jargon (interface terminology) used in clinical practice proved to be different and distant from standardized terminologies such as ontologies (reference terminologies) and even from ICD-10 (aggregation terminology). This research described some differences in syntax and semantics that posed obstacles to achieving interoperability between information health systems. To reduce these differences, we propose using existing knowledge representation resources in the information science field and the assistance of clinical librarians.

We identified several issues with spelling, punctuation, and typographical errors in the analyzed text. We realized the difficulties in applying NLP techniques to real-world texts and foresaw that ontology could reduce the peculiarity of human notes, helping to achieve the goal of harmonization. As

[4] **Funding:** No grant supported this research
Declarations Ethics approval: The study was approved by the local institutional review board (CAAE:03384418.0.0000.51259).
Competing interests: The authors declare that they have no competing interests.

an additional contribution, we created a computational lexicon (corpus in healthcare) in Portuguese, which can help create algorithms for the domain of gynecology.

One of the main aspects explored in the research was the issue of semantics and syntax of the terms. In this, we aimed to address a primary difficulty in analyzing the medical jargon used in interface terminology, namely, its epistemological aspects, which depend heavily on the medical context. Thus, ontology is an artifact that should be used in seeking a solution to this difficulty.

6. References

[1] S. W. Smith, R. Koppel. Healthcare information technology's relativity problems: A typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ. J Am Med Inform Assoc. 21(2014):117-31. doi: 10.1136/amiajnl-2012-001419.

[2] N. F. de Keizer, A. Abu-Hanna, J.H. Zwetsloot-Schonk. Understanding terminological systems. I: Terminology and typology. Methods Inf Med. 39 (2000):16-21.

[3] J. Rogers. Using Medical Terminologies. (2005). Available from: http://www.cs.man.ac.uk/~jeremy/HealthInf/RCSEd/terminology using. Htm. Accessed on: 05 Mar. 2019.

[4] J. A. Miñarro-Giménez, R. Cornet, M. C. Jaulent, H. Dewenter, S. Thun, K. R. Gøeg, D. Karlsson, and S. Schulz. Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform. 123 (2019):37-48. doi: 10.1016/j.ijmedinf.2018.12.011.

[5] S. Schulz, J. M. Rodrigues, A. Rector, C. G. Chute. Interface Terminologies, Reference Terminologies, and Aggregation Terminologies: A Strategy for Better Integration. Stud Health Technol Inform, 245(2017):940-944.

[6] A. L. Rector. Clinical Terminology: Why is it so Hard? Methods of Information in Medicine, Stuttgart, 38:147-157, 1999.

[7] F. Farinelli. et al. OntONeo: The Obstetric and Neonatal Ontology. In: Conference: International Conference on Biomedical Ontology 2016, ICBO, Corvallis, Oregon, USA, August 2016. Available at: https://www.researchgate.net/publication/304254064_OntONeo_The_Obstetric_and_Neonatal_Ontology.

[8] A. D. Souza. Clinical Practice Discourse and Standardization Terminologies: Investigating the Connection. Pós-Graduação em Gestão e Organização do Conhecimento, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte. 2021. [Portuguese]. Available at: http://hdl.handle.net/1843/38044. Accessed on: 03 Jul. 2023.

[9] S. R. Willis. NLM Classification 2019 Summer Edition Now Available. The NLM Technical Bulletin, Bethesda, n.430, e4, Sep-Oct 2019. Available at: https://www.nlm.nih.gov/class/index.html. Accessed on: Jul. 21, 2020.

[10] Wikipédia, a enciclopédia livre. Lista de sintomas médicos. Esta página foi editada pela última vez às 04h13min de 13 de janeiro de 2019b. [Portuguese].Available at: https://pt.wikipedia.org/wiki/Lista_de_sintomas_m%C3%A9dicos.

[11] Wikipédia, a enciclopédia livre. Sinal médico. Esta página foi editada pela última vez às 18h56min de 17 de agosto de 2018.[Portuguese] Available at: https://pt.wikipedia.org/wiki/Sinal_m%C3%A9dico.

[12] J. O. A. Falcão Júnior *et al.* Ginecologia e obstetrícia: assistência primária e saúde da família. Rio de Janeiro: MedBook, 2017.[Portuguese]

[13] Associação Brasileira de Normas Técnicas. Relatório técnico ISO/TR 12300: Informática em saúde – princípios de mapeamento entre sistemas terminológicos. 28.11.2016. Rio de Janeiro: ABNT, 2016.[Portuguese]

[14] Health Sciences Descriptors: DeCS [Internet]. 2017 ed. São Paulo (SP): BIREME / PAHO / WHO. 2017 [updated 2017 May; cited 2017 Jun 13]. Available at: http://decs.bvsalud.org/I/homepagei.htm

[15] National Center for Biotechnology Information, U.S. National Library of Medicine. (2022). MeSH (Medical Subject Headings). Bethesda: NLM. Available at: https://www.ncbi.nlm.nih.gov/mesh/.

[16] R. Arp, B. Smith, and A. D. Spear. Building Ontologies with Basic Formal Ontology. Cambridge, MA: The MIT Press, 2015.

[17] ICD-10. (2022). Symptoms and signs involving the genitourinary system R30-R39. Available at: https://www.icd10data.com/ICD10CM/Codes/R00-R99/R30-R39.