From ANVISA leaflets to extended interoperability with global health databases: some pitfalls and success stories

Márcia Jacobina Andrade Martins[1], **Claudia Bauzer Medeiros**[1]

[1]Institute of Computing – University of (UNICAMP) Campinas – SP – Brazil

m905106@dac.unicamp.br, cmbm@unicamp.br

Abstract. *Data interoperability in Health Information Systems (HIS) has long been recognized as a challenge, sometimes even within a single institution, given the number of different databases, systems, standards and requirements adopted. In Brazil, this is aggravated by the lack of standardized, consensual data sources in Portuguese. This paper describes a hands-on approach to overcome these hurdles, thereby helping researchers and practitioners interested in these issues. It shows how, starting from official sites of the Brazilian Health Ministry, we built HealDB – an open Portuguese language database with hundreds of thousands of instances. HealDB supports interoperability across multiple international data sources, providing a core for the construction of federated HIS in Brazil. Within this context, the paper contains two main contributions: (1) a discussion of successive approaches to derive disease ICD-10 codes from text in drug leaflets, identifying their pros and cons; and (2) a case study of linkage to RxNorm, a normalized naming system for clinical drugs maintained by the U.S. National Library of Medicine, thereby illustrating potential extensions.*

1. Introduction

Health Information Systems (HIS) are systems that manage healthcare data, with the primary objective of supporting patient care [Haux 2006], in addition to the associated administrative and managerial tasks. Their proliferation in medical facilities, and dependence on customized tools and data sources, present a major interoperability challenge – even when we restrict ourselves to textual data. Moreover, a large portion of these sources concern private patient data, and thus, subject to data protection laws. Finally, most standardized data sources are in English, requiring adaptation to use in environments that require other languages.

To overcome some of these issues, we have designed and implemented the HealDB database [Martins and Medeiros 2025]. The main objective of constructing HealDB is to provide a curated and interoperable health database to support integration with HIS, thereby benefiting healthcare professionals and researchers. Its core tables, in Portuguese, are based on processing and curating data from all drug leaflets and medications authorized by the ANVISA[1] system. Its data include medications, their active ingredients, therapeutic indications, and drug and food interactions. Additional HealDB data involve symptoms, drug composition, and other related data. This basic curated data infrastructure aims to enable linkage to arbitrary external biomedical data sources, thereby helping researchers and practitioners working in the Brazilian health scenario.

[1]The Brazilian Health Regulatory Agency – https://www.gov.br/anvisa

The construction of HealDB went through several phases, encompassing almost 2 years. The first version appeared in a short paper in SBBD 2023 [Martins and Medeiros 2023]; it was limited to a few thousand records and data on diseases, symptoms, and medications. This was subsequently extended in June 2024 to hundreds of thousands of instances, and additional tables on drug-drug and drug-food interactions [Martins and Medeiros 2024]. Since then, we have performed major curation and extension tasks, which among others extend its interoperability to major external biomedical data sources.

In this context, the paper's main contributions are: (1) the discussion of our several approaches to identify disease ICD-10 codes from descriptive text in drug leaflets – see Section 4; and (2) an interoperability case study, which discusses how we included support for integration with RxNorm[2], a normalized naming system for clinical drugs created by the U.S. National Library of Medicine – see Section 5.

2. Related Work

The term "interoperability" has become semantically charged. While originally conceived to refer to software systems, it is now also used to refer to the "interaction" among data sources. The survey on software systems' interoperability by [Maciel et al. 2024] proposes a classification of 36 types of interoperability. Our approach falls into their classes "technical", "semantic" and, "interface".

One of the limitations of healthcare interoperability in Brazil is the lack of structured biomedical data in Portuguese. HealDB addresses this challenge by extracting and standardizing medical information from Portuguese-language official data sources from the Brazilian Health Ministry, in addition to enabling interoperability with international terminologies. Initiatives such as LeME-PT [Simões and Gamallo 2021] have helped to build a corpus of annotated drug leaflets in European Portuguese. Although distinct in scope, this research shares with HealDB a focus on extracting standardized information from domain-specific texts in Portuguese.

There are relatively few papers on non-English NLP for biomedicine and clinical studies. [Shaitarova et al. 2023] provide a review of biomedical and clinical Natural Language Processing (NLP) studies from 2020 to 2022 focusing on *low resource languages*, covering topics such as Named Entity Recognition (NER), large language-specific and multilingual Pretrained Language Models, and entity linking. It includes some research in Portuguese, but highlights the impressive advances made in Spanish.

There are two approaches to handling low-resource languages: language-specific models and multilingual models. The work presented in [Sousa et al. 2023] identifies clinical entities in oncology texts written in European Portuguese. The methodology used consists of automatic annotation, curation by medical experts, and training of NER models using a specific BERT-based model. [Sallauka et al. 2025] exemplifies a multilingual model – they developed a weakly supervised multilingual NER pipeline to extract symptoms, tests, and treatments from patient-reported outcomes and other medical texts. Besides Portuguese, it processes seven other languages, using a BERT-based model.

Another relevant aspect of interoperability concerns the use of structured clini-

[2]https://www.nlm.nih.gov/research/umls/rxnorm/index.html

cal data in Electronic Health Records (EHRs). Depending on the research scope, studies may use data directly extracted from EHRs or from information derived from them, e.g., [Sohn and Liu 2014]. While EHRs provides access to many kinds of clinical data and patient profiles, they are seldom standardized and, for ethical reasons, closed to access. Many research efforts take advantage, instead, of MIMIC data[3] a publicly available database of deidentified real patient data from intensive care units. HealDB can interoperate with any EHR that contains ICD-10 identifiers (and thus MIMIC), as well as with those that contain Brazilian-approved medications or standardized drug names or symptoms in Portuguese.

3. An overview of HealDB

HealDB is an open relational database whose core was built using open data obtained from primary Brazilian official data sources – among others, drug leaflets and medications in Portuguese, provided by ANVISA[4] . Figure 1 shows the evolution of this development, starting from ANVISA data. It does not represent the organization of the database, nor its data model – it serves to illustrate how the core concepts originated.

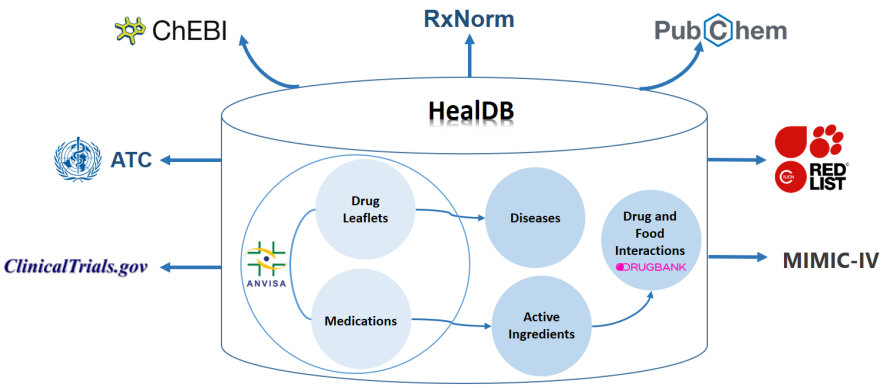

Figure 1. HealDB – a schematic illustration.

Descriptive text in leaflets was used to identify the diseases treated by each medication, leading to standardized ICD-10 codes – see Section 4, which describes our disease identification alternatives and our final implementation choices.

Each medication is composed of one or more active ingredients. We used them, in turn, to identify hundreds of thousands of drug-drug and drug-food interactions. Interactions were elicited via extensive work of cross-checking with DrugBank[5], a public database on drugs, their mechanisms, interactions, and targets. This was described in [Martins and Medeiros 2024] in June 2024.

Since then, we have performed major curation and extension tasks, which among others included providing HealDB with additional identifiers and tables that allow it to be linked to several major external health-related data sources, as well as performed several competency queries to validate it – e.g., ChEBI[6], or RxNorm – see Section 5.

[3]https://physionet.org/content/mimiciii/1.4/

[4]https://dados.gov.br/dados/conjuntos-dados/medicamentos-registrados-no-brasil and https://consultas.anvisa.gov.br/

[5]https://go.drugbank.com/

[6]https://www.ebi.ac.uk/chebi/

4. Methodology for Disease Extraction from Drug Leaflets

The structured association between medications and diseases is essential for the integration of clinical components of HIS (e.g., prescription and diagnosis), as well as for medical decision support and the automation of analyses. In Brazil, this relationship is not included in structured public databases, appearing only in free text in ANVISA's drug leaflets. Its standardization allows interoperability with electronic medical records, the identification of therapeutic inconsistencies, and integration with international vocabularies.

4.1. Alternative Approaches for Disease Extraction

Drug leaflets contain free text descriptions of diseases for which the drug is indicated, under sections "what this medication is indicated for" and "how this medication works". Disease extraction corresponds to the following text matching problem – *given the free text of drug leaflet that describes a set of diseases and symptoms, what are the associated ICD-10 codes?* Though seemingly straightforward, this is not a simple task. Among many challenges, there is no standard vocabulary to textually describe diseases or symptoms – and thus, a problem of Natural Language Processing.

We went through four distinct approaches for this matching task, looking for both matching accuracy and less dependence on human intervention. This section describes the different procedures we tried; we include this to help researchers and practitioners who attempt similar tasks to better understand some alternatives. Our goal here was not to compare models, but to adopt a solution that would require less expert annotation effort.

Our initial approach for this extraction, described in [Martins and Medeiros 2023], used the SpaCy NLP library. Though adequate for a first validation, it proved to be unsatisfactory for promoting overall interoperability. Of ANVISA's total of 7,476 valid leaflets, diseases and symptoms were identified in only 57%. Thus, we searched for alternative extraction strategies. We first considered using a NER model, and perform the fine-tuning of a model pretrained in SpaCy. However, we were unable to find a pretrained SpaCy model specialized in medical terms and available in Portuguese, thus we discarded this solution.

Next, we considered disease identification using BioBERTpt [Schneider et al. 2020], a BERT-based language model pretrained on Portuguese biomedical corpora. Here, we would also need to perform fine-tuning, annotating approximately 1,200 leaflets.

Still looking for a less human time-consuming alternative, we tested disease extraction with both the GPT-3.5 and GPT-4.0 models. The accuracy of GPT-4.0, as expected, was superior, returning a more precise range of ICD codes from text. Its extraction resulted in 42,031 records, of which 26,829 were diseases and 15,160 were symptoms. However, validation with help of experts revealed many inconsistencies in the ICD codes, and thus the need for improving our inputs (and, again, need for experts).

Finally, we tested Amazon Comprehend Medical [Bhatia et al. 2019], a web service provided by Amazon, specialized in NLP and pretrained to extract health information from text. Unlike models behind GPT and similar, it does not generate text, and cannot be trained — rather, it is a service that identifies health-related entities (diseases, symptoms, etc.), and classifies them according to standardized vocabularies (e.g., ICD-10-CM

or SNOMED-CT). We used its *InferICD10CM* API[7], which identifies medical entities and their ICD-10 codes, and found it adequate for our needs – namely, not requiring help from experts for fine-tuning. Given its robustness, and "domain-expertise", we incorporated this solution into our disease extraction methodology, as described below.

4.2. Disease Extraction Workflow

Figure 2 shows the workflow for extracting standardized codes from drug leaflets. In the first step, "Extract and Translate Leaflet Sections", we retrieved the full text of each drug leaflet stored in HealDB, identifying and extracting two specific sections, namely "what this medication is indicated for" and "how this medication works", to be processed by Amazon Comprehend Medical. However, since Amazon Comprehend Medical only works with texts in English, and our leaflets are in Portuguese, it was necessary to translate these sections into English using GPT-4.0 (via API)

In the second step, the translated texts were submitted as input to Amazon Comprehend Medical. For each leaflet text, it returns a JSON "table" identifying possible diseases and symptoms and their respective ICD-10-CM codes, together with three different scores indicating the confidence the system has in the result (overall score, entity score, and trait score). Overall scores indicate confidence in the relationship between code and disease descriptive text; entity scores indicate confidence in identifying the ICD code; and trait scores indicate confidence that the text has a given trait - e.g., it is a diagnosis, a hypothesis, a symptom, or a negation.

In the final step, "Link Medications to Diseases", we applied a multi-criteria strategy to analyze each result, and selected the association "descriptive text - ICD code" that attained our highest confidence score. These codes were then linked to medications and stored in HealDB. [8].

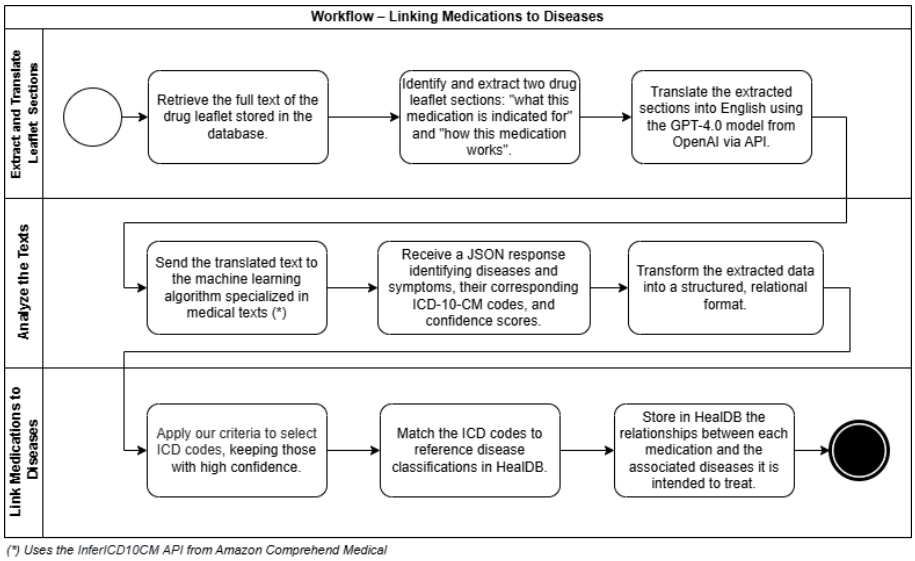

Figure 2. Workflow - Linking Medications to Diseases.

[7]https://docs.aws.amazon.com/comprehend-medical/latest/api/API_InferICD10CM.html

[8]Due to size limits, we cannot explain these scores, nor how we considered their combination to select the most appropriate ICD-10 code.

A total of 7,476 drug leaflets went through the disease extraction process. Of these, 5,808 contained valid disease mentions according to our criteria. Overall, our disease identification process produced 43,194 valid pairs of medication-disease associations, covering 2,404 distinct ICD-10 codes.

4.3. Some results and Insights

HealDB data enable several kinds of analyses. Figure 3 shows one such example: a result of checking the association between medications and diseases. It shows the top 15 ICD-10 categories associated with the highest number of ANVISA-approved medications (out of the 2,404 ICD codes we identified). As we can see, depressive episodes, epilepsy, and glaucoma (respectively ICD-10 categories F32, G40, and H40) each have indications for approximately 650 different authorized medications, while acute sinusitis ranks 15th in number of different medications. This may suggest that certain diseases have a larger variety of approved potential treatment alternatives in Brazil; alternatively, they may be more common (and thus have a wider range of medications available). Such findings may help, for instance, prioritize diseases for future research or select the best treatment for a given disease, among other health-related analyses.

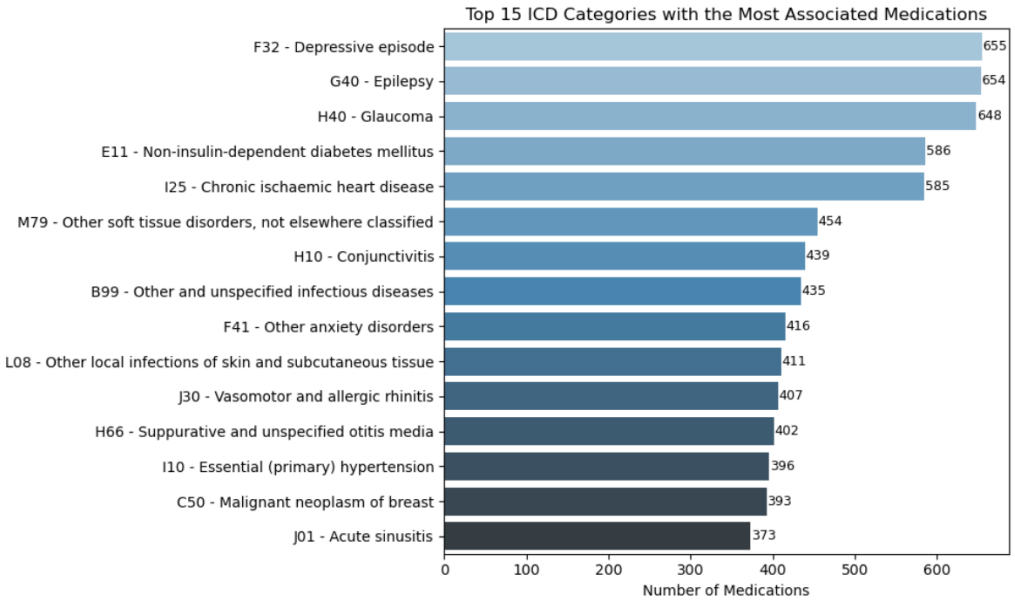

Figure 3. Medication–disease association in HealDB.

5. Enabling Interoperability in HealDB – Use Case around RxNorm

While Section 4 presents how we identified disease ICD-10 codes from leaflet texts, and used them to process queries *within* HealDB, this section shows how we are extending the database so that it can interoperate *with external sources*. The basic idea is to import into HealDB, for each external data source, its corresponding identifier(s). This is a time-intensive process, since each external data source has several identifier attributes, with specific semantics and roles. Hence, this is not just a matter of creating a new column in HealDB; rather, each data source needs to be analyzed and understood as regards usages, users and attribute semantics, so that the most appropriate identifiers can be imported. Moreover, each external source has distinct access interfaces – again, demanding

a previous analysis to choose the most appropriate processing method. We now show how we implemented interoperability with RxNorm using information in HealDB active ingredients tables.

5.1. Use of HealDB Active Ingredients to import appropriate RxCUI from RxNorm

RxNorm, developed and maintained by the US National Library of Medicine, is a "standardized naming system for generic and brand-name drugs; and a tool to support semantic interoperation between drug terminologies and pharmaceutical knowledge base systems"[9]. We analyzed its several APIs (each of which returning distinct features) and 43 properties. For interoperability, we chose RxCUIs attributes (RxNorm Concept Unique Identifiers) linked to concepts related to US generic and brand-name drugs.

One of HealDB internal tables contains standardized active ingredient names in English, to allow us to support interoperability with external English-language data sources [Martins and Medeiros 2024]. We employed these names in English to call the RxNorm API, retrieving the corresponding RxCUI and storing it in HealDB.

5.2. Extending HealDB beyond its boundaries

Next, we used the RxCUIs retrieved in the previous step to access additional RxNorm APIs. These APIs returned complementary information about the associated active ingredients, such as the preferred name, synonyms, and the date of the last update in RxNorm. Additionally, these APIs provide related information that connect each active ingredient to other RxNorm concepts, such as clinical drugs and branded drugs.

This exemplifies how to support semantic enrichment of active ingredients, thereby integrating with additional biomedical databases and ontologies – e.g., via queries. Through RxNorm alone, this allows the following competency queries, among others: what is the standardized name, concept type, and status of an active ingredient (e.g., "aspirin", IN, active); which clinical drug presentations (strength and form) are associated with it (e.g., "aspirin 325 MG Oral Tablet"); what brand drug concepts are related to this compound (e.g., "Bayer Aspirin", "St. Joseph Aspirin"); and what is the current status of the ingredient in the RxNorm database (e.g., active, current release since 04/2005).

First, these examples show that HealDB enables alternative standardizations - e.g., active ingredient names according to RxNorm. Second, it supports enrichment of the clinical context through the representation of concentration and pharmaceutical forms. Third, the inclusion of synonyms allows for more accurate search processes that can combine drug leaflets, medical notes, and patient records.

6. Conclusions and Ongoing Work

This paper describes our ongoing effort to construct HealDB, a curated database to serve as an integration hub for health information systems that is centered on authoritative official Brazilian sources. As shown here, we are able to provide interconnections to distinct large health vocabularies and databases, thereby facilitating queries that would require extensive navigation across such sources. We spent considerable time refining data curation tasks and performing AI-assisted tasks to extend HealDB with means to link with

[9]https://www.nlm.nih.gov/research/umls/rxnorm/overview.html

major international data sources and ontologies. Our examples show how we can gain new insights into the Brazilian health scenario, even without considering EHR.

Ongoing work involves processing distinct kinds of queries and correlations across additional data sources, and validating them with the help of domain experts. One such task involves cross-referencing Brazilian drugs and diseases with data from ClinicalTrials.gov.

References

Bhatia, P., Celikkaya, B., Khalilia, M., and Senthivel, S. (2019). Comprehend medical: A named entity recognition and relationship extraction web service. In *Proc. 18th IEEE International Conference On Machine Learning And Applications*, pages 1844–1851.

Haux, R. (2006). Health information systems–past, present, future. *International journal of medical informatics*, 75(3–4):268–281.

Maciel, R. S. P., Valle, P. H. D., Santos, K. S., and Nakagawa, E. Y. (2024). Systems Interoperability Types: A Tertiary Study. *ACM Computing Surveys*, 56(10).

Martins, M. J. A. and Medeiros, C. B. (2023). Linking Heterogeneous Health Data Sources in Brazil Centered on Drug Leaflet Processing. In *Proc. XXXVIII Brazilian Database Symposium*, pages 366–371. SBC - Brazilian Computer Society.

Martins, M. J. A. and Medeiros, C. B. (2024). Construction of Open Data Sources for Data Interoperability in Brazilian Health Information Systems. In *Proc. 28th European Conference on Databases and Information Systems – ADBIS 2024 - DOING workshop (Intelligent data - from data to knowledge)*, pages 117–129. Springer - CCIS vol 2186.

Martins, M. J. A. and Medeiros, C. B. (2025). HealDB - an open Portuguese language database for health information systems, V1. https://doi.org/10.25824/redu/24I1FH.

Sallauka, R., Arioz, U., Rojc, M., and Mlakar, I. (2025). Weakly-supervised multilingual medical ner for symptom extraction for low-resource languages. *Applied Sciences*, 15(10):5585.

Schneider, E., de Souza, J., Knafou, J., Copara, J., e Oliveira, L., Gumiel, Y., de Oliveira, L., Teodoro, D., Paraiso, E., and Moro, C. (2020). Biobertpt - a portuguese neural language model for clinical named entity recognition. In *Proc. 3rdClinical Natural Language Processing workshop*, pages 65–72.

Shaitarova, A., Zaghir, J., Lavelli, A., Krauthammer, M., and Rinaldi, F. (2023). Exploring the Latest Highlights in Medical Natural Language Processing across Multiple Languages: A Survey. *Yearbook of medical informatics*, 32(1):240–243.

Simões, A. and Gamallo, P. (2021). Leme-pt: a medical package leaflet corpus for portuguese. In *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*, pages 10–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Sohn, S. and Liu, H. (2014). Analysis of medication and indication occurrences in clinical notes. *AMIA Annu Symp Proc*, 2014:1046—1055.

Sousa, H., Mario Jorge, A., Pasquali, A., Santos, C., and Lopes, M. (2023). A biomedical entity extraction pipeline for oncology health records in portuguese. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, page 950–956.