POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Executive Summary of the Thesis

Development of a Mapping Tool between EMDN and GMDN Nomenclatures to Support Post-Market Surveillance of Medical Devices

Laurea Magistrale in Computer Science and Engineering

Author: Riccardo Gibello

Advisor: Prof. Enrico Gianluca Caiani

Co-advisor: Eng. Yijun Ren

Academic year: 2021-2022

1. Introduction

The current landscape of Medical Device standardization is characterized by a **variety** of **adopted nomenclatures**. This tendency highly affects **regulatory processes** (e.g., post-market surveillance), aimed at ensuring **safety** and **effectiveness** of products, by monitoring the market through proper device identification.

In light of introducing new European Regulations, that reinforce the high-risk medical devices monitoring processes, the need for higher nomenclature interoperability became even more relevant to standardize the identification of these products, when present in world markets using different nomenclature systems.

Starting from these observations, this Master's Thesis project aims at investigating possible **automated solutions** to promote **mapping** between standards, possibly enhancing tasks related to the mentioned regulatory processes.
This work focuses on two broadly adopted nomenclatures:

1. the **European Medical Device Nomenclature** (EMDN), the standard that identifies medical device categories in the European market. It is based on a **tree structure**, in which every branch represents a particular device category.
2. the **Global Medical Device Nomenclature** (GMDN), a 5 code digits standard, widely adopted in many countries (e.g., in the USA, Australia, UK, etc.), which is characterized by a high-granularity **flat** structure.

The proposed mapping algorithm considers the results of a preliminary study, proposed by WHO in 2021 [1], investigating possible modifications to overcome the limitations of that approach.

The adopted methodology applies different aspects of Computer Science to this practical problem, implementing a **client-server infrastructure**, aimed at **promoting accessibility** to the utilization of the proposed algorithm in an easy and **user-friendly** environment. As of today, similar solutions are not available on the market yet. Therefore, there is hope for possible future adoption of this tool in a real setting.

2. Materials and Methods

2.1. Materials

The used materials during the project are all **publicly available data sources**, which correspond to:

- a complete **list of EMDN codes**, downloaded from the European Commission website and contained in a CSV file.
- an **Italian CSV dataset of devices**, maintained by the Italian Ministry of Health, and constantly updated. It contains the **EMDN code** for each medical device.
- the **Global Unique Device Identification Database** (GUDID), administered by the FDA and containing device identification information submitted by manufacturers. This information includes the **GMDN Term Name** and **Definition** of each medical device, and is provided in an XML dataset.

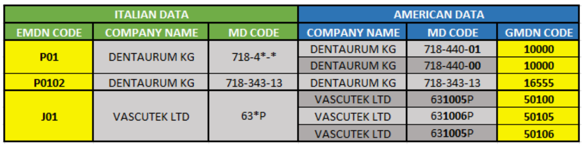
2.2. Methods

2.2.1 Cleansing and Transformation of Heterogeneous Data Sources

In this first stage, the Italian CSV dataset of Medical Devices required a proper pipeline to **select**, among the provided table **features** (=columns), only the required ones. Moreover, **cleansing** of the annotated **"Catalogue Code"** was necessary due to both **malformed** values, and **conventions** used to aggregate codes (e.g., "DA(L) X - A(L) Y", where X and Y are possible codes).

On the other hand, the FDA provided a well-formatted dataset of XML files. Therefore, an easier pipeline was implemented to transform these files into a CSV one, resembling the Italian tabular version.

At this point, **two homogeneous data sources** were obtained, representing the Italian and the American markets. Figure 1 summarizes the most meaningful information of the tables.



ITALIAN DATA			AMERICAN DATA		
EMDN CODE	COMPANY NAME	MD CODE	COMPANY NAME	MD CODE	GMDN CODE
P01	DENTAURUM KG	718-4*-*	DENTAURUM KG	718-440-01	10000
			DENTAURUM KG	718-440-00	10000
P0102	DENTAURUM KG	718-343-13	DENTAURUM KG	718-343-13	16555
J01	VASCUTEK LTD	63*P	VASCUTEK LTD	631005P	50100
			VASCUTEK LTD	631006P	50105
			VASCUTEK LTD	631005P	50106

Figure 1: The main information extracted by Data Cleansing and Transformation. Note that the asterisk wildcard represents families of medical devices, enhancing the number of matches between the two markets.

2.2.2 Data Integration

This second phase is based on previous work, presented by WHO in 2021 [1], which proposed to build a **"bridge" between different nomenclatures** by cross-checking markets of Medical Devices. This liaison between datasets was identified by researchers in the Unique Device Identifier (UDI), which represents a globally used system to unequivocally identify a commercialized medical device.

The adoption of the UDI as a linking key between the American and Italian market datasets, as stated in [2], presents some limitations. This is mainly due to a **different** level of **UDI standard adoption** between Europe and the USA. While the latter reports, for every device, a proper UDI code, in Europe the adoption of this identification system is in its early stages, since the introduction of two new Regulations, the "Regulation (EU) 2017/745 on Medical Devices" [3] and the "Regulation (EU) 2017/746 on In Vitro Diagnostics" [4]. Therefore, many devices do not present any UDI definition yet, preventing the identification of potential medical devices correspondences.

Accordingly, the proposed solution **overcomes this barrier**, matching devices based on the **"Catalogue Code"**, provided by any manufacturer.

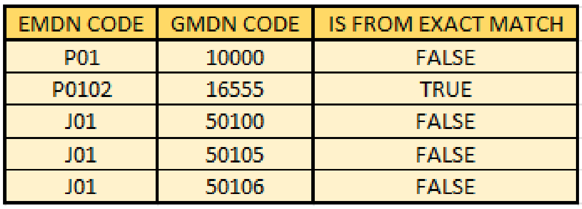
By doing so, it was possible to extract valuable information from the given data sources. Firstly, concerning all active manufacturers in both markets, the following datasets were produced:

- A CSV dataset, related to devices sold only on the Italian market, containing 125,781 devices for 1,425 different companies.
- A CSV dataset, related to devices sold only on the American market, containing 1,089,614 devices for 1,507 different compa-

nies.

- A CSV dataset, related to **devices** sold on **both markets**, containing **209,556** devices for 783 different companies.

Starting from the last produced dataset, a similar data source, as the one proposed by the feasibility study, was easily created. This revealed to be a valuable source of information, representing a possible **ground knowledge** to be fed to the **mapping algorithm**, thus enhancing its decision process.



EMDN CODE	GMDN CODE	IS FROM EXACT MATCH
P01	10000	FALSE
P0102	16555	TRUE
J01	50100	FALSE
J01	50105	FALSE
J01	50106	FALSE

Figure 2: An exemplification of the GMDN-EMDN translations extracted from the starting data sources, and stored in a `Python SQLite` `"mapping"` table.

By doing so, the final version of the extracted vocabulary contains **9863 translations**, among which 4931 were inferred from "exact" catalogue code matches.

2.2.3 Implementation of a Nomenclature Mapping Algorithm

Before implementing a mapping algorithm between different nomenclatures, their structures were analyzed, to possibly find important features to be exploited.
The EMDN is designed with a **tree structure**, as reported by Figure 3.

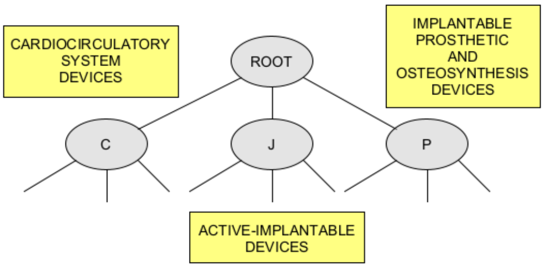

CARDIOCIRCULATORY SYSTEM DEVICES
ROOT
IMPLANTABLE PROSTHETIC AND OSTEOSYNTHESIS DEVICES
C
J
P
ACTIVE-IMPLANTABLE DEVICES

Figure 3: An exemplification of the first level of the EMDN tree, representing three top levels out of 22.

As stated by the official documentation, **each branch** of the tree represents a particular **category** of a medical device. From this characteristic, and performing exploratory analysis on the EMDN data, it is evident that the extraction of **semantic meaning** from text is crucial to find the proper mapping of a GMDN code.
In fact, as reported by Figure 4, and Figure 5, EMDN codes related to different categories, carry different semantic meanings too.

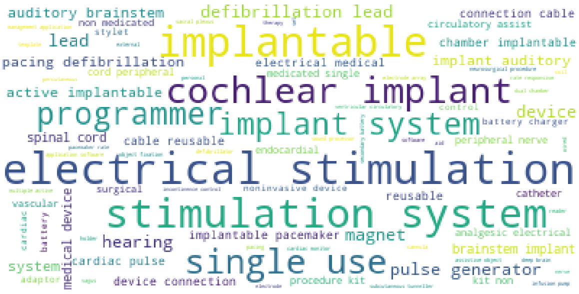


Figure 4: A word map of the EMDN descriptions related to the "J - ACTIVE-IMPLANTABLE DEVICES" top-level.

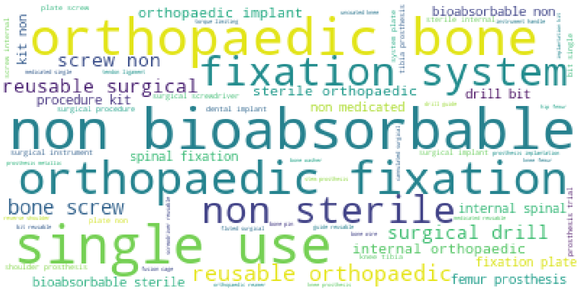


Figure 5: A word map of the EMDN descriptions related to the "P - IMPLANTABLE PROSTHETIC AND OSTEOSYNTHESIS DEVICES" top-level.

For example, the word map extracted from EMDN descriptions related to "P - IMPLANTABLE PROSTHETIC AND OSTEOSYNTHESIS DEVICES" shows a prevalence of terms semantically related to "non-bioabsorbable", "orthopaedic fixation", "orthopaedic bone" and "surgical drill" concepts.
Conversely, the word map related to "J - ACTIVE-IMPLANTABLE DEVICES" shows a tendency towards terms like "electrical stimulation systems", "cochlear implant", and "stimulation system".

Therefore, a proper **algorithmic solution** was implemented to evaluate the **semantic**

similarity of biomedical strings. In this way, it was possible to automatically identify the most **suitable mapping** of a **GMDN code** into the EMDN nomenclature, and vice-versa.

Among the possible tested solutions during the project life-cycle, adopting a **pre-trained** Language Model was considered the most reasonable option, because:

- a pre-trained model represents a useful **kick-off** in solving a Natural Language Processing problem. In fact, it is a tool already trained by researchers on particular text corpora, and therefore it can provide baseline performances.
- it **saves computation** time, since the training of a model from scratch requires a high amount of time and resources.
- a suitable training corpus, specifically focused on the medical device field, was **unavailable** at the time of the writing.

Different pre-trained solutions were tested, among which the best one was identified in **"MPNet"** model, implemented by Microsoft and presented in [5]. In particular, this solution is one of the state-of-the-art options, mixing the benefits of different previous architectures, such as Masked Language Modeling and Permuted Language Modeling.

The implemented algorithm could be defined as "two-way" because both mapping directions are provided. Figure 6 reports one way of mapping, which can be easily reversed to get the other one.

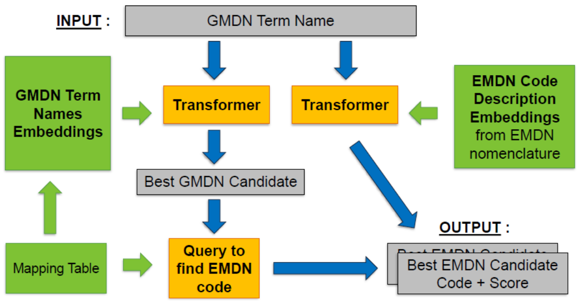

INPUT :
GMDN Term Name
GMDN Term Names Embeddings
Transformer
Transformer
EMDN Code Description Embeddings from EMDN nomenclature
Best GMDN Candidate
Mapping Table
Query to find EMDN code
OUTPUT :
Best EMDN Candidate Code + Score

Figure 6: A box-flow representing the mapping procedure of a GMDN code into the EMDN nomenclature. The "`mapping`" table contains all the GMDN-EMDN translations.

The mapping algorithm is based on the use of **embeddings**, that are vector representations of sentences. These mathematical Language Model structures are properly stored, to be computed only once, and used to evaluate the **cosine similarity** between sentences.

3. A Client-Server Infrastructure to Enable Mapping Algorithm Usability

Besides the actual implementation of the mapping algorithm, an additional part of the project was focused on enabling its **usability** to end-users, in order to **provide** them with **real-time mappings** from GMDN to EMDN codes, and vice-versa.

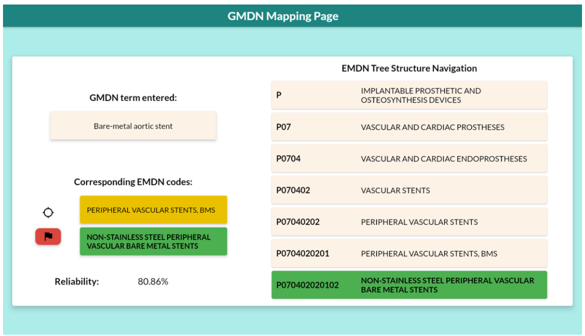
In particular, it seemed reasonable to design a client-server infrastructure, where the `Python`-based server, encapsulating the Machine Learning algorithm, interacts with a **lightweight client**, developed using `Flutter` framework.

The application **access** is **role-based**, and includes the following actors:

- **Server Administrator**: the unique role that can access the Server Dashboards, and start the Information Extraction pipeline, described in Section 2.2.
- **Base User**: the role of every new user, pending Server Administrator confirmation of the registration.
- **Base Authorized User**: the role given to any user whose mapping service subscription has been accepted. In this case, the user can access the mapping dashboards to request EMDN, or GMDN, mappings.
- **Rejected User**: the role given when a registration request is rejected. In this case, no further login is allowed.

The core application functionality is represented by the possibility of sending **GMDN codes** for mapping to the server. Afterwards, the user is redirected to the page shown in Figure 7. In this case, the dashboard displays:

1. On the left, a panel reporting the GMDN input code, the EMDN candidate mapping codes, and the reliability of the selected candidate (i.e., the semantic similarity between the input GMDN Term Name and the EMDN description).
2. On the right, a navigational tool. By **horizontally scrolling** the codes, the user can inspect different EMDN code categories.



GMDN Mapping Page

GMDN term entered:

Bare-metal aortic stent

Corresponding EMDN codes:

PERIPHERAL VASCULAR STENTS, BMS

NON-STAINLESS STEEL PERIPHERAL VASCULAR BARE METAL STENTS

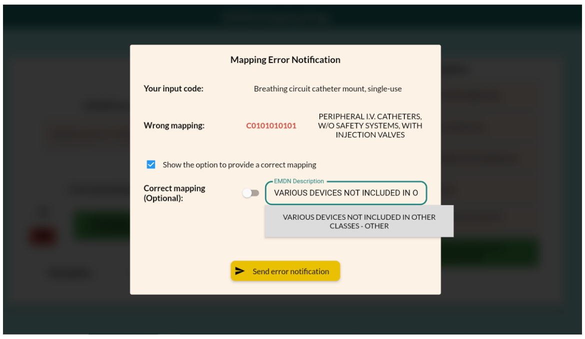
Reliability: 80.86%

EMDN Tree Structure Navigation

P	IMPLANTABLE PROSTHETIC AND OSTEOSYNTHESIS DEVICES
P07	VASCULAR AND CARDIAC PROSTHESES
P0704	VASCULAR AND CARDIAC ENDOPROSTHESES
P070402	VASCULAR STENTS
P07040202	PERIPHERAL VASCULAR STENTS
P0704020201	PERIPHERAL VASCULAR STENTS, BMS
P070402020102	NON-STAINLESS STEEL PERIPHERAL VASCULAR BARE METAL STENTS

Figure 7: The dashboard shown to the user when mapping a GMDN code.

Moreover, in order to **increase** the level of **user engagement**, another page, reported in Figure 8, is implemented to let any user **report an error** in the proposed mapping. Additionally, the user is provided with the possibility to send a **different**, possibly right, **mapping proposal**.



Mapping Error Notification

Your input code: Breathing circuit catheter mount, single-use

Wrong mapping: C0101010101 PERIPHERAL I.V. CATHETERS, W/O SAFETY SYSTEMS, WITH INJECTION VALVES

☑ Show the option to provide a correct mapping

Correct mapping (Optional):

EMDN Description: VARIOUS DEVICES NOT INCLUDED IN O

VARIOUS DEVICES NOT INCLUDED IN OTHER CLASSES - OTHER

Figure 8: The notification page, where the wrong mapping couple of codes is reported. An optional EMDN mapping can be provided by the user.

This functionality is intended to **enrich** the GMDN-EMDN **vocabulary of translations**, and, therefore, to dynamically **update** the mapping algorithm **performances**. Given the high influence of such notification services, manual inspection of Server Administrators is required to make the changes effective in the implementation.

4. Results

During the mapping algorithm validation, data scarcity played a crucial role in the encountered difficulties in properly evaluating the model.

The GMDN-EMDN **vocabulary** was considered as the **Gold Standard** (GS) for the algorithm validation since it was not used for Language Model training, and it was derived from the automatic processing of **manually annotated data**. Therefore, the **GMDN-EMDN** validation set consisted of **2863** unique GMDN Term Names translations, extracted from the identified sources.

	l.1	l.2	l.3	l.4
% of matches	81.94	72.93	56.06	32.20

Table 1: Percentage of mapping matches per minimum EMDN level, on a total test sample of 2863 records.

Considering the GMDN-EMDN validation results, given a GMDN code test sample, the proposed EMDN code can partially match the GS, up to a certain level of the EMDN hierarchy. Considering this particular event, Table 1 reports some aggregated results of the validation. It is clear that, by **deepening the EMDN predicted levels**, the **accuracy decreases** due to an increase in the specificity of each code description. On the other hand, the algorithm was able to identify the right EMDN category (first level) of a GMDN code in 81.94% of the validation samples.

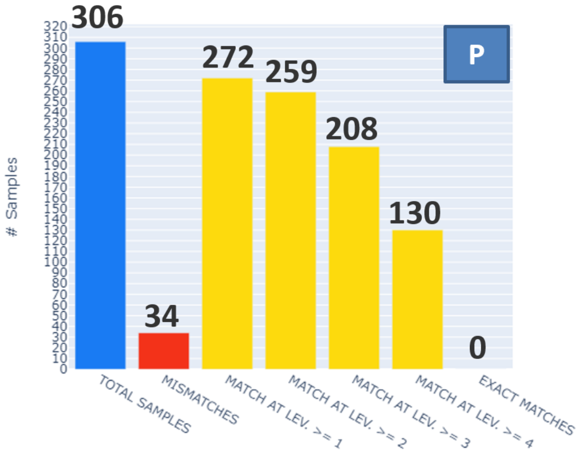


Figure 9: Results of the validation for the "P" EMDN category.

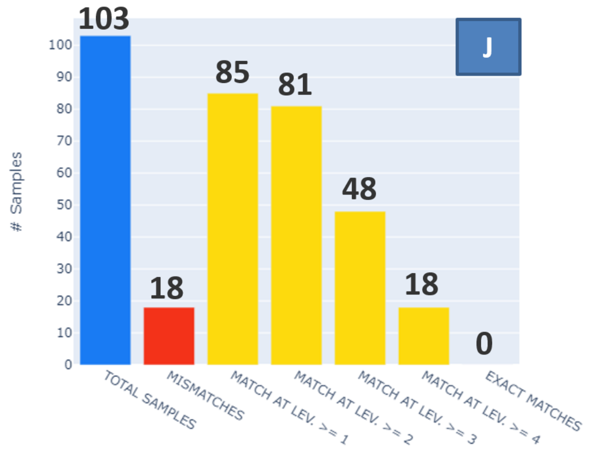


Figure 10: Results of the validation for the "J" EMDN category.

Moreover, since the introduction of the new European Regulations, requiring a higher clinical evidence level for high-risk medical devices, there is a growing interest of manufacturers, and regulatory organizations, towards these products. Figure 9 and Figure 10 report the validation results of the specific EMDN categories related to this critical type of medical devices, which can be identified in "J - ACTIVE-IMPLANTABLE DEVICES" and "P - IMPLANTABLE PROSTHETIC AND OSTEOSYNTHESIS DEVICES".

5. Conclusions

This Master's Thesis project was aimed at developing a tool to **ease** some tasks related to the **regulatory processes** of Medical Devices, and, as of today, still carried out by hand. Therefore, it represented an opportunity both to **develop** an **algorithm** to solve this problem, and to face the challenges related to its **usability** and **encapsulation** in a fully working end-user application.

The proposed application represents a solution that was not investigated before. In fact, the only available scientific research on the topic consists of the mentioned WHO feasibility study [1], which was used as starting point to solve the given problem.

A similar service to this tool is offered by the GMDN Agency, with payment of an annual fee. In fact, they allow subscribed users to access a dedicated Web interface, and to add GMDN codes to a "cart". Then, the user can download, in an XML document, the EMDN translations of the selected GMDN codes.

Moreover, this application was developed using **only open data**, promoting **reusability** of the tool for other research studies.

To conclude, proper validation of the implemented mapping algorithm and code testing was carried out, whereas the end-user application has not been validated yet.
On the other hand, the final results of this thesis have been presented in **separate sessions** to several **stakeholders**, whose tasks are strictly related to the nomenclatures standardization. **Positive comments** were expressed during these meetings on the application **usability**. Particular attention was given towards the **adopted methodology**, and the implemented **feedback functionality** to simultaneously **engage** users and **enhance** the algorithm performances. This gives hope about a possible future adoption of this tool in a real setting.

References

[1] World Health Organization. *Webinar: Medical devices nomenclature mapping*. 2021. URL: https://www.loom.com/share/18847ab410d444a68f9da2ee88bc6e07 (visited on 02/23/2023).

[2] World Health Organization. “Standardization of medical devices nomenclature”. In: (May 2022). URL: https://apps.who.int/gb/ebwha/pdf_files/WHA75/A75_11-en.pdf (visited on 04/05/2023).

[3] European Commission. *Medical Devices Regulation (EU) 2017/745*. 2017. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745 (visited on 04/05/2023).

[4] European Commission. *In Vitro Diagnostic Devices Regulation (EU) 2017/746*. 2017. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0746 (visited on 04/05/2023).

[5] Kaitao Song et al. “MPNet: Masked and Permuted Pre-training for Language Understanding”. In: *Advances in Neural Information Processing Systems* 33 (Apr. 2020), pp. 16857–16867.