POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Development of an IT tool for assisting Medical Devices Post-Market Surveillance: application to the Italian scenario

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Michele Amedeo Bertoldi

Advisor: Prof. Enrico Gianluca Caiani

Academic year: 2020-2021

1. Introduction

In EU, medical technologies are strictly regulated by laws that govern the safety and performance of devices throughout their life cycle, before and after they are placed on the market. Specifically, the EU Medical Device (MD) sector has been regulated by the Medical Device Regulation (MDR) EU 2017/745 since May 26th, 2021. One of the main focuses of this regulation is the Post Market Surveillance (PMS), i.e., the process of monitoring incidents once products are put on the market and used by patients, through the collection of the notices that the manufacturers must report.

In this context, this project aims at developing a tool to support expert panels during the Post Market Surveillance process of EU devices [2]. In order to achieve this goal, we developed a model focusing on the Italian MD market. In fact, the sources to access the information of the Italian MD market are official and publicly available. Moreover, the European Union decided to use the Italian CND classification nomenclature to categorise devices of all EU Member States.

The final product of this thesis is a graphical user interface for clustering and visualising the medical devices notices through different functionalities such as filtering the devices by one or more manufacturers or by one or more CND categories. By accurately selecting and visualising the portion of the dataset of interest, expert panel members are helped in the time-consuming initial MD screening process and, through the available graphs and statistics, are supported in the process of studying the phenomenon.

2. Proposed Methodology and Implementation

2.1. Data Sources

Two datasets were created for the purpose of this project. Dataset 1 was downloaded from the register of all the MDs available on the Italian market that is publicly available on the Ministry of Health website. This dataset contains all the information about each medical device (Figure 1).

Dataset 2 was extracted from the official notices present on the Ministry of Health website, using a computerised bot (see Section 2.2). This dataset contains the reports of incidents, the Field Safety Corrective Actions (FSCAs) and the Field Safety Notices that each manufacturer has to submit to the Ministry of Health.

The merging of these two datasets allows to re-

	FABBRICANTE_ASSEMBLATORE	DENOMINAZIONE_COMMERCIALE	CODICE_CATALOGO_FABBR_ASS	PROGRESSIVO_DM_ASS	CLASSIFICAZIONE_CND	DESCRIZIONE_CND
0	IDACO S.R.L.	PENNARELLI DERMOGRAFICI	PD01R	1221	V9004	MATITE DERMOGRAFICHE
1	JOHNSON & JOHNSON MEDICAL HOLDING S.P.A.	VACUDRAIN	V400	1241	A060299	SISTEMI DI DRENAGGIO PERCUTANEO - ALTRI
2	BUERO INNO MED	IRRIGAZIONE/ASPIRAZIONE PER TECNICA BI...	BSD 500/510/560	1262	K01020110	CANNULE MONOUSO PER ASPIRAZIONE/IRRIGAZIONE IN...
3	ACTIMEX SRL	RINOCICLINA SPRAY NASALE	ACDM-01-07	1281	G030199	DISPOSITIVI NASOFARINGEI - ALTRI
4	JOHNSON & JOHNSON MEDICAL HOLDING S.P.A.	VACUDRAIN	V410	1301	A060299	SISTEMI DI DRENAGGIO PERCUTANEO - ALTRI
...	...	...	...	...	...	...
1474970	HOFER GMBH & CO KG	TRIAL INTEOS PLACCA CALCAGNO M3-3,5 CON LOOP A...	731-111-135-002-T	2100906	L091099	STRUMENTARIO PLURIUSO PER OSTEOSINTESI - ALTRO
1474971	HOFER GMBH & CO KG	TRIAL INTEOS PLACCA CALCAGNO M3-3,5 CON LOOP A...	731-111-135-003-T	2100907	L091099	STRUMENTARIO PLURIUSO PER OSTEOSINTESI - ALTRO
1474972	MHK MEDIKAL TEKSTIL SAN. TIC. LTD. STI	KIT PER BIOPSIA MHK	212812	2103908	T0202	KIT CHIRURGICI (sono compresi solo i Kit cert...
1474973	MHK MEDIKAL TEKSTIL SAN. TIC. LTD. STI	KIT PER ANGIOGRAFIA MHK	212811	2103909	T0202	KIT CHIRURGICI (sono compresi solo i Kit cert...
1474974	MHK MEDIKAL TEKSTIL SAN. TIC. LTD. STI	KIT PER DRENAGGIO MHK	212813	2103910	T0202	KIT CHIRURGICI (sono compresi solo i Kit cert...

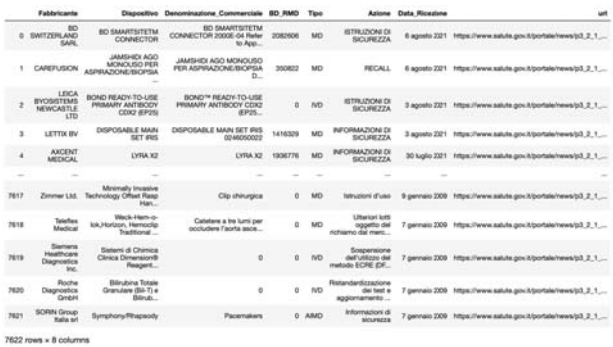
1474975 rows × 6 columns

Figure 1: Example of samples of Dataset 1. In the figure are showed only the most relevant features, the ones used for this project.

construct, for each element in Dataset 2, the CND classification assigned during registration.

2.2. Web Scraping

Since the notices of Dataset 2 were not immediately available in an analyzable format but only viewable as text in the Ministry of Health website, web scraping has been used. Web scraping is a computer technique for extracting data from a website by means of software programs. Typically, such programs simulate human navigation on the World Wide Web using the HTTP or through browsers, such as Internet Explorer or Google Chrome. The data scraping technique used for this project is HTML parsing as Dataset 2 is organized in a standardized way and updated daily. From the HTML file of the Ministry of Health website, a bot "scrapes" the parameters of interest in an automated way. The bot was implemented using the programming language Python and, more specifically, the Python library Selenium which allows HTML parsing and is based on the Google Chrome Webdriver named Chromedriver. The main features of Dataset 2 extracted from each notice are reported in Figure 2



	Fabbricante	Dispositivo	Denominazione_Commerciale	BD_RMD	Tipo	Azione	Data_Ricezione	url
0	BD SWITZERLAND SARL	BD SMARTSITETM CONNECTOR	BD SMARTSITETM CONNECTOR 2000E-04 Refer to App...	2082606	MD	ISTRUZIONI DI SICUREZZA	6 agosto 2321	https://www.salute.gov.it/portale/news/p3_2_1_...
1	CAREFUSION	JAMSHIDI AGO MONOUSO PER ASPIRAZIONE/BIOPSIA ...	JAMSHIDI AGO MONOUSO PER ASPIRAZIONE/BIOPSIA D...	350822	MD	RECALL	6 agosto 2321	https://www.salute.gov.it/portale/news/p3_2_1_...
2	LEICA BIOSYSTEMS NEWCASTLE LTD	BOND READY-TO-USE PRIMARY ANTIBODY CDX2 (EP25)	BOND™ READY-TO-USE PRIMARY ANTIBODY CDX2 (EP25...	0	IVD	ISTRUZIONI DI SICUREZZA	3 agosto 2321	https://www.salute.gov.it/portale/news/p3_2_1_...
3	LETTIX BV	DISPOSABLE MAIN SET IRIS	DISPOSABLE MAIN SET IRIS 0246050022	1416329	MD	INFORMAZIONI DI SICUREZZA	3 agosto 2321	https://www.salute.gov.it/portale/news/p3_2_1_...
4	AXCENT MEDICAL	LYRA X2	LYRA X2	1936776	MD	INFORMAZIONI DI SICUREZZA	30 luglio 2321	https://www.salute.gov.it/portale/news/p3_2_1_...
...	...	...	...	...	...	...	...	...
7617	Zimmer Ltd.	Minimally Invasive Technology Offset Resp Han...	Clip chirurgica	0	MD	Istruzioni d'uso	9 gennaio 2309	https://www.salute.gov.it/portale/news/p3_2_1_...
7618	Teleflex Medical	Weck-Hem-o-lok,Horizon, Hemoclip Traditional ...	Celetere a tre lumi per occludere l'aorta asce...	0	MD	Ulteriori lotti oggetto del richiamo dal merc...	7 gennaio 2309	https://www.salute.gov.it/portale/news/p3_2_1_...
7619	Siemens Healthcare Diagnostics Inc.	Sistemi di Chimica Clinica Dimension® Reagent...	0	0	IVD	Sospensione dell'utilizzo del metodo ECRE (DF...	7 gennaio 2309	https://www.salute.gov.it/portale/news/p3_2_1_...
7620	Roche Diagnostics GmbH	Bilirubina Totale Granulare (Bil-T) e Bilirub...	0	0	IVD	Ristandardizzazione dei test e aggiornamento ...	7 gennaio 2309	https://www.salute.gov.it/portale/news/p3_2_1_...
7621	SORIN Group Italia srl	Symphony/Rhapsody	Pacemakers	0	AIMD	Informazioni di sicurezza	7 gennaio 2309	https://www.salute.gov.it/portale/news/p3_2_1_...

7622 rows × 8 columns

Figure 2: Example of samples of Dataset 2 and their features

Once the data had been acquired, we realised that the feature enabling devices from Dataset 2 to be associated with Dataset 1 was present in only about 30% of the 7622 alerts on the website. This feature is called BD/RMD and is a unique identification number of the device on the Italian market, that corresponds to the PROGRESSIVO DM/ASS of Dataset 1. For this reason, we used natural language processing to compare the other features of the two datasets and associate them with each other.

2.3. Data Cleansing and Standardization

Data cleansing process consists of validating and correcting data to make make it easily analyzable by removing duplicates and missing values, correcting typos and transforming it into the right format for further processing. For this process a Python string-processing library, namely NLTK, that contains a large number of functions for cleansing methods, was used.
The data transformation workflow using NLTK libraries includes:

- Removing extra spaces: remove extra spaces from each element by using regular expressions.
- Removing punctuation: punctuation, when attached to any word, creates problems in differentiating with other words.
- Case Normalization: converts the case of all characters in the text to either upper or lower case, as Python is case-sensitive.
- Tokenization: each sentence is split into words and becomes a list of words. The name of the function used is *word_ tokenize*.
- Removing Stopwords: stopwords (e.g., and, but, was, were, being, have, etc.) are words that do not add meaning to the data and so should be removed. This reduce the dimension of our data removing noise.
- Stemming: it is a technique that takes the word to its root form, by removing suffixes from the words.

The stemming part was necessary to standardize the names of different departments of the same company (e.g., Siemens Diagnostic and Siemens Healthineers). Data Cleansing and Standardization made the information of Dataset 2 uniform and allowed the comparison between strings of the two datasets.

Dataset 1

	FABBRICANTE_ASSEMBLATORE	DENOMINAZIONE_COMMERCIALE	CODICE_CATALOGO_FABBR_ASS
0	id	pennarelli dermografici	pd01r
1	johnson	vacudrain	v400
2	buerki	cannule irrigazioneaspirazione per tecnica bim...	bsd 500510560
3	actimex	rinociclina spray nasale	acdm0107
4	johnson	vacudrain	v410
...	...	...	...
1474970	hofer	trial inteos placca calcagno m33 5 con loop as...	731111135002t
1474971	hofer	trial inteos placca calcagno m33 5 con loop as...	731111135003t
1474972	mhk	kit per biopsia mhk	212812
1474973	mhk	kit per angiografia mhk	212811
1474974	mhk	kit per drenaggio mhk	212813

1474975 rows × 6 columns

Dataset 2

	Fabbricante	Dispositivo	Denominazione_Commerciale
0	bd	bd smartsitetm connector	bd smartsitetm connector 2000e04 refer to appe...
1	carefusion	jamshidi ago monouso per aspirazionebiopsia de...	jamshidi ago monouso per aspirazionebiopsia de...
2	leica	bond readytouse primary antibody cdx2 ep25	bond(tm) readytouse primary antibody cdx2 ep2...
3	lettix	disposable main set iris	disposable main set iris 0246050022
4	axcent	lyra x2	lyra x2
...	...	...	...
7617	zimmer	minimally invasive technology offset rasp hand...	clip chirurgica
7618	teleflex	weckhemolok horizon hemoclip traditional e tr...	catetere a tre lumi per occludere l'aorta asce...
7619	siemens	sistemi di chimica clinica dimension(r) reagen...	0
7620	roche	bilirubina totale granulare bilt e bilirubin...	0
7621	sorin	symphonyrhapsody	pacemakers

7622 rows × 8 columns

Figure 3: Compared features of the two datasets. Features highlighted with the same color are merged together.

2.4. Entity Resolution

Entity Resolution (ER) is the problem of extracting, matching and resolving entity links in structured and unstructured data. Our task is a problem of Single-entity ER because all mentions correspond to a single entity type as we only deal with the MDs [1]. Moreover the decision to match a pair of mentions is made independently of other mentions so our use case falls into pairwise ER category. Figure 3 shows how Pairwise ER is performed on the two datasets. Features highlighted with the same color are merged together.

2.4.1 Natural Language Processing - String Matching

To compare the strings of figure 3, a particular branch of Natural Language Processing (NLP), called string matching was used. Specifically, we used approximate matching techniques because the formatting and nomenclature adopted in the two datasets were often inconsistent but similar. Approximate matching allows to match strings with a similar, but not identical, pattern. Approximate string matching is typically divided into two sub problems: finding approximate sub-string matches inside a given string and finding dictionary strings that match the pattern approximately. Among the approximate string-matching techniques, fuzzy logic was chosen because, despite other methods, it does not provide a Boolean result (i.e., "similar" or "dissimilar"), but provides a score from 0 to 100 that better represents the degree of similarity between two entities and allows to handle more complex scenarios.
This thesis's fuzzy system was implemented using the Fuzzy-wuzzy library. In particular, token_set_ratio() build-in function was used. It removes the common tokens and then calculates the Levenshtein distance between strings.
In our projects the strings are variable in length and can be sorted in a very heterogeneous way. To make an example, the model has to search the manufacturer's catalogue code of Dataset 1 (i.e., a single string) in the product description of Dataset 2 which is a much longer sets of strings. Additionally, some strings repeated multiple times within the same product description. The algorithm uses specific fuzzy rules to overcome these problems and provides an accurate similarity score from 0 to 100.

2.4.2 Implemented Entity Resolution Algorithm

To compare the samples of Dataset 1 with those of Dataset 2, it was decided to use fuzzy string comparisons [3] of the features represented in Figure 3.
The idea behind the algorithm is to associate each item in Dataset 2 to the corresponding MD of Dataset 1, in order to assign through the BD/RDM the corresponding CND to each medical device in Dataset 2. About 30% of the items of Dataset 2 have the BD/RMD information available, so that the association between the two datasets is straightforward. For all the remaining items, the BD/RMD is missing. In this cases, the fuzzy method described in the previous chapters was used to calculate the similarity scores between one items of Dataset 2 and all the items of Dataset 1. The developed algorithm works iteratively. In particular, the algorithm set a similarity score threshold and creates an association for all the items that get a similarity score within this threshold. At the next iteration, the threshold is lowered down and all the items of Dataset 2 that are were not yet assigned to a MD in Dataset 1 are analyzed. Figure 4 shows the flowchart of the implemented Entity Resolution algorithm.

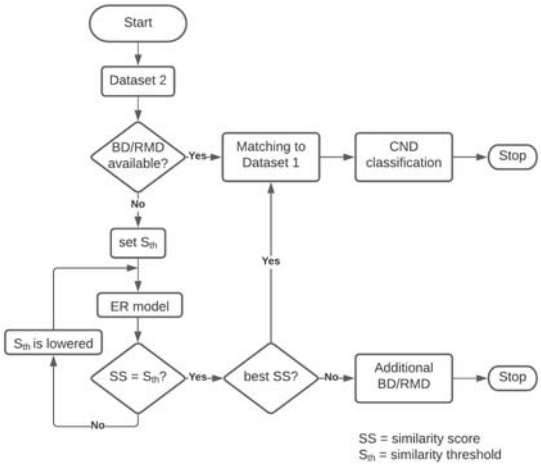


Figure 4: Flowchart of the implemented Entity Resolution algorithm.Each arrow concerns all the samples of Dataset 2 that satisfy the specified condition.

Originally, a similarity score threshold of 100 was considered. Thus, all the items of Dataset 2 that got a similarity score of 100 were associated with the corresponding item of Dataset 1, assigning the same BD/RMD. If an item of Dataset 2 resulted in a similarity score of 100 with multiple items of Dataset 1, the item in Dataset 2 was randomly assigned to one of the BD/RMD of Dataset 1. The remaining items of Dataset 1 with whom the element of Dataset 2 also got 100, were separately stored in a features called "additional BD/RMD". Subsequently, the similarity score threshold was lowered between 99 and 95 and the algorithm searched, among the items of Dataset 2 that have not been associated to an MD in Dataset 1 yet, for those that had a similarity score within the range. Each item of Dataset 2 was associated to the item of Dataset 1 with which it received the higher similarity score within the considered range. The BD/RMD of all the other MDs of Dataset 1, with which the item of Dataset 2 got a similarity score within the threshold but not the higher one, were separately stored in the "additional BD/RMD". The same process was repeated for 94-90, 89-85, 84-80, 79-75, 74-70, 69-65 and 64-60. Lower ranges were not considered because they were found to deteriorate the accuracy of the model and because they returned too many possible matches.

2.4.3 Graphical User Interface

In order to visualize the data and facilitate its analysis, a Graphical User Interface (GUI) was created with the aim of displaying the MD notices of Dataset 2, filtered and enriched with the information of Dataset 1. This is a business intelligence problem, and for this purpose a Microsoft tool called PowerBI was used.
The features of the two dataset that have been represented in the GUI in order to support the PMS process are:

- Fabbricante: Name of the manufacturer of the device.
- Dispositivo: Name of the device
- Classificazione_cnd_1: First level of the CND classification of the device.
- Classificazione_cnd_2: Second level of the CND classification of the device.
- Classificazione_cnd_3: Third level of the CND classification of the device.

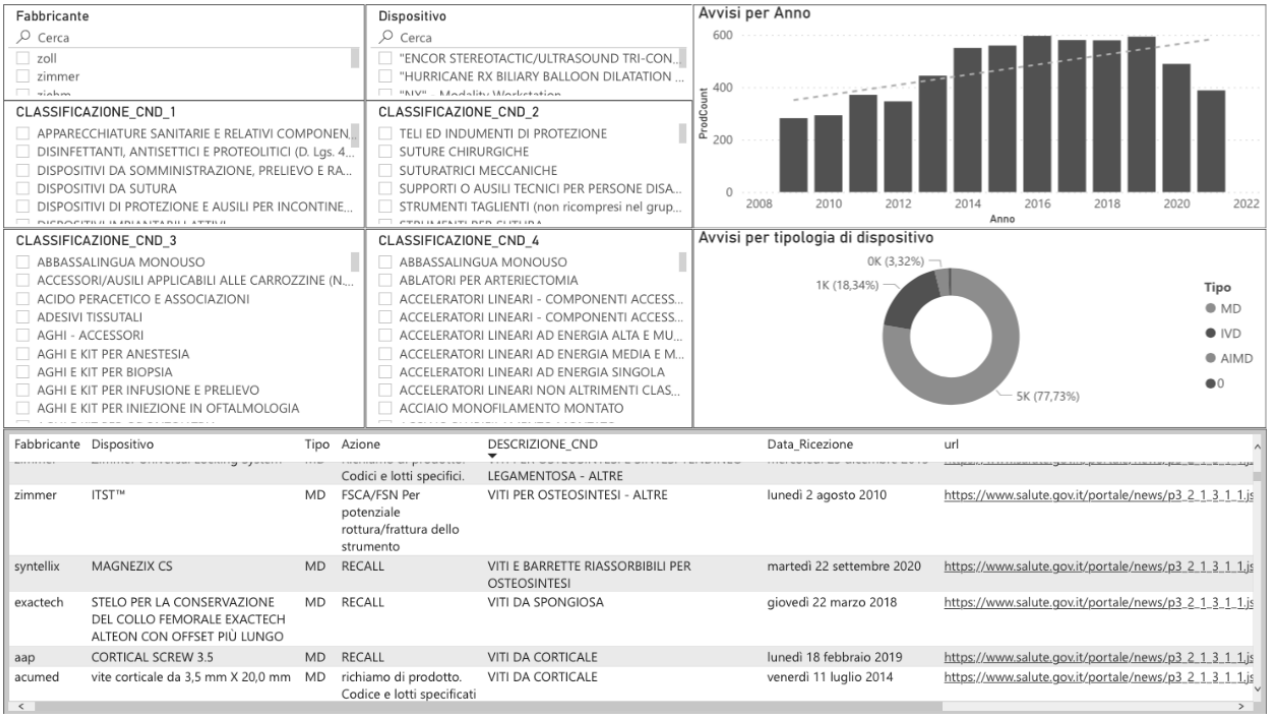


Figure 5: Screenshot of the PMS tool.

- Classificazione_cnd_4: Fourth level of the CND classification of the device.
- Azione: Typology of the notice.
- Data_Ricezione: Date of the notice.
- URL: link to the original notice on the Ministry of Heatlh Website.

This visualization tool can be used to create effective queries and obtain the notices of interest. For each query, two charts are obtained from the aforementioned features and visualized, namely:

- Avvisi per Anno: an histogram representing the number of alerts per year identified by the query;
- Avvisi per tipologia di dispositivo: a diagram representing the distribution of MDs, IVDs and AIMDs identified by the query.

Figure 5 represents the PowerBI GUI tool for PMS implemented for this thesis project.

3. Results

A fundamental process to understand the quality of our tool was to test the accuracy of the NLP algorithm to reconstruct the links between the elements of the two datasets. For this validation process, the subset of Dataset 2 that contained only those samples that possessed the number of registration BD/RMD was used. This allowed to have a reference target variable for 2440 samples out of the 7622 samples of Dataset 2, to test the ER algorithm explained in Section 2.4.

Three tests were carried out as part of the validation process. The first test aimed at measuring the algorithm's ability to correctly assign the BD/RMD code; the second test investigated if the correct BD/RMD exceeded the algorithm's similarity threshold when the BD/RMD was assigned worngly; the third test measured the ability to correctly assign the CND class. This last test, in particular, allowed us to understand whether the number of samples present in the clusters selected by the queries of our graphical user interface were accurate. The entity resolution algorithm assigned correctly the BD/RMD in the 68.89% of the samples of Dataset 2. This means that the notice of Dataset 2 was associated with the correct MD of Dataset 1, and, therefore, all the CND classification levels (up to CND level 4) were assigned correctly. As already mentioned, the second test checked if the correct BD/RMD were included in the list that contains all the BD/RMDs of MDs of Dataset 1 with which an item of Dataset 2 received a similarity score within the considered thresholds but was not associated with because it was not the best score. This test showed that in 98.72%

of the cases the correct BD/RMD was present in the list. The final test measured the difference between the CND classification level of the predicted sample with respect to target sample. Figure 6 shows the accuracy of the algorithm in assigning the CND classifications levels for the entire Dataset 2. The 88.48% of the samples were classified with the correct fourth level of CND classification. If we also consider those samples with the correct third level of CND classification, the cumulative accuracy rises to 92.95%. If we consider the accuracy of getting the CND class 2 correct, this rises to 94.75%, and considering the CND Class 1 correct the accuracy rises to 95.9%. The first CND level of classification was wrongly assigned in 4.10% of the total samples of Dataset 2, meaning that 100 out of 2440 devices were assigned to a completely wrong category.

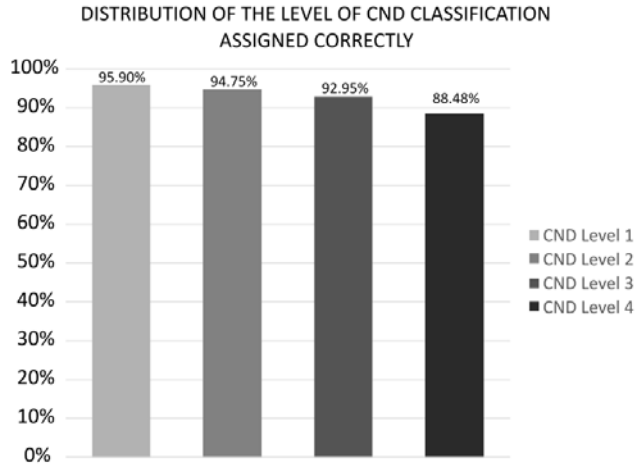


Figure 6: Results of the Entity Resolution algorithm on Dataset 2.

The same tests were conducted specifically for the cardiovascular and orthopaedic device categories. The results for these two sub-sets were consistent with those obtained for the entire dataset.

4. Discussion and Conclusions

The aim of this project was to develop a tool to potentially support expert panels during the certification process of EU medical devices to decide if a new device would need an additional clinical review of the provided data, based on signals of possible increased risks that could be derived from Post-Market Surveillance data. As far as the result obtained is concerned, we believe it is a very positive result. In fact, with the developed tool, it is possible to obtain a level of detail of over 90% which allows the members of the export panels to be facilitated in their research as they could reach almost complete information in a very short time. However, this work certainly has limitations, the reduced size of Dataset 2 did not allow us to use machine learning methods. In fact, only 2440 samples in Dataset 2 presented the target variable (BD/RMD). At the same time, comparing about 5000 samples of Dataset 2 (i.e., the samples without BD/RMD) with more than 1.5 million samples of Dataset 1 with a classical calculation method resulted in a very expensive task, requiring optimisation (i.e, parallelization) to achieve an acceptable computational time. Another limitation of this work is that notices were not categorized in terms of safety concerns for the patient. Actually this information could be accurately obtained only by examining the content of the .pdf file from the notice webpage. Starting from this last aspect, as future works, it is clear that extrapolating and analysing the contents of the pdf files describing the individual notifications would allow to categorise the notices by the type and severity of the device's technical problem. Another aspect that could be explored in detail is the action taken by the manufacturer after the incident, e.g., the device may be recalled, undergo a modification or require a software update.

References

[1] Vassilis Christophides, Vasilis Efthymiou, Themis Palpanas, George Papadakis, and Kostas Stefanidis. An overview of end-to-end entity resolution for big data. *ACM Computing Surveys (CSUR)*, 53(6):1–42, 2020.

[2] Rima Izem, Matilde Sanchez-Kam, Haijun Ma, Richard Zink, and Yueqin Zhao. Sources of safety data and statistical strategies for design and analysis: postmarket surveillance. *Therapeutic innovation & regulatory science*, 52(2):159–169, 2018.

[3] Vilém Novák. Fuzzy logic in natural language processing. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2017.