Article

Leveraging natural language processing to aggregate field safety notices of medical devices across the EU

Check for updates

Yijun Ren [1] & Enrico Gianluca Caiani [1,2] ✉

The European Union (EU) Medical Device Regulation and In Vitro Medical Device Regulation have introduced more rigorous regulatory requirements for medical devices, including new rules for post-market surveillance. However, EU market vigilance is limited by the absence of harmonized reporting systems, languages and nomenclatures among Member States. Our aim was to develop a framework based on Natural Language Processing capable of automatically collecting publicly available Field Safety Notices (FSNs) reporting medical device problems by applying web scraping to EU authority websites, to attribute the most suitable device category based on the European Medical Device Nomenclature (EMDN), and to display processed FSNs in an aggregated way to allow multiple queries. 65,036 FSNs published up to 31/12/2023 were retrieved from 16 EU countries, of which 40,212 (61.83%) were successfully assigned the proper EMDN. The framework's performance was successfully tested, with accuracies ranging from 87.34% to 98.71% for EMDN level 1 and from 64.15% to 85.71% even for level 4.

The field of medical technologies, including medical devices and in vitro medical devices, plays a major role in the healthcare systems in the European Union (EU), with more than 500,000 medical technologies available in hospitals, community care settings and at home[1]. While pre-market evaluation of product quality, safety and performance is part of the conformity assessment procedure, manufacturers must continuously monitor the performance of their medical devices after they have been released into the market through post-market surveillance (PMS)[2,3].

Several EU public health safety issues involving medical devices, such as Poly Implant Prothèse breast implant[4,5] and metal-on-metal hip replacements[6,7], have highlighted the weakness of PMS as codified by the previous Medical Device Directive 93/42/EEC. The fact that, up to 2016, devices approved first in the EU and then in the United States (US) were associated with an increased risk (adjusted hazard ratio equal to 2.9) of safety alerts and recalls once on the market, compared to devices approved first in the US and then in the EU, strengthens this observation[8].

To invert such a tendency towards a more regulated EU market, as well as for greater regulatory transparency for patients and healthcare professionals[9], the EU Medical Device Regulation (MDR) 2017/745[10] and the In Vitro Medical Device Regulation (IVDR) 2017/746[11] have entered into force since 26 May 2021 and 26 May 2022, respectively, introducing stricter rules for risk classification, more rigorous clinical evaluation procedures, and an emphasized codification of PMS requirements throughout the full product lifecycle[12].

According to the MDR, the PMS plan must be suited to actively and systematically gathering, recording, and analysing relevant data on the quality, performance, and safety of a device through its life cycle, enabling any preventive and corrective actions[13]. In this new perspective, the PMS system constitutes the pillar of the manufacturer's quality management system, and it is intended to continually re-verify and re-validate the results of the development phase by gathering real world data from several sources, in order to improve the safety and performance of the device and allow for early detection of possible problems. As a consequence, both proactive and reactive (i.e., vigilance) approaches should be implemented by the manufacturer. Once being notified by final users about possible problems relevant to its device and after having performed an internal assessment, as a response to any serious incident to the device, any field safety corrective action put in place by the manufacturer is described into a Field Safety Notice (FSN), that will be issued electronically through the EUropean DAtabase on MEdical Devices (EUDAMED), a newly created secure and web-based portal that aims to collate and process information about devices and manufacturers at a central level. Such FSNs will become accessible to the public as to the national competent authorities, and could be utilized as part of the data collected by every manufacturer for the proactive PMS, to study trends and deviations for their own product category or similar devices.

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy. [2]IRCCS Istituto Auxologico Italiano, Ospedale S. Luca, Milan, Italy. ✉e-mail: enrico.caiani@polimi.it

The creation of EUDAMED and the setting of the European Medical Device Nomenclature (EMDN)[14,15], as common EU nomenclature system, represent key aspects of the MDR and IVDR to provide standardization in the data collection process and enhance overall transparency and coordination between Member States. EUDAMED, constituted by six modules, was initially scheduled to be fully operational by May 2020, but has experienced several delays: the new expected completion date for the last module is the third quarter of 2026[16]. In accordance with the transitional provisions set out in Regulation (EU) 2024/1860 amending the medical devices regulations, the mandatory use of each module will start 6 months after it is declared functional following an independent audit, and the publication of a Commission notice to that effect in the Official Journal of the European Union. However, the effective future use of EUDAMED, in particular of its vigilance and market surveillance modules relevant to the FSNs, is undermined by the following factors: 1) their limitations are currently unknown and unpredictable, due to several delays in their development; 2) up to that date, utilization of EMDN and common reporting form of FSNs is voluntary; 3) once mandatory, EUDAMED will not contain historical data, now collected by each national competent authority, meaning that generating trends for proactive PMS or safety signal detection would require several years of data accumulation; 4) once functional, no link with other market jurisdictions outside EU will be provided.

As a result, no integrated computerized infrastructure is currently supporting the automated collection of safety information from different EU countries[17], and no global database to access such information is now available[18].

Natural Language Processing (NLP), a branch of Artificial Intelligence, focuses on the design and implementation of systems and algorithms able to interact through human language[19]. The significance of NLP in regulatory science is widely recognized, particularly in extracting useful information from regulatory documents[20,21]. In the context of the EU project Coordinating Research and Evidence for Medical Devices[22], given the current and future limits of EUDAMED, we hypothesized that NLP methods could be exploited to cope with the unstructured and incomplete nature of historical FSN data collected by each national competent authority, to still allow its potential utilization for trend analysis useful in PMS and signal detection. As a result, we proposed a first framework to automatically collect, to classify based on the EMDN, and to display in an aggregated way the official Italian FSN data[23] publicly available through the website of the Italian Ministry of Health. Such approach was then further tested in a pilot study on the FSNs of the Netherlands, characterized by a higher degree of partial data unavailability and unstructuredness[24].

Consolidating on these preliminary results, the aim of this paper was to propose and validate an extended framework based on NLP to automatically collect the publicly available historical FSN information by web scraping from each national competent authority's website for the 27 EU Member States to: i) create a structured global database from such unstructured FSN data; ii) assign for each FSN the EMDN code to the corresponding device; iii) based on specific queries focused on the EMDN code, manufacturer or device, to aggregate such information to allow trends visualization and further post-processing, with possible use in proactive PMS and signal detection.

Results

Initially, a screening was conducted for all the EU 27 national competent authorities to determine whether they regularly updated publicly available FSNs on their official websites. The results showed that only 16 (59%) out of the 27 EU countries do so; this analysis also showed that they currently use different standards, formats, and criteria for reporting (see Supplementary Table 1). Within the proposed methodological framework depicted in Fig. 1 and described in detail in the "Methods" section, the official websites of these 16 EU national competent authorities were explored, and their publicly published FSNs were retrieved automatically by web-scraping, by collecting all the historical information, beginning from the first available FSN for each country and setting as ending date December 31st, 2023. Table 1 provides an overview of the total number of FSNs retrieved for each country (see also Supplementary Fig. 1), along with the issue date of the first available FSN, the estimated number of issued FSN/month, and the number of FSNs for which it was possible to assign the relevant EMDN code to the corresponding device in it.

A total of 65,036 FSNs were retrieved from different EU countries; the total number of retrieved FSNs varied significantly across the countries, also due to the different coverage timeframe in which the FSNs were made publicly available, with a larger number of estimated FSN/month for Germany, followed by Spain and Italy. The outlier value for Portugal is justified by the limited number of medical devices marketed in Portugal due to economic restrictions and of inspections to identify unsafe devices[25], and by the fact that old data stays available online only for a limited period of time.

On average, the EMDN code was successfully assigned to the relevant devices in 40,212 (61.83%) FSNs, with a median value per country equal to 60.47%, with a 95% bias-corrected and accelerated bootstrap confidence interval from 51.80% to 71.83%.

Figure 2 shows the distribution of EMDN level 1 for the FSNs with assigned EMDN levels ($n = 40,212$): it is possible to appreciate how the first three categories with the highest number of FSNs present in the resulting structured database were "Z—Medical equipment and related accessories, software and consumables" ($n = 13,097$; 32.57%), "W—In vitro diagnostic medical devices" ($n = 9723$; 24.18%), and "P—Implantable prosthetic and osteosynthesis devices" ($n = 4046$; 10.06%).

Results of the framework validation

Different tests were conducted to evaluate the performance of the proposed framework in order to assess its ability:
- in identifying named entities from the FSN unstructured text, by the Named Entity Recognition (NER) subtask;
- in assigning the corresponding EMDN code to the device to which the FSN referred, by the Entity Resolution (ER) subtask.
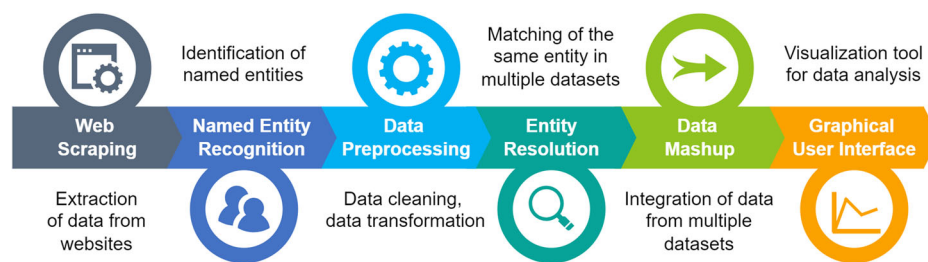

Fig. 1 | Schematic overview of the methodological framework. This figure shows the proposed framework. First, data relevant to FSNs were retrieved from different national competent authorities' websites using web scraping. Afterwards, named entity recognition was applied if the names of manufacturers and devices were not provided as structured fields. After data preprocessing, entity resolution was performed to match the same real-world items. Data integration was then performed using a mashup module, and a graphical user interface with several types of queries was build to present the generated harmonized and centralized database to users.

Table 1 | Number of Field Safety Notices (FSNs) retrieved for each official Member State public website included in the analysis, together with their subset for which it was possible to assign an EMDN code to the medical device product category to which they were referring

| Country	Issue date of the first retrieved FSN	Total number of retrieved FSN	Estimated number of issued FSN/month	Total number of FSN with assigned EMDN codes
Croatia	20 Apr 2012	1867	13.33	1341 (71.83%)
Czechia	02 Jun 2015	3463	33.62	2230 (64.40%)
Denmark	26 Jan 2012	4831	33.78	3560 (73.69%)
Estonia	01 Dec 2018	1103	18.08	330 (29.92%)
France	31 Oct 2007	3659	18.86	1830 (50.01%)
Germany	13 Dec 2004	15117	66.01	8439 (55.82%)
Greece	03 Jan 2007	1005	4.93	338 (33.63%)
Ireland	21 Mar 2002	7340	28.12	4149 (56.53%)
Italy	07 Jan 2009	9193	51.07	6622 (72.03%)
Latvia	21 Mar 2010	1794	10.87	1398 (77.93%)
The Netherlands	27 May 2015	4245	41.21	2774 (65.35%)
Poland	08 Mar 2007	4780	23.66	2561 (53.58%)
Portugal	11 Jan 2021	76	2.11	36 (47.37%)
Slovenia	03 Jan 2019	1505	25.08	788 (52.36%)
Spain	04 Jan 2018	4038	56.08	3090 (76.52%)
Sweden	22 Jun 2021	1020	34.00	726 (71.18%)
Total	-	65036		40212 (61.83%)
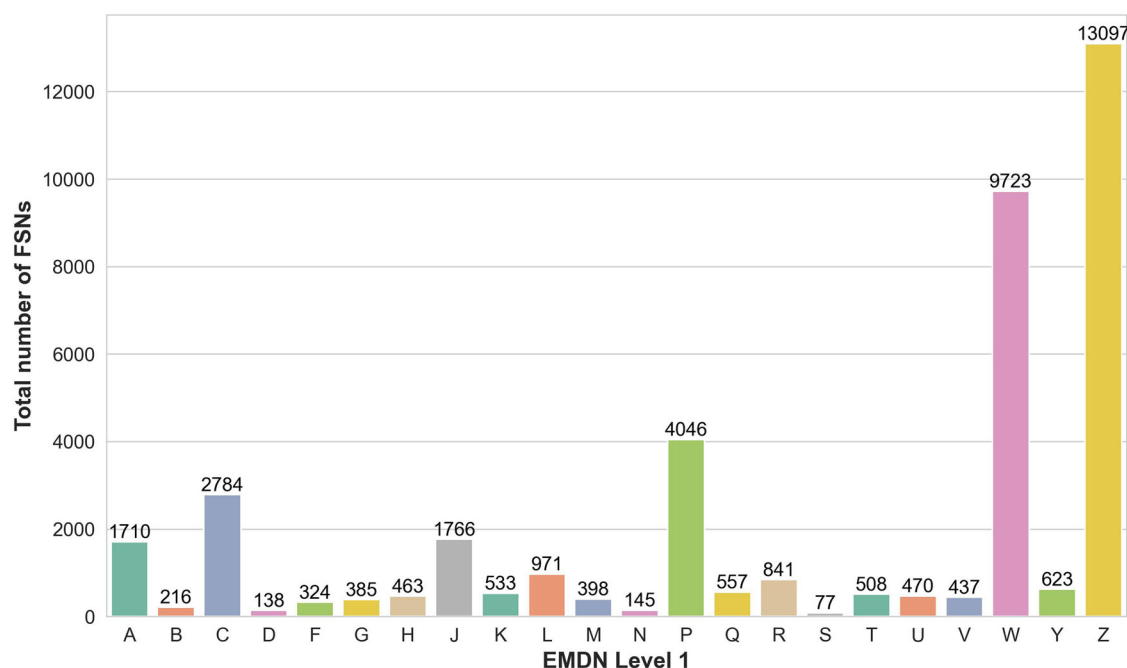
Fig. 2 | Distribution of EMDN categories at level 1 for FSNs in which the EMDN code was successfully assigned ($n = 40,212$). This plot shows the distribution of EMDN categories within the generated centralized dataset. (See Supplementary Table 2 for the term descriptions). The x-axis represents the different categories, while the y-axis represents the count of FSNs for each category. The figure pointed out that the first three categories with the highest number of FSNs were "Z—Medical equipment and related accessories, software and consumables" ($n = 13,097$; 32.57%), "W—In vitro diagnostic medical devices" ($n = 9723$; 24.18%), and "P— Implantable prosthetic and osteosynthesis devices" ($n = 4046$; 10.06%).

The proposed general validation strategy is described in the Methods section, and it is based on the availability of the original PDF files of the published FSNs, along with the HTML text in the scraped webpage. As Greece and Portugal provided only standardized PDF files of the FSNs, they needed to be excluded from the validation process, not respecting such constraints. As the validation results for Italy and the Netherlands were already described in detail in previous work[23,24], they will not be repeated here.

For the remaining countries, due to no common reporting forms, multiple languages, alphabets and different nomenclatures among Member States, in absence of the availability of a gold standard providing in a structured way the name entities of interest (manufacturer and device

Table 2 | Performance of the Named Entity Recognition model on different named entities (i.e., Manufacturer and Device) considering the "same FSNs" (Jaccard similarity index = 1) as reference standard for validation, in those countries in which a country-specific approach was needed. Exact-match means perfect string correspondence with ground truth, while relaxed-match allows partial string overlap (see text for details)

Country	Number of retrieved "same FSNs"	Named entity	Precision		Recall		F1-score	
			Exact-match	Relaxed-match	Exact-match	Relaxed-match	Exact-match	Relaxed-match
Croatia	74	Manufacturer	0.96	1.0	1.0	1.0	0.98	1.0
		Device	0.24	0.97	1.0	1.0	0.39	0.99
Estonia	24	Manufacturer	0.29	1.0	1.0	1.0	0.45	1.0
		Device	0.21	1.0	1.0	1.0	0.34	1.0
France	99	Manufacturer	0.78	1.0	1.0	1.0	0.88	1.0
		Device	0.14	0.98	1.0	1.0	0.25	0.99
Germany	247	Manufacturer	0.77	1.0	1.0	1.0	0.87	1.0
		Device	0.26	0.98	1.0	1.0	0.42	0.99
Poland	87	Manufacturer	0.74	1.0	1.0	1.0	0.85	1.0
		Device	0.15	1.0	1.0	1.0	0.26	1.0

Table 3 | Performance of the Named Entity Recognition model on different named entities (i.e., Manufacturer and Device) considering the "potentially equal FSNs" (Jaccard similarity index ≥0.6, cosine similarity ≥0.9, and fuzzy string matching ≥60) as reference standard for validation purposes in those countries in which a country-specific approach was needed. Exact-match means perfect string correspondence with ground truth, while Relaxed-match allows partial string overlap (see text for details)

Country	Number of retrieved "potentially equal FSNs"	Named entity	Precision		Recall		F1-score	
			Exact-match	Relaxed-match	Exact-match	Relaxed-match	Exact-match	Relaxed-match
Croatia	475	Manufacturer	0.85	1.0	1.0	1.0	0.92	1.0
		Device	0.18	0.99	1.0	1.0	0.30	0.99
Estonia	221	Manufacturer	0.26	1.0	1.0	1.0	0.41	1.0
		Device	0.20	1.0	1.0	1.0	0.34	1.0
France	607	Manufacturer	0.70	1.0	1.0	1.0	0.82	1.0
		Device	0.12	0.98	1.0	1.0	0.22	0.99
Germany	1764	Manufacturer	0.76	1.0	1.0	1.0	0.86	1.0
		Device	0.31	0.99	1.0	1.0	0.47	0.99
Poland	732	Manufacturer	0.64	1.0	1.0	1.0	0.78	1.0
		Device	0.12	0.99	1.0	1.0	0.22	1.0

names) as well as the corresponding EMDN codes, an alternative approach was proposed to derive a silver standard for comparison. This was possible by exploiting the previous results obtained from the Italian data, in which the EMDN codes were successfully assigned to the Italian FSNs with already identified named entities[23]: if the same FSN was found published both in Italy and in another country (i.e., the manufacturer issues the same FSN to all the national competent authorities where the device is on the market), then the already identified entities were compared with those obtained from the NER subtask. In addition, the known assigned EMDN codes, at different levels of its hierarchical tree-structure, were compared to the output of the ER process.

To retrieve such corresponding documents for comparison, the "same FSNs" issued in different countries from the same manufacturer for the same device and, most importantly, for the same reported problem needed to be identified. As different problems could arise over time, thus generating different FSNs for the same device, we firstly hypothesized to capture the same FSN by comparing the accompanying PDF document, that should have been identical except for the source language. This hypothesis was empirically not confirmed, due to possible discrepancies in the issue dates in the different countries, additional numbers, or pages. To overcome this limit, a certain range of possible discrepancies between two FSNs was set to

define "potentially equal FSNs", to be used in the validation process (see 'Methods' for more details).

For the NER subtask validation, in Table 2 the results of the performance in correctly identifying the Manufacturer and Device named entities, considering as reference those FSNs defined as the "same" published in Italy and in each specific country for which a country-specific approach was needed (see 'Methods' for more details), are presented, both considering the exact-match or the relaxed-match evaluation metrics. While the exact-match metrics varied significantly across countries and entity types, all relaxed-match metrics were higher than 0.9. As these results could be affected by a bias due to the small number of available documents retrieved as "the same FSNs", a second test was performed based on the "potentially equal FSNs". The previously achieved performance was confirmed by the measured high values of the relaxed-match evaluation metrics, as shown in Table 3.

To validate the ER subtask, the "same FSNs" published in Italy with an assigned EMDN code (i.e., the reference for comparison) and in each specific country were retrieved. A first test was based on comparing the EMDN codes for those FSNs by counting the EMDN codes assigned by ER for the specific country that matched with the Italian EMDN code, at different levels of description in the EMDN hierarchical tree-structure.

Table 4 | Performance of the Entity Resolution model in attributing an EMDN code to the medical device category relevant to the "same FSNs" that matched with the Italian EMDN code, assumed as reference for that device, at different levels of the EMDN hierarchical tree up to the fourth (see text for details)

Country	Number of "same FSNs" with non-empty EMDN	EMDN level 1	EMDN level 2	EMDN level 3	EMDN level 4
Croatia	55	98.18%	98.18%	92.73%	83.64%
Czechia	87	95.40%	91.95%	87.36%	79.31%
Denmark	160	95.62%	93.75%	90.0%	80.62%
Estonia	9	100.0%	100.0%	100.0%	88.89%
France	69	94.20%	94.20%	89.86%	78.26%
Germany	170	94.71%	92.35%	87.65%	80.0%
Ireland	10	90.0%	70.0%	70.0%	60.0%
Latvia	38	97.37%	94.74%	94.74%	81.58%
Poland	60	91.67%	90.0%	83.33%	76.67%
Slovenia	42	97.62%	97.62%	97.62%	90.48%
Spain	106	94.34%	92.45%	89.62%	83.96%
Sweden	41	100.0%	100.0%	95.12%	87.80%

Table 5 | Performance of the Entity Resolution model in attributing an EMDN code to the medical device category relevant to the "potentially equal FSNs" that matched with the Italian EMDN code, assumed as reference for that device, at different levels of the EMDN hierarchical tree up to the fourth (see text for details)

Country	Number of "potentially equal FSNs " with non-empty EMDN	EMDN level 1	EMDN level 2	EMDN level 3	EMDN level 4
Croatia	327	94.80%	94.19%	90.83%	79.82%
Czechia	530	96.23%	93.77%	90.94%	81.70%
Denmark	813	95.57%	92.99%	89.42%	79.83%
Estonia	79	87.34%	83.54%	82.28%	72.15%
France	404	93.81%	92.33%	86.88%	76.98%
Germany	1194	93.97%	91.54%	87.44%	78.81%
Ireland	53	88.68%	84.91%	77.36%	64.15%
Latvia	259	96.14%	94.59%	91.51%	85.71%
Poland	434	94.93%	91.47%	86.18%	73.73%
Slovenia	229	96.51%	94.32%	90.83%	82.10%
Spain	659	96.05%	94.23%	89.07%	80.73%
Sweden	232	98.71%	95.69%	93.97%	85.34%

As reported in Table 4, a high percentage of success (median [25th; 75th]: 81.1% [78.5%; 86.8%]) even when considering the fourth level was found, thus indicating very good performance. Again, as these results could be biased due to the limited number of available FSNs, a second test based on the retrieved "potentially equal FSNs" was performed. Relevant results are reported in Table 5, where similar results for matched EMDN codes up to the fourth level (79.8% [74.5%; 82.0%]) were present, thus indicating that the proposed methodology was still able to assign the proper EMDN codes of the corresponding device category for these "potentially equal FSNs", regardless of the unavailability or incompleteness of the datasets, thus confirming the previously observed very good performance.

Discussion

The proposed extended framework represents the first attempt to use web scraping and NLP to automatically collect and classify, based on the standard EMDN nomenclature, publicly available unstructured FSN data to create a structured global database from which, through queries focused on EMDN code, manufacturer or device names, aggregated data are made available to the user to allow trends visualization and further post-processing, with possible use in proactive PMS[26] and signal detection.

Considering the actual unavailability of EUDAMED for vigilance and market surveillance, and its future limitations due to not inclusion of historical data, this work shows that it is potentially possible to retrieve and aggregate such historical FSN data, despite it being provided by the relevant competent authorities in a fragmented manner, in different languages according to the Member State, without a common reporting form or nomenclature. To do so, country-specific strategies were developed for web scraping purposes, as well as for recognizing entities of interest (Manufacturer and Device names) in the unstructured text, with very high accuracy when using a relaxed-approach, capable of not restricting the research to a very specific string so to take into account the variations in names in different countries. Standardizing and categorizing historical FSN data is essential for providing consistency in long-term data analysis, allowing for the seamless integration of historical data with current or future standardized data, and enabling meaningful comparisons across extensive historical timelines to identify long-term trends and patterns. Furthermore, the inclusion and analysis of such data, especially from similar devices (i.e., within the same EMDN categories), could serve several purposes for manufacturers, such as risk management, improvement of device safety and performance, and regulatory compliance.

We noticed that only 16/27 EU competent national authorities are making publicly available the FSNs through their website, thus evidencing in general a transparency problem[9] with the impossibility for manufacturers to manually dig for potentially useful information for proactive PMS for 11 Member States.

Among all countries included in the analysis, the fact that the highest number of FSNs were retrieved from Germany could be explained by the spanned timeframe of publicly available data (since 2004), and by the fact that Germany holds the largest share in the European medical device market[1]. The availability of a centralized database could be beneficial to all EU competent authorities, as well as device manufacturers, by expanding the vision outside what's happening in its own country and specific market, thus potentially fostering greater transparency and cooperation towards more harmonized regulatory standards and practices, and a more effective and proactive PMS.

By leveraging the available resources, together with specific NLP approaches, the developed framework allowed to assign the proper EMDN code relevant to the category of the device in the retrieved FSNs, even when the information was partially missing or not properly structured. The EMDN category with the highest number of FSNs was the "Z—Medical equipment and related accessories, software, and consumables", followed by "W—In vitro diagnostic medical devices". Indeed, category Z includes a wide range of products, from bioimaging and radiotherapy instruments to instruments for functional explorations and therapeutic interventions, software, and non-specific consumables for diagnostic instruments, that generated a high volume of FSNs, suggesting the need to put particular attention in the certification process and PMS for such devices. The numbers associated with category W underline the need to exercise higher scrutiny, vigilance, and PMS for these products, as now regulated by the IVDR.

When comparing the EU system and available PMS data with the US, where the regulatory oversight is managed by the Food and Drug Administration (FDA), the European complexity and heterogeneity strikingly appear. The FDA has established several structured databases and standardized information collection systems to ensure the safety and effectiveness of medical devices, two of which are relevant to PMS: the Manufacturer and User Facility Device Experience (MAUDE) and Recall database[27]. These

databases are structured, with the possibility of several different querying and direct data download, which makes these data actionable, as testified by several commercial software in the US market that ingest such data and generate further value from its processing. However, comparing, matching or aggregating safety information from the EU and the US (and from other world markets like China, Russia, Brazil, the United Kingdom, Canada, Singapore, South Korea, Japan) remains a challenging task due to several key factors, including differences in device nomenclature, regulatory frameworks, and reporting awareness[28]. However, our proposed framework has the potential to be applied to these databases by replacing the first step of web scraping with direct data download, thereby taking a preliminary step toward exploring possible regulatory harmonization.

The spontaneous reporting systems (i.e., the FDA MAUDE, future EUDAMED), are crucial for safety signal detection, allowing the identification and analysis of potential adverse events associated with medical products. The FDA has issued guidance to establish a process for identifying and assessing emerging signals defined as new information about medical devices that may impact patient management or the benefit-risk profile, but the guidance did not include specific methods for these evaluations[29]. Although our tool currently does not directly identify possible anomalies, the richness of the generated dataset serves as the foundational starting point for further analysis. Instead of interrogating 16 different national databases, utilizing only one centralized database with possible queries about manufacturer, device, or EMDN, and obtaining results in a structured format could allow for a better understanding of the medical device market by moving from the currently narrow single competent authority perspective to a more global vision. This holistic view significantly simplifies the data retrieval process and helps in identifying patterns, trends and insights that may be obscured when data is dispersed across multiple national databases.

While standardization and classification in reporting could ensure consistency in how PMS data is presented to users, some inherent limitations of passive surveillance systems could still be present, and make cross-country comparisons (in particular among different markets – i.e., USA and EU) difficult. For instance, it may be difficult to determine the root causes of reported problems due to limited information and the absence of access to the actual devices. Also, the lack of a denominator indicating the number of people exposed to the devices makes the calculation of the incidence rate unfeasible; therefore, it is difficult to identify outliers devices with higher incidence of safety problems[30]. As publication of FSN is inherently delegated to the manufacturer itself, and thus connected to its propensity to properly react, this cannot be standardized, and maybe affected by external factors. Overreporting may be driven by media coverage, where medical devices with well-known adverse event problems are more likely to be exposed, thus leading to an excess of reports that do not necessarily indicate a true increase in the risk for such devices. On the other hand, underreporting may occur due to several reasons, such as lack of awareness of the requirement for reporting, uncertainty about the reasons behind the problems after interal assessment, and discretion exercised by manufacturers due to the absence of a standardized reporting threshold[18,30]. The degree to which these reporting biases are present could significantly differ from one country to another, and also the extent of these biases remains unclear. For instance, the tendency of overreporting may be more pronounced in countries where certain devices are extensively used, or where national clinical practices involve more detailed monitoring, or in countries with stringent regulatory oversight, heightened public awareness, and legal repercussions. Conversely, in countries where reporting is less emphasized, underreporting might be more prevalent. This disparity in reporting behavior, stemming from differences in the reporting requirements, allocating responsibilities among different stakeholders, and balancing central and regional control[2], makes global analysis to assess the safety and effectiveness of devices complicated even when data is formatted in a standardized manner, thus constituting an inherent limitations of passive surveillance systems. However, despite such challenges in data quality, meaningful analysis could still feasible with careful consideration. For instance, these issues could be mitigated by complementing the original data also using external sources, such as clinical registries, to achieve a more accurate and robust analysis by using all the possible information to maximize the potential of identifying outliers for a certain category of medical device, both in reporting behaviour and in relevant risk for the patient.

Our proposed framework is characterized by several strengths: 1) while the study focused on FSNs retrieved until 31 December 2023, it enables automatic retrieval of new FSNs over time to keep the database updated; 2) a Graphical User Interface (GUI) was developed to allow for querying the standardized and categorized data based on different user's needs (i.e., country, manufacturer, device, type, action, EMDN levels, time interval), retrieving both cumulative results, trends, as well as the specific FSNs made available as structured fields, with reference to the original web link (see Supplementary Fig. 1). This visualization tool provides many advantages for exploratory analysis, as a first step for potentially supporting the identification of patterns, trends, and anomalies. The integration of direct links to original sources not only enhances transparency and allows for real-time verification, but also facilitates a more thorough examination of the original information, thereby contributing to the overall reliability of the analysis.

While the current study has provided a valid solution to transform unstructured, incomplete and dispersed safety information into a standardized and centralized database, it is important to acknowledge some potential limitations, the first of them stemming from the partial availability and completeness of retrieved public data, as the analysis heavily relies on the availability of such data using a cross-learning approach. Also, the information about the relationship between companies, such as subsidiaries or parent companies, is absent. Therefore, querying for a specific manufacturer's name will return only notices pertinent to that specific manufacturer, and not those of possible company's subsidiaries. Also, the feasibility of attributing an EMDN code representing the most suitable category for the devices in the retrieved FSN was assessed, with median results around 60% and mainly based on the completeness and quality of publicly available data. These results could constitute the reference of comparison for further algorithm improvements. While this limitation could affect the interpretation of possible extracted trends based on the EMDN code as representing only a subsample of the available data, this information could still be valuable when considering that nowadays every national competent authority refers only to FSN in its own country which, for a specific device category, is largely underrepresenting the total number of FSN in all European market. Moreover, while the query of the generated database focused on the EMDN code would miss part of the potential FSNs, other types of queries, such as that for a specific manufacturer or for a specific device, will not be affected as relating to the NER subtask, thus allowing to derive specific trends useful for analyzing the corresponding risk profile over time. Finally, the possibility of searching directly for a specific device model, as well as querying for specific causes of malfunction relevant to a category of devices is still missing. The retrieval of such related information would require a manual screening of the results. Future development in analyzing the content of the PDF files to extract other important information, such as models and causes of malfunctions, will need to address this limitation, dealing with varying formats, languages and styles of PDF files.

Considering the findings from the current study, several recommendations highlighting current needs in the regulatory process could be proposed. Firstly, a standardization of the reporting format for the information made publicly available online by Member States could help and enhance the retrieval process and relevant exploitation of such data for multiple purposes. Second, a consistent requirement for reporting across various countries should be defined to avoid inconsistency between regions, which could lead to different levels of data completeness. Third, FSNs from manufacturers should incorporate more crucial information about the reported incident, such as unique device identifiers or the International Medical Device Regulators Forum code for Adverse Event Terminology[31], to improve the traceability of devices and the potential utility of the retrieved information. In this way, the overall efficacy and effectiveness of the developed framework would be increased by providing other ways of

Table 6 | Relevant information included in the Dataset of Devices (DoD) by standardizing information in the available original list of devices

Country	Original Variable Name	Modified Variable Name in the DoD
Italy	Manufacturer/assembler	Manufacturer
	Commercial name	Device
	Catalogue code	Model
	Progressive DM/ASS	Device ID
	CND (the Italian nomenclature for medical devices)	EMDN
Portugal	Manufacturer	Manufacturer
	Model	Device
	Brand	
	NPDM (the Portuguese nomenclature for medical devices)	EMDN

aggregating data and reporting results based on severity of the issue. While the recommendations provided are based on the current study focusing on the EU Member States, extending the analysis to non-EU countries could help ensure that the recommendations are reflective of a broader scenario. Such expansion could also assist in aligning different regulatory standards and enhancing possible global harmonization.

The proposed approach could constitute a valid solution to have access to historical data from national authorities while EUDAMED would start to be populated with more standardized and complete information. The multifaceted utility of the developed approach could provide a reliable foundation for well-informed decision-making to regulatory bodies, especially Expert Panels for assessing high-risk medical devices, through a systematic approach, as well as to manufacturers in their quest for data to comply to the proactive evaluation of their devices throughout their market lifecycle as part of the PMS, thus contributing to a collective advancement toward evidence-based practices in the field of regulatory science.

Methods
This section presents a detailed and comprehensive description of the proposed methodological steps involving NLP techniques, as depicted in Fig. 1. Briefly, data relevant to FSNs were retrieved from different national competent authorities' websites using web scraping. Afterwards, if the names of manufacturers and devices were not available in the extracted text as structured fields, NER was applied. After data preprocessing, ER was performed to match the same real world items in different datasets to be able to associate the device in the FSN with the respective EMDN code representing its category. Data integration was then performed using a mashup module and presented to the final users through a graphical interface that allows several types of queries.

Finally, the description of the validation strategy adopted to evaluate the performance of the proposed framework for the different countries is presented.

Data sources
The deployment of the proposed framework requires the availability of two crucial types of dataset sources: the Dataset of Notices (DoN) and the Dataset of Devices (DoD). The former consists of all FSNs retrieved from the national competent authority website as detailed in Supplemental Table 1. The latter contains information about devices available on a national market, and only a few EU countries make it available: Italy (including 1,703,175 devices, as of January 2023), Portugal (1,346,977 devices, as of January 2023), and France (77,502 devices, but outdated). In addition of being partial and outdated, as the French list of devices referred to the Global Medical Device Nomenclature[32], it was not considered further. As a result, with the term

DoD, we refer to the set composed by the Italian and Portuguese list of devices in which the EMDN code of the corresponding category is provided.

Table 6 reports the original variables for the Italian and Portuguese list of devices and their modified names in the DoD after preprocessing; in particular, the field *Device* in the Portuguese DoD was populated by merging the original fields *Brand* and *Model*, if not equal, due to possible data inconsistency and imprecision found in the published list.

Web scraping
Web scraping, also known as web crawling, is the procedure of automatically extracting data from websites and transforming unstructured web data into a structured format for presentation or storage[33,34]. In the proposed approach, web scraping was necessary as the information about FSNs was available only as HTML text on the competent authorities' websites. To automate web scraping on the Chrome browser, the Python library *Selenium* (https://pypi.org/project/selenium/) and *ChromeDriver* (https://chromedriver.chromium.org/) were used.

Due to the high heterogeneity of the scraped websites, country-specific approaches were developed to effectively extract structured information about the FSNs, involving the analysis of the unique layout and content organization of each website as well as developing customized extraction methods. The heterogeneity was reflected not only in the different formats and styles of reporting but also in the level of precision and quantity of the publicly available information. The extracted data was used to populate the following fields of the DoN, if the corresponding information was available:

Title: webpage title that contains information about the manufacturer and device.
Manufacturer: Name of the manufacturer that reported the incident.
Device: device name.
Model: device model.
Category: The device category could be a broad category or a specific identifier that can be linked to any device nomenclature.
Device ID: device identifier provided by the country.
Description: description of the problem reported.
Date: date of receipt of reports.
URL: web address of the specific FSN page.

Named Entity Recognition (NER)
NER, a subtask of NLP, aims to recognize mentions of rigid designators in text belonging to predefined semantic types, such as person, geographical location, organization, etc[35]. NER has been proven crucial in various NLP applications such as information retrieval[36,37], automatic text summarization[38], question answering[39], machine translation[40], etc.

The four main streams of techniques applied in NER are: rule-based[41], unsupervised learning[42], feature-based supervised learning[43], and deep learning (DL)-based approaches[44]. In recent years, DL-based approaches have become dominant and achieved state-of-the-art results due to their ability to automatically discover complex and intricate features, their effectiveness in learning, and the possibility of training in an end-to-end approach. In particular, the Transformer[45], the first transduction model, relies solely on self-attention and has demonstrated impressive effectiveness across different NLP tasks. Self-attention is an attention mechanism relating different positions of a single sequence to compute a representation of the sequence and effectively capture long-term dependencies. Another key element in the development of these models involves pre-training on a large corpus, followed by fine-tuning on a limited labeled dataset for the target, to address the lack of an adequate large dataset by transferring the knowledge gained from the pre-trained model to the new one[46].

The required named entities to successfully deploy the proposed framework were *Manufacturer* and *Device*. As these entities were not presented as structured fields in the FSNs from certain countries, NER was required to identify and extract them within the text of the FSN, typically in the field *Title*. Due to the country-specific differences, several approaches were adopted to obtain the best results by leveraging the available information on a case-by-case basis, using rule-based, DL-based, or PDF parsing techniques.

The DL-based approach was applied to Dutch FSNs by fine-tuning the pre-trained transformer model XLM-RoBERTa, a multilingual masked language model pre-trained on text in 100 languages with a massive dataset[47]. More specifically, the model was fine-tuned on the NER task using Dutch FSNs in which the information about *Manufacturer* and *Device* was already identified, to detect such information in other FSNs through the field *Title*[24]. As XLM-RoBERTa was trained from multiple languages, the same fine-tuned model was also applied to extract *Manufacturer* from the Estonian FSNs, where a proper training dataset was unavailable. Despite this, the fine-tuned model performed well because the style of the field *Title* in Estonian FSNs was similar to that in Dutch FSNs.

The PDF parsing was only applied for FSNs from Greece, as these documents have a standardized PDF format that shares a uniform structure, ensuring that the needed information is consistently located in the same position of the document.

On the other hand, FSNs from all the other countries that necessitated NER were accompanied by PDF files redacted by manufacturers, possibly including translations into the respective national languages. As a result, country-specific rule-based approaches, not requiring a proper training dataset, were applied for the remaining countries requiring the NER: Croatia, Estonia (to extract *Device* only), France, Germany, Poland, and Portugal.

For other countries with already available structured information about manufacturers and devices, such as Czechia, Denmark, Ireland, Italy, Latvia, Slovenia, Spain and Sweden, the NER step was omitted.

Data preprocessing

As all variables in the DoN and DoD are string variables, except *Date* in the DoN, proper text processing techniques were applied to address poor quality issues, such as text transliteration, using the Python library *Unidecode* (https://pypi.org/project/Unidecode/), removal of brackets, extra spaces and punctuations, and lowercasing. The company name parsing, aiming at providing cleaned names by stripping away terms referring to organization type, was applied only to the field *Manufacturer* using the Python package *cleanco* (https://github.com/psolin/cleanco).

Entity Resolution (ER)

ER describes the problem of extracting, matching and resolving entity mentions in structured and unstructured data, thus identifying items in multiple data that refer to the same real world entity[48]. In our scenario, the identification of the same device in the DoD and DoN was crucial to associate each device in the latter with the corresponding EMDN code reported only in the former. As not all countries provided the list of devices with information about the EMDN code publicly available, the following possible country-specific situations could emerge:

Case 1: the national list of devices was available and regularly updated, i.e., Italy and Portugal.

Case 2: a partial national list of devices was available but outdated.

Case 3: the national list of devices was not available.

In the first case, the national list of devices was utilized to perform ER[23]. In the second case, the national list of devices was initially used to perform ER, and then DoD was used to perform ER for those records lacking matches. Finally, for the last case, the DoD was considered for the matching[24].

Then, the existence of a direct linkage, represented by the field *Device ID*, between the DoD and the DoN was checked. So, for each FSN in the DoN, if the field *Device ID* was present, the EMDN nomenclature code corresponding to the device category was automatically retrieved by finding the devices with the same *Device ID* in the DoD. Note that this approach was feasible only for about one third of the Italian FSNs, as they occasionally provide *Device ID*. For the other countries, the problem was tackled in two phases to avoid unnecessary computational complexity and higher computational time: 1) identify similar manufacturers in the DoN and DoD; 2) among the subset of records with identified similar manufacturers, identify similar devices.

In the first step, the field *Manufacturer* in the DoN and DoD was mapped into term frequency-inverse document frequency (TF-IDF)

representations[49], and then the cosine similarity was computed between two TF-IDF vectors[50]: they were considered similar if the similarity score was equal to or higher than 0.90. In the second step, the comparison of both the fields *Device* and *Model* was performed to evaluate device similarity using approximate string matching. In particular, fuzzy string matching[51] with a similarity score from 0 to 100 based on the Levenshtein distance was applied using the Python package *FuzzyWuzzy* (https://pypi.org/project/fuzzywuzzy/). As the field *Model* was not always present in DoN or DoD, a comparison based on this field was made only if it existed in both datasets. As a result, the similarity calculated for two devices, one in the DoD and one in the DoN, was defined as the highest value between the similarities calculated from the fields *Device* and *Model*. The minimum similarity threshold was set to 60, and the pair with the highest similarity score was identified as the corresponding one.

Note that the identification of similar manufacturers was crucial for the analysis and could dramatically impact the identification of similar devices, as devices with different manufacturers would not be compared. The use of TF-IDF representations with the cosine similarity ensured a robust approach to this purpose. At the same time, the adoption of fuzzy string matching allowed for the identification of similar devices, even in the presence of minor discrepancies. The conditional comparison based on the field *Model* provided flexibility in the matching process and leveraged the available information.

Data mashup and graphical user interface

A data mashup merges different homogenous or heterogeneous data sources into a unique content page[52]. Information within the DoN and DoD was combined to provide a comprehensive database to users through a GUI. The interface was developed using Flutter, built and open-sourced by Google, that combines a reactive framework with customizable widgets to create cross-platform applications using a single code base[53].

Validation strategy

To evaluate the proposed framework's performance, different tests were performed to validate the NER and the ER subtasks based on the "same FSNs" and "potentially equal FSNs" found in both Italy (i.e., assumed as the reference for comparison) and the specific country of interest. To ensure the correspondence between the two FSNs, their attached PDF files were analyzed to verify whether the same manufacturer published them for the same device and, most importantly, for the same reason. The similarity between the files was assessed by the basic Jaccard similarity index[54] evaluated on the numeric characters found within the files, thus avoiding the uncertainty related to the translation process. Therefore, if two FSNs published in different markets had a Jaccard similarity index equal to 1, it means they were referring to the "same FSN". Empirical observations revealed that, despite the correspondence between two FSNs, the calculated Jaccard similarity index might result <1 due to discrepancies in the dates, additional numbers, or pages in the respective PDF files. Therefore, FSNs that could be "potentially equal" were identified by loosening the condition on the Jaccard similarity index, and including additional conditions for manufacturer and device similarities, which were calculated using the cosine similarity and fuzzy string matching, respectively. As a result, two FSNs were considered "potentially equal" if Jaccard similarity index ≥ 0.6, cosine similarity ≥ 0.9, and fuzzy string matching ≥ 60[24].

For the NER validation, FSNs from the Netherlands had already been tested by obtaining favourable results, thanks to the possibility of creating a proper training dataset to fine-tune the pre-trained model[24]. For other countries requiring the NER step, the similarities between the identified named entities *Manufacturer* and *Device*, reported by these "same FSNs" and "potentially equal FSNs", were used as a measure of model performance. The comparison between identified entities and the ground truth (the Italian FSNs) was quantified by either exact-match or relaxed-match[35]. More specifically, with exact-match evaluation, a named entity was considered correctly recognized only if it exactly matches the ground truth. Precision, Recall and F1-score were computed based on the number of true positives (TP), false positives (FP), and false negatives (FN), defined as:

TP: entities recognized by NER and match ground truth.
FP: entities recognized by NER, but do not match ground truth.
FN: entities not recognized by NER, but annotated in the ground truth.

The relaxed-match evaluation considered instead a correct matching if there was an overlap between the identified entity with the ground truth. In our scenario, a named entity was considered as correctly recognized using the relaxed-match approach if there is an overlap with the ground truth regardless of its boundaries. As an example, the device names "sistema SpaceOAR" from an Italian FSN (ground truth) and "SpaceOAR and SpaceOAR Vue Systems" from a German FSN, would be considered as a correct match according to the relaxed-match approach, but not according to the exact one.

The validation process for the ER subtask was straightforward for the Italian FSNs because of the direct linkage between DoN and DoD through the field *Device ID*. All FSNs with a non-empty value for *Device ID* were used as the gold standard to evaluate the model's performance, as previously reported[23]. For other countries, the comparison of the EMDN code assigned by the ER subtask with the Italian reference at different levels of description in the EMDN hierarchical tree-structure for the subsets of the "same FSNs" and of the "potentially equal FSNs" was used as a measure of the model's performance.

Data availability
All data supporting the findings of this study are publicly available and retrievable from the national competent authorities' websites using web scraping techniques as described.

Code availability
The code developed for this study using Python (version 3.11.5), and the graphical user interface, built using Flutter, are restricted for access. However, results relevant to specific queries could be obtained by the authors upon reasonable request.

References
1. MedTech Europe. *The European Medical Technology Industry in Figures* (2022).
2. Kramer, D. B., Tan, Y. T., Sato, C. & Kesselheim, A. S. Postmarket surveillance of medical devices: a comparison of strategies in the US, EU, Japan, and China. *PLoS Med.* **10**, e1001519 (2013).
3. World Health Organization. *Guidance for Post-market Surveillance And Market Surveillance Of Medical Devices, Including In Vitro Diagnostics* (World Health Organization, 2021).
4. Berry, M. & Stanek, J. J. The PIP mammary prosthesis: a product recall study. *J. Plast. Reconstr. Aesthet. Surg.* **65**, 697–704 (2012).
5. Dieterich, M. et al. Ruptured poly-implant protheses breast implant after aesthetic breast augmentation: diagnosis, case management, and histologic evaluation. *Aesthet. Plast. Surg.* **37**, 91–94 (2013).
6. Smith, A. J., Dieppe, P., Vernon, K., Porter, M. & Blom, A. W. Failure rates of stemmed metal-on-metal hip replacements: analysis of data from the National Joint Registry of England and Wales. *Lancet* **379**, 1199–1204 (2012).
7. Howard, J. Balancing innovation and medical device regulation: the case of modern metal-on-metal hip replacements. *Med. Devices Evid. Res.* **9**, 267–275 (2016).
8. Hwang, T. J., Sokolov, E., Franklin, J. M. & Kesselheim, A. S. Comparison of rates of safety issues and reporting of trial outcomes for medical devices approved in the European Union and United States: cohort study. *BMJ* i3323. https://doi.org/10.1136/bmj.i3323 (2016)
9. Fraser, A. G. et al. The need for transparency of clinical evidence for medical devices in Europe. *Lancet* **392**, 521–530 (2018).
10. *Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and Repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA Relevance)Text with EEA Relevance* (2023).
11. *Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in Vitro Diagnostic Medical Devices and Repealing Directive 98/79/EC and Commission Decision 2010/227/EU (Text with EEA Relevance)Text with EEA Relevance* (2023).
12. Martelli, N. et al. New european regulation for medical devices: what is changing? *Cardiovasc. Interv. Radiol.* **42**, 1272–1278 (2019).
13. Pane, J. et al. EU postmarket surveillance plans for medical devices. *Pharmacoepidemiol. Drug Saf.* **28**, 1155–1165 (2019).
14. Iadanza, E., Cerofolini, S., Lombardo, C., Satta, F. & Gherardelli, M. Medical devices nomenclature systems: a scoping review. *Health Technol.* **11**, 681–692 (2021).
15. Franzò, M. et al. Does a medical device nomenclature suitable for all purposes exist? Twenty years of Italian experience with the CND and its adoption in EUDAMED at European level. (2020).
16. European Commission. Questions and Answers on in vitro diagnostics and the European Database on Medical Devices (EUDAMED) (2024).
17. Faraulo, F. & Griesinger, C. *Exploiting Globally Available Safety Information on Medical Devices to Support EU Market Surveillance/Vigilance: An Analysis of Available Data Sources and Their Systematic and Consistent Use*. https://doi.org/10.2760/006806 (2017).
18. Pane, J. et al. Challenges associated with the safety signal detection process for medical devices. *Med. Devices Evid. Res.* **14**, 43–57 (2021).
19. Lauriola, I., Lavelli, A. & Aiolli, F. An introduction to deep learning in natural language processing: models, techniques, and tools. *Neurocomputing* **470**, 443–456 (2022).
20. Everhart, A. O., Sen, S., Stern, A. D., Zhu, Y. & Karaca-Mandic, P. Association between regulatory submission characteristics and recalls of medical devices receiving 510(k) clearance. *JAMA* **329**, 144 (2023).
21. Wunnava, S., Miller, T. A., Narang, C., Nathan, M. & Bourgeois, F. T. US Food and Drug Administration Approval of high-risk cardiovascular devices for use in children and adolescents, 1977-2021. *JAMA* **328**, 580 (2022).
22. Fraser, A. G. et al. Improved clinical investigation and evaluation of high-risk medical devices: the rationale and objectives of CORE–MD (Coordinating Research and Evidence for Medical Devices). *EFORT Open Rev.* **6**, 839–849 (2021).
23. Ren, Y., Bertoldi, M., Fraser, A. G. & Caiani, E. G. Validation of CORE-MD PMS support tool: a novel strategy for aggregating information from notices of failures to support medical devices' post-market surveillance. *Ther. Innov. Regul. Sci.* **57**, 589–602 (2023).
24. Ren, Y. & Caiani, E. G. Development of a framework dealing with partial data unavailability and unstructuredness to support post-market surveillance. In *Proc. IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* 1–4. https://doi.org/10.1109/BHI58575.2023.10313402 (IEEE, 2023)
25. Pires, C., Duarte, D. & Cavaco, A. Analysis of medical device alerts issued by the Portuguese medicines agency: scoping the purpose of new regulatory recommendations. *Acta Méd. Port.* **34**, 201–208 (2021).
26. Kearney, B. & McDermott, O. Challenges faced by manufacturers with clinical evaluation under the new European Medical Device Regulations. *Cogent Eng.* **10**, 2261236 (2023).
27. Liebel, T. C., Daugherty, T., Kirsch, A., Omar, S. A. & Feuerstein, T. Analysis: using the FDA MAUDE and medical device recall databases to design better devices. *Biomed. Instrum. Technol.* **54**, 178–188 (2020).
28. Fink, M. & Akra, B. Comparison of the international regulations for medical devices–USA versus Europe. *Injury* **54**, 110908 (2023).

29. Food and Drug Administration. Public notification of emerging postmarket medical device signals ("emerging signals"). https://www.fda.gov/media/95125/download (2016).

30. Rajan, P. V., Kramer, D. B. & Kesselheim, A. S. Medical device postapproval safety monitoring: where does the United States stand? *Circ. Cardiovasc. Qual. Outcomes* **8**, 124–131 (2015).

31. IMDRF Adverse Event Terminology Working Group. *IMDRF Terminologies for Categorized Adverse Event Reporting (AER): Terms, Terminology Structure and Codes* (2020).

32. Anand, K., Veermaram, C., Saini, S. K. & Singh, B. K. Global medical device nomenclature: the concept for reducing device-related medical errors. *J. Young. Pharm.* **2**, 403–409 (2010).

33. Khder, M. Web scraping or web crawling: state of art, techniques, approaches and application. *Int. J. Adv. Soft Comput. Appl.* **13**, 145–168 (2021).

34. Mitchell, R. *Web Scraping with Python: Collecting More Data from the Modern Web* (O'Reilly Media, Inc., 2018).

35. Li, J., Sun, A., Han, J. & Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **34**, 50–70 (2022).

36. Guo, J., Xu, G., Cheng, X. & Li, H. Named entity recognition in query. In *Proc. 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 267–274. https://doi.org/10.1145/1571941.1571989 (ACM, 2009).

37. Petkova, D. & Croft, W. B. Proximity-based document representation for named entity retrieval. In *Proc. Sixteenth ACM Conference on Information and Knowledge Management* 731–740. https://doi.org/10.1145/1321440.1321542 (ACM, 2007).

38. Khademi, M. E. & Fakhredanesh, M. Persian Automatic Text Summarization Based on Named Entity Recognition. *Iran. J. Sci. Technol. Trans. Electr. Eng.* 1–12. https://doi.org/10.1007/s40998-020-00352-2 (2020).

39. Mollá, D., van Zaanen, M. & Smith, D. Named Entity Recognition for question answering. In *Proc. Australasian Language Technology Workshop* (eds Cavedon, L. & Zukerman, I.) 51–58 (2006).

40. Babych, B. & Hartley, A. Improving machine translation quality with automatic Named Entity Recognition. In *Proc. 7th International EAMT Workshop on MT And Other Language Technology Tools, Improving MT through Other Language Technology Tools, Resource and Tools for Building MT at EACL* (2003).

41. Zhang, S. & Elhadad, N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. *J. Biomed. Inform.* **46**, 1088–1098 (2013).

42. Etzioni, O. et al. Unsupervised named-entity extraction from the web: an experimental study. *Artif. Intell.* **165**, 91–134 (2005).

43. Settles, B. Biomedical Named Entity Recognition using conditional random fields and rich feature sets. In *Proc. International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)* (eds. Collier, N., Ruch, P. & Nazarenko, A.) 107–110 (COLING, 2004).

44. Huang, Z., Xu, W. & Yu, K. Bidirectional LSTM-CRF models for sequence tagging. Preprint at http://arxiv.org/abs/1508.01991 (2015).

45. Vaswani, A. et al. Attention is all you need. in *Advances in Neural Information Processing Systems* Vol. 30 (Curran Associates, Inc., 2017).

46. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies, Vol. 1 (Long and Short Papers), 4171–4186, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423 (Association for Computational Linguistics, 2019).

47. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451 (2020) Online. https://doi.org/10.18653/v1/2020.acl-main.747

48. Getoor, L. & Machanavajjhala, A. Entity resolution: theory, practice & open challenges. *Proc. VLDB Endow.* **5**, 2018–2019 (2012).

49. Cohen, W. W., Ravikumar, P. & Fienberg, S. E. A comparison of string distance metrics for name-matching tasks. *IIWeb* **3**, 73–78 (2003).

50. Hapke, H., Howard, C. & Lane, H. *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python* (Simon and Schuster, 2019).

51. Zadeh, L. A. Fuzzy logic. *Computer* **21**, 83–93 (1988).

52. Koschmider, A., Torres, V. & Pelechano, V. Elucidating the Mashup Hype: definition, challenges, methodical guide and tools for mashups. In *Proc. 2nd Workshop on Mashups, Enterprise Mashups and Lightweight Composition on the Web at WWW* 1–9 (2009).

53. Windmill, E. *Flutter in Action* (Simon and Schuster, 2020).

54. Costa, L. da F. Further generalizations of the Jaccard Index. Preprint at http://arxiv.org/abs/2110.09619 (2021).

Acknowledgements

Author contributions

Y.R. and E.G.C. contributed to the conceptualization and design of the study. Y.R. carried out the technical details. Y.R. and E.G.C. wrote the manuscript. All authors have read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01337-9.

Correspondence and requests for materials should be addressed to Enrico Gianluca Caiani.

Reprints and permissions information is available at http://www.nature.com/reprints

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