Development of a Benchmark Corpus for Medical Device Adverse Event Detection

Susmitha Wunnava[*,†], **David Harris**[†], **Florence T. Bourgeois**[†,‡], **Timothy A. Miller**[†,‡]
[*]Harvard-MIT Center for Regulatory Science, Harvard Medical School, Boston, MA, USA
[†]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA
[‡]Department of Pediatrics, Harvard Medical School, Boston, MA, USA
susmitha_wunnava@hms.harvard.edu
{david.harris, florence.bourgeois, timothy.miller}@childrens.harvard.edu

Abstract

The U.S. Food and Drug Administration (FDA) collects real-world adverse events, including device-associated deaths, injuries, and malfunctions, through passive reporting to the agency's Manufacturer and User Facility Device Experience (MAUDE) database. However, this system's full potential remains untapped given the extensive use of unstructured text in medical device adverse event reports and lack of FDA resources and expertise to properly analyze all available data. In this work, we focus on addressing this limitation through the development of an annotated benchmark corpus to support the design and development of state-of-the-art NLP approaches towards automatic extraction of device-related adverse event information from FDA Medical Device Adverse Event Reports. We develop a dataset of labeled medical device reports from a diverse set of high-risk device types, that can be used for supervised machine learning. We develop annotation guidelines and manually annotate for nine entity types. The resulting dataset contains 935 annotated adverse event reports, containing 12252 annotated spans across the nine entity types. The dataset developed in this work will be made publicly available upon publication.

Keywords: medical devices, adverse event, natural language processing

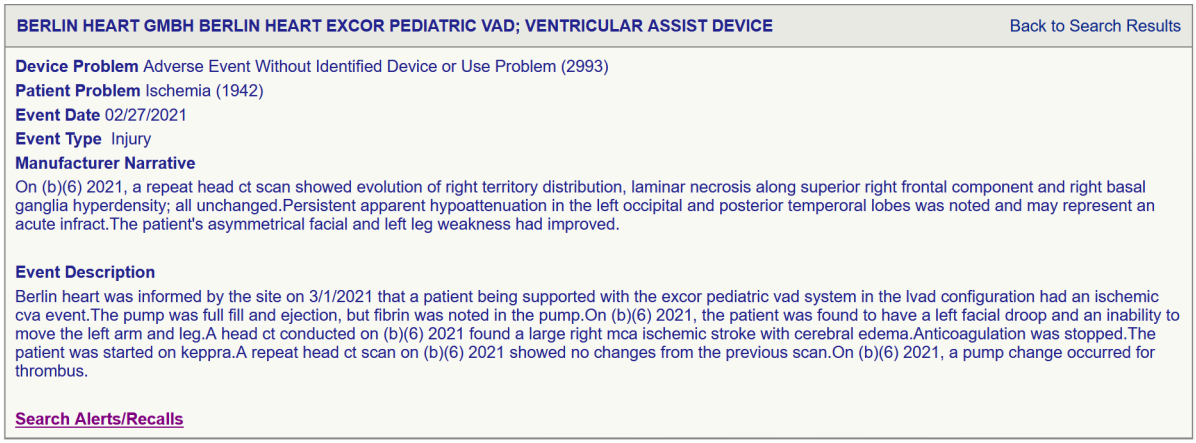
1. Introduction

Medical device adverse events are undesirable, unexpected events that occur during or after the use of a medical device. The United States Food and Drug Administration (FDA) uses a multifaceted approach to monitor the safety and effectiveness of marketed devices. The Manufacturer and User Facility Device Experience (MAUDE) database is a passive surveillance system and the FDA's primary post-market surveillance tool to capture real-world device-related deaths, serious injuries, and malfunctions. Other data sources include premarket clinical trials, and analysis of real-world data such as information from electronic health records (EHRs)(FDA, 2018). To support large-scale use of EHRs similar to what has been achieved in pharmacovigilance for drugs, the FDA introduced the unique device identification (UDI) system to enable identification and monitoring of devices. However, the program remains in an early phase given the slow adoption of UDIs by healthcare systems and the inability to efficiently identify and track device use (Kinard and McGiffert, 2020; Concato and Corrigan-Curay, 2022; Salazar and Redberg, 2020). Thus, the FDA continues to seek additional methodologies to support device surveillance activities.

Adverse event reporting enables FDA to take corrective action on problematic devices when safety concerns are identified (Levinson, 2009). The FDA's MAUDE database, which contains all device-related adverse event reports dating back to 1991, is publicly available on the FDA's website. The FDA makes the MAUDE database available to "provide patients and health care professionals with important information they can utilize to make more informed medical decisions." While spontaneous reports have limitations, most notably underreporting, many important safety signals have been initially identified using this information (Chung et al., 2020). The MAUDE database remains the primary mechanism for identifying safety signals for devices that require enforcement action, and was the most frequent source of device safety information leading to Medical Device Safety Warnings issued from 2011 to 2019 (Tau and Shepshelovich, 2020; Tomes, 2020). The database has also been used by investigators to assess the safety of specific devices across medical specialties (Coelho and Tampio, 2017; Tambyraja et al., 2005; Mahmoud et al., 2021).

Existing methods for safety signal detection from adverse event reports use statistical data mining methods such as disproportionality analysis, statistical process control, and sequential probability tests. These methods depend on structured data in the reports. While the data are rich in details regarding the specifics of the adverse event, most of it is free-form, unstructured text that requires processing and conversion into structured information for analysis. The few studies addressing device adverse event information extraction from text

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 240–245



BERLIN HEART GMBH BERLIN HEART EXCOR PEDIATRIC VAD; VENTRICULAR ASSIST DEVICE Back to Search Results

Device Problem Adverse Event Without Identified Device or Use Problem (2993)
Patient Problem Ischemia (1942)
Event Date 02/27/2021
Event Type Injury
Manufacturer Narrative
On (b)(6) 2021, a repeat head ct scan showed evolution of right territory distribution, laminar necrosis along superior right frontal component and right basal ganglia hyperdensity; all unchanged.Persistent apparent hypoattenuation in the left occipital and posterior temperoral lobes was noted and may represent an acute infract.The patient's asymmetrical facial and left leg weakness had improved.

Event Description
Berlin heart was informed by the site on 3/1/2021 that a patient being supported with the excor pediatric vad system in the lvad configuration had an ischemic cva event.The pump was full fill and ejection, but fibrin was noted in the pump.On (b)(6) 2021, the patient was found to have a left facial droop and an inability to move the left arm and leg.A head ct conducted on (b)(6) 2021 found a large right mca ischemic stroke with cerebral edema.Anticoagulation was stopped.The patient was started on keppra.A repeat head ct scan on (b)(6) 2021 showed no changes from the previous scan.On (b)(6) 2021, a pump change occurred for thrombus.

Search Alerts/Recalls

Figure 1: Example of adverse event report narrative from MAUDE

use rule-based methods consisting of user-defined rules for pattern matching to the raw text for information extraction (Alemzadeh et al., 2016; Penz et al., 2007). Rule-based systems are valued for their interpretability and ability to incorporate domain knowledge, but manually creating rules covering all possible information categories is labor intensive and requires high-level human expertise. The rules also apply to a small number of event types, making generalization expensive. Even fewer studies have applied supervised machine learning-based approaches towards device adverse event information extraction (Xie et al., 2018; Callahan et al., 2019). The few studies that automate information extraction have focused on specific device(s), limited data types, and a pre-determined set of basic adverse events.

Application of natural language processing (NLP) techniques to adverse event information extraction may provide an effective way to augment current approaches for post-marketing safety monitoring (Harpaz et al., 2014; Karimi et al., 2015). In this work we describe the development of a new dataset that will allow for fine-grained device-related adverse event information extraction, including important data types such as patient problems, device problems, reported patient outcomes and device information mentioned in the reports.

2. Background

In the context of drug safety surveillance and pharmacovigilance, many open challenges and shared tasks were conducted to assess and advance the state of the art in NLP for extraction of adverse drug events from clinical narratives (Uzuner et al., 2011; Henry et al., 2019; Jagannatha et al., 2019; Weissenbacher et al., 2019). Besides providing a venue for researchers to develop comparative systems on the same data and tasks, the challenges also made a variety of annotated adverse drug events datasets available for future researchers to learn and build on the state-of-the-art systems. On the other hand, NLP for medical device adverse event detection is unexplored. Research in this area is also impeded by a lack of curated medical device adverse event detection datasets for developing NLP models, and limited research in device signal detection methods from unstructured text. This work is therefore addressing an unmet need, since it is the first to describe the creation of a novel medical device adverse event detection NLP benchmark dataset, a data genre that is medical but different from adverse drug events, EHRs, and other biomedical text.

3. Data and Preparation

3.1. Data Source

We use the FDA's MAUDE database, a publicly accessible resource with over 10 million records on medical device safety. Each report has structured fields that capture patient problem and device problem codes, but also two unstructured fields – manufacturer narrative and adverse event description (Figure 1). The adverse event information in the MAUDE reports might not be well-captured by the structured data. Detailed information about the adverse event in the unstructured part of the reports may play a key role in identifying additional events and safety signals that are missed in the structured data (Figure 2).

Natural language annotation (i.e., tagging text such as patient problems, product problems, and patient outcomes) is a key step for training machine learning models to automatically extract adverse event information from large-scale corpora. This requires the following steps we detail below: 1) Identifying important information from the reports, defining the entities that reflect this information, and

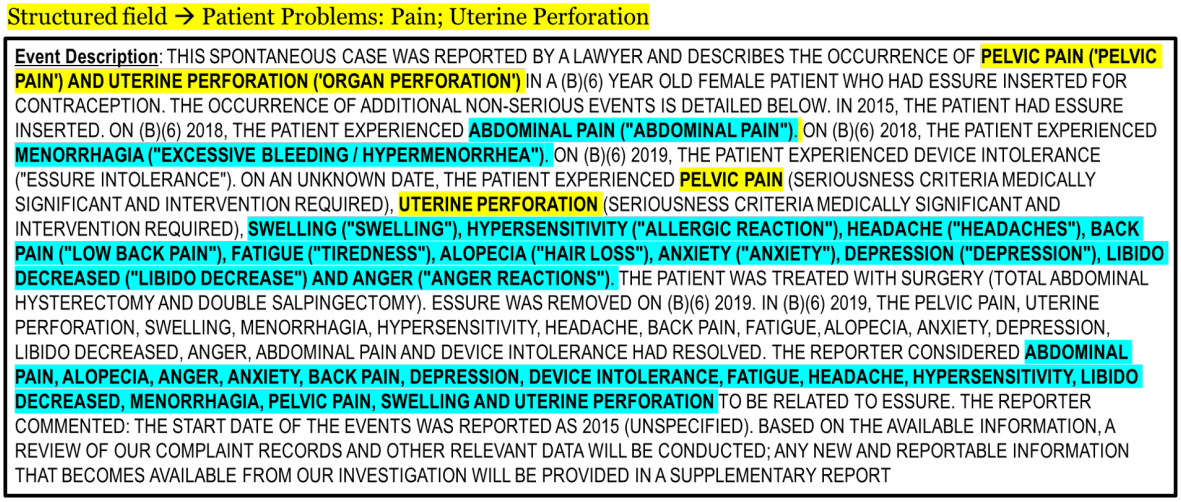


Figure 2: A sample report showing potential adverse events described in an unstructured "Event Description" field. Yellow highlighting indicates events that overlap with the structured data for the report, while blue indicates adverse events without corresponding structured data, and hence potential safety signals that were missed in the structured data.

creating annotation standards for annotators on the entities 2) Manually annotating a sample of reports with these entities. The key data extracted from the MAUDE database for this work is the unstructured device adverse event report narratives submitted to the FDA.

3.2. Dataset Creation

We create a large, diverse dataset of class III (high-risk) medical device adverse event reports from the FDA MAUDE database. Class III devices (e.g., pacemakers, blood vessel stents, cochlear implants) are implantable and/or life-sustaining devices that require premarket clinical safety and effectiveness data for approval. Any problems with these devices could lead to significant adverse outcomes for the patients. While class III devices constitute only 6.7% of all the devices, they make up more than 35.2% of device adverse event reports. Our sample includes reports of Class III devices with clinical safety and effectiveness data to maximize data usefulness for subsequent tasks. Devices are assigned to one of 491 "product categories", to ensure a representative sample of devices, we include all product categories with at least one adverse event report and include up to a maximum of six reports per product category. Finally, we select reports that include narrative descriptions of the adverse event.

4. Annotation Protocol

4.1. Named Entity Annotations

We created annotation guidelines for the following nine named entities:

1. **Manufacturer.** The manufacturer of a device
2. **Device.** Type of device. Common/Generic name of device implanted/explanted, used in the diagnosis, cure, mitigation, treatment, or prevention of disease. And/Or The Proprietary/Trade/Brand name of the medical device (as used in device labeling or in the catalog).
3. **Device Problem.** The product problems that were reported to the FDA if there was a concern about the quality, authenticity, performance, or safety of any medication or device.
4. **Treatment.** Treatment of event the patient received. Medications/Device Therapy/Surgery in response to the adverse event. Name(s) of the drugs/ devices/ therapies mentioned in the treatment of the adverse event.
5. **Procedure.** Medical procedure for/during which the device is used. A device is either implanted, explanted, replaced, or applied.
6. **Adverse Event.** Adverse side-effects of the device on the patient (a.k.a. patient problems). These are medical conditions, signs or symptoms resulting from use (implanting/explanting/application) of the device.
7. **Indication.** Medical sign or symptom that is the basis or direct cause of treatment. Alternatively, it can be described as a medical condition for which a device implant/explant has been prescribed in the past or present.
8. **Other Medical Conditions (OMC).** Medical signs, symptoms, or disease names that are

neither being actively treated (Indications) nor are they adverse side effects (patient problems) of using a device.

9. **Outcome.** Outcome associated with the adverse event for a patient.

Manufacturer and Device categories are important for device name normalization because device names can be difficult to parse and are not consistently used across reports. Adverse Events, Indications, and Other Medical Conditions are all essentially "medical problems," and so can appear superficially similar, but have crucial differences in how they should be interpreted. As a result, distinguishing them may make for a challenging task, but one that is vital for truly understanding the report. Treatment and Procedure categories are important to extract and distinguish since they are also superficially similar, yet have different interpretations, and which can also relate to the Outcome category. Overall, this set of categories attempts to capture the most important pieces of information in a report, potentially allowing for a variety of downstream applications.

4.2. Annotation Quality Control

We developed annotation guidelines and provided them to the annotators. We created rigorous annotation guidelines in an iterative process. The first draft guidelines included entity type descriptions, examples, and detailed instructions for challenging scenarios. Any ambiguous situations that arose during the annotation conflict resolution exercises were documented as examples for the guidelines. To ensure consistency and correctness, two annotators (a dedicated staff annotator with expertise in medical coding, and a regulatory scientist with experience working in the biomedical text mining domain) independently annotated a sample of reports after training, performed a check for agreement, and adjusted the annotation instructions to improve subsequent annotations. We use a web-based annotation tool called Label Studio (Maxim Tkachenko et al., 2023) to label the reports.

5. Annotation Results

To assess the quality of the manual annotations, we measure the inter-annotator agreement between the annotators using precision, recall, and F-measure, the performance metrics commonly applied in information retrieval tasks (Hripcsak and Rothschild, 2005). The two annotators labeled 130 reports with 2606 entity labels. The inter-annotator agreement yielded a precision of 0.71 and a recall of 0.68. In total, we labeled 935 adverse event reports with a total of 12252 labels spanning the nine entities. We further split the corpus into train/test, resulting in 822, and 113 reports, respectively. The training/test set split is stratified such that the test set consists of devices that were not part of the training set. This split allows for domain adaptation-style experiments where an evaluation can be broken down into performance on devices that have been previously seen versus those that are new. All annotations are stored as JSON files as well as in CONLL2003 (Sang and De Meulder, 2003) data format suitable for the named entity recognition task. We report statistics on the labels. The occurrence of each named entity type is provided in Table 1.

Entities	#Labels	Avg #Labels Per Report
Adverse Event	2993	4.17
Device	3410	3.99
Device Problem	964	2.52
Indication	385	2.01
Manufacturer	280	1.56
OMC	461	3.27
Outcome	70	1.46
Procedure	3144	4.05
Treatment	545	2.75

Table 1: Number of annotations per entity type in the dataset.

6. Conclusions

Medical devices are more complex than pharmaceutical drugs, and faulty design and manufacturing are often the cause of device-related injuries. New devices are less likely to have their safety established clinically before they are marketed. Effective postmarket surveillance of high-risk medical devices is vital for early warning about safety issues. Reportable adverse events suggest that the device may have caused or contributed to a death or serious injury. Spontaneous reporting of adverse events is an important surveillance tool. Natural Language Processing (NLP) techniques can provide an effective way of post-marketing safety monitoring, but large domain-specific corpora are needed to train and assess high-performance NLP models. This work aims to address this unmet need by developing a benchmark corpus and annotated dataset for training and evaluating NLP approaches to extract adverse event information from medical device safety reports and help in improving the medical device safety surveillance process. The dataset can also be used for other natural language processing tasks such as text classification or question answering, among others. The dataset developed in this work will be made publicly available upon publication.

7. Bibliographical References

Homa Alemzadeh, Jaishankar Raman, Nancy Leveson, Zbigniew Kalbarczyk, and Ravishankar K. Iyer. 2016. Adverse Events in Robotic Surgery: A Retrospective Study of 14 Years of FDA Data. *PLOS ONE*, 11(4):e0151470.

Alison Callahan, Jason A. Fries, Christopher Ré, James I. Huddleston, Nicholas J. Giori, Scott Delp, and Nigam H. Shah. 2019. Medical device surveillance with electronic health records. *npj Digital Medicine*, 2(1):94.

Gary Chung, Katherine Etter, and Andrew Yoo. 2020. Medical device active surveillance of spontaneous reports: a literature review of signal detection methods. *Pharmacoepidemiology and Drug Safety*, 29(4):369–379.

Daniel H. Coelho and Alex J. Tampio. 2017. The Utility of the MAUDE Database for Osseointegrated Auditory Implants. *Annals of Otology, Rhinology & Laryngology*, 126(1):61–66.

John Concato and Jacqueline Corrigan-Curay. 2022. Real-World Evidence — Where Are We Now? *New England Journal of Medicine*, 386(18):1680–1682.

FDA. 2018. FDA Medical Device Safety Action Plan: Protecting Patients, Promoting Public Health.

Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam H. Shah. 2014. Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art. *Drug safety : an international journal of medical toxicology and drug experience*, 37(10):777–790.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 27(1):3–12.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 12(3):296–298.

Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. 2019. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). *Drug safety*, 42(1):99–111.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015. Text and Data Mining Techniques in Adverse Drug Reaction Detection. *ACM Computing Surveys*, 47(4):1–39.

Madris Kinard and Lisa McGiffert. 2020. Medical Device Tracking—How It Is and How It Should Be. *JAMA Internal Medicine*.

Daniel R Levinson. 2009. Adverse event reporting for medical devices. *Office of Inspector General*.

Karim Mahmoud, Sreenivasulu Metikala, Kathryn M. O'Connor, and Daniel C. Farber. 2021. Adverse events related to total ankle replacement devices: an analysis of reports to the United States Food and Drug Administration. *International Orthopaedics*.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2023. Label Studio: Data labeling software. Original-date: 2019-06-19T02:00:44Z.

Janet F. E. Penz, Adam B. Wilcox, and John F. Hurdle. 2007. Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics*, 40(2):174–182.

James W. Salazar and Rita F. Redberg. 2020. Leading the Call for Reform of Medical Device Safety Surveillance. *JAMA Internal Medicine*, 180(2):179.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Rabindra R. Tambyraja, Michael A. Gutman, and Cliff A. Megerian. 2005. Cochlear Implant Complications: Utility of Federal Database in Systematic Analysis. *Archives of Otolaryngology–Head & Neck Surgery*, 131(3):245.

Noam Tau and Daniel Shepshelovich. 2020. Assessment of Data Sources That Support US Food and Drug Administration Medical Devices Safety Communications. *JAMA Internal Medicine*, 180(11):1420–1426.

Madris Tomes. 2020. Identification and Market Removal of Risky Medical Devices. *JAMA Internal Medicine*, 180(11):1426–1427.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics.

Jiaheng Xie, Xiao Liu, and Daniel Dajun Zeng. 2018. Mining e-cigarette adverse events in social media using Bi-LSTM recurrent neural network with word embedding representation. *Journal of the American Medical Informatics Association*, 25(1):72–80.