Received: 16 March 2021 | Revised: 9 December 2021 | Accepted: 13 December 2021

DOI: 10.1002/pds.5402

ORIGINAL ARTICLE

WILEY

Assessing adverse event reports of hysteroscopic sterilization device removal using natural language processing

Jialin Mao[1] | Art Sedrakyan[1] | Tianyi Sun[1] | Maryam Guiahi[2] | Scott Chudnoff[3] | Madris Kinard[4] | Stephen B. Johnson[5]

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA

[2]Department of Obstetrics and Gynecology, Division of Family Planning, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

[3]Department of Obstetrics and Gynecology, Stamford Hospital, Stamford, Connecticut, USA

[4]Device Events, York, Pennsylvania, USA

[5]Department of Population Health, New York University Langone Health, New York, New York, USA

Correspondence
Jialin Mao, Department of Population Health Sciences, Weill Cornell Medical College, 570 Lexington Ave, 10 Floor, New York, NY 10022, USA.
Email: jim2012@med.cornell.edu

Funding information
This study was supported by the Agency for Healthcare Research and Quality (Grant number 1R03HS026291-01, PI, Mao) and partially supported by the U.S. Food and Drug Administration (Grant number U01FD005478, PI, Sedrakyan). The funder had no influence on the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Abstract

Objective: To develop an annotation model to apply natural language processing (NLP) to device adverse event reports and implement the model to evaluate the most frequently experienced events among women reporting a sterilization device removal.

Methods: We included adverse event reports from the Manufacturer and User Facility Device Experience database from January 2005 to June 2018 related to device removal following hysteroscopic sterilization. We used an iterative process to develop an annotation model that extracts six categories of desired information and applied the annotation model to train an NLP algorithm. We assessed the model performance using positive predictive value (PPV, also known as precision), sensitivity (also known as recall), and F_1 score (a combined measure of PPV and sensitivity). Using extracted variables, we summarized the reporting source, the presence of prespecified and other patient and device events, additional sterilizations and other procedures performed, and time from implantation to removal.

Results: The overall F_1 score was 91.5% for labeled items and 93.9% for distinct events after excluding duplicates. A total of 16 535 reports of device removal were analyzed. The most frequently reported patient and device events were abdominal/pelvic/genital pain ($N = 13\ 166$, 79.6%) and device dislocation/migration ($N = 3180$, 19.2%), respectively. Of those reporting an additional sterilization procedure, the majority had a hysterectomy or salpingectomy ($N = 7932$). One-fifth of the cases that had device removal timing specified reported a removal after 7 years following implantation ($N = 2444/11\ 293$).

Conclusions: We present a roadmap to develop an annotation model for NLP to analyze device adverse event reports. The extracted information is informative and complements findings from previous research using administrative data.

KEYWORDS
female sterilization, medical device safety, research methods

Key Points

- Device adverse event reports can provide independent evidence when no other data is available or complement current evidence from claims data but are difficult to analyze. This study presents a roadmap for developing an annotation model for NLP to analyze device adverse

This study was presented as an abstract at the 2020 AcademyHealth Annual Meeting (Virtual).

event reports and demonstrates the utility of adverse event reports in device safety research and the challenges in working with this data source.

- Studies based on claims data could not provide accounts of adverse events associated with reoperations due to the data collection and diagnostic coding system. Device adverse event reports can be analyzed to determine the occurrence of patient events and device complications among cases that reported a device removal, shedding light on the nature of events associated with device removal.

Plain Language Summary

This study analyzed 16 535 adverse event reports from the Manufacturer and User Facility Device Experience database from January 2005 to June 2018 related to device removal following hysteroscopic sterilization. We developed an annotation model and applied it to train a natural language processing (NLP) algorithm to extracts six categories of desired information. Using information extracted by this method, we found that the most frequently reported patient and device events associated with hysteroscopic sterilization device removal were abdominal/pelvic/genital pain and device dislocation/migration. Of women reporting an additional sterilization procedure, the majority had a hysterectomy or salpingectomy. One-fifth of the cases that had device removal timing specified reported a removal after 7 years following implantation. This study helps better understand the nature of patient and device events associated with hysteroscopic sterilization device removal. It also presents a roadmap to develop an annotation model for NLP to analyze device adverse event reports.

1 | BACKGROUND

The hysteroscopic sterilization device had a promising safety profile in early clinical trials and was approved by the U.S. Food and Drug Administration (FDA) in 2005.[1,2] After wide commercialization, safety concerns related to the device emerged.[3] Following the advisory committee meeting in 2015, the FDA-mandated a box warning and a decision checklist to be included for device users in 2016.[4] The manufacturer decided to withdraw the device from the US market in 2018,[5] but hundreds of thousands of women are implanted with the device and continued review of patient outcomes is necessary.[6]

Previous studies based on administrative and claims data demonstrated that hysteroscopic sterilization was associated with an increased risk of reoperation when compared with traditional laparoscopic sterilization.[7–10] However, claims data are limited in providing detailed accounts of adverse events associated with reoperations due to the data collection and diagnostic coding system. The Manufacturer and User Facility Device Experience (MAUDE) database is a reporting system mandated by the FDA for postmarket device surveillance and collects device adverse event reports, including injuries and device malfunctions, from mandatory and voluntary reporters.[11] A previous study investigating the feasibility of using an application to collect voluntary reports showed that these reports contained specific patient and device events reported by women receiving the sterilization device.[12] Information extracted from adverse event reports can provide independent evidence when no other data is available or complement current evidence from claims data.

Due to the narrative nature of reports submitted to the MAUDE database, previous studies based on these reports have mostly relied on keyword search and manual reviews.[13–18] Keyword search requires a comprehensive knowledge of all possible variations of event reporting, which poses significant difficulties in study design. Manual review is labor-intensive, particularly if there is a high volume of adverse event reports. Natural language processing (NLP) has the capacity to process free-text structures and has been used to identify adverse drug events[19–21] and postoperative complications[22] from electronic health records (EHRs). Two recent studies reported the use of NLP to analyze MAUDE reports but lacked details on the development and implementation of the NLP system.[23,24] In this article, we present a roadmap for developing an annotation model for NLP to analyze MAUDE reports. We applied the developed tool to evaluate the most frequently reported patient and device events among patients who reported a device removal.

2 | METHODS

2.1 | Data source and overview of study design

Our data source was the MAUDE database, which houses reports submitted by mandatory, such as manufacturers, and voluntary reporters, such as patients. Full reports in the MAUDE database have patients' direct and indirect identifiers redacted and are publicly available. Reports were extracted from the MAUDE database using the keyword search feature and cleaned using Device Events.[25] Device Events is a tool that aggregates, normalizes, and cleans medical device reports for easier data extraction. We included 16 783 reports from January 2005 to June 2018 related to device removal following

hysteroscopic sterilization. Only reports related to device removal were included because our goal was to assess events related to sterilization device removal.

Since these adverse event reports have no structure and are different from layperson language or medical notes, human experts need to annotate text with labels to locate the information of interest. Labeled texts can then be used to train and validate the NLP algorithm. We used an iterative process to develop the annotation model and NLP (Figure 1). We first manually reviewed 50 randomly selected adverse event reports to determine the type of information that can be extracted and created an annotation model. We implemented the developed annotation rules for 100 reports and tested an NLP algorithm to assess and improve the annotation. This step was to ensure that the annotation model and software worked as intended and finalize the annotation model. Next, a two-stage approach of double annotation was adopted to improve and assess the inter-annotator agreement (IAA). We then annotated a larger corpus (N total = 1000) to train the NLP model, applied the trained NLP to the entire dataset, and classified the extracted events for the statistical analysis. This study was determined as exempt by Weill Cornell Medicine IRB and informed consent was waived.

2.2 | Annotation model specification and inter-annotator agreement

After iterations of review, annotation, NLP testing, and model revision, we established an annotation model with six primary categories of events (Table S1): (1) reporting sources, (2) medical confirmation or legal process related to the case, (3) patient events (abdominal/pelvic/genital pain, bleeding, menstrual disorder, allergy to metal, pregnancy, and others), (4) device events (organ perforation, device dislocation/migration, device embedment, device breakage, and device shape alteration), (5) implantation and removal timing, and (6) additional procedures performed. Three additional categories were used as modifiers of events: (1) device, (2) location where device events or procedures happened, and (3) degree of uncertainty. We developed a data dictionary and detailed annotation guidelines to assist human annotators when performing annotation (Figure S1). Initially, we intended to assess the most relevant patient and device events, and prespecified patient and device events of interest based on the endpoints in the FDA-mandated post-approval study for the sterilization device.[26] After the pilot implementation of the annotation model and NLP, we found a significant number of patient events that were not prespecified. Therefore, we decided to

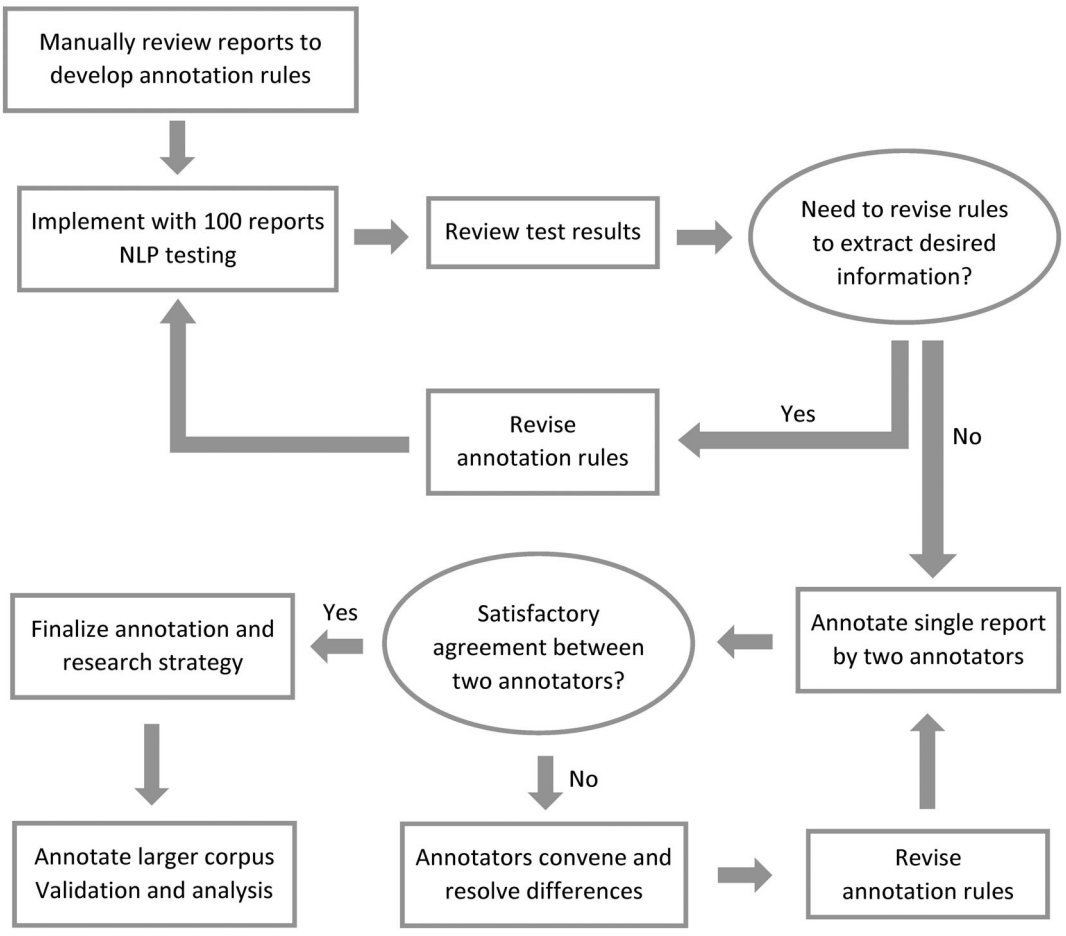


FIGURE 1 The process of annotation model and natural language processing (NLP) development

classify all output from NLP into subtypes of events to provide a comprehensive review. Patient event subtypes were defined based on the adverse event terminologies proposed by the International Medical Device Regulators Forum (IMDRF),[27] adapted to the current context. Annotation and adjudications were performed with a software tool that shows the different categories used to label portions of a text using different colors of highlighting (multi-document annotation environment).[28] To ensure high reliability when multiple annotators are involved, we used a two-stage process to improve the consistency in annotation and assessed the IAA (detailed in Data S1).

2.3 | Model implementation with NLP

We used named entity recognition (NER), which identifies the boundary and categories of entities (i.e., labeled events specified in the annotation model). The code used for NER was written in Python using the SpaCy library.[29] SpaCy allows users to train customized NER based on user-provided training data. The SpaCy model architecture is based on word embedding, which uses statistical predictions to classify words based on encoded word vectors, taking into account the sentence and document context. The model processes the annotated entities in a *BIO* annotation format, indicating which words *Begin* an entity or are *Inside* or *Outside* of any entity.[29] In order to train a NER, the user must provide an annotated input document (gold standard), which contains the category and textual span of each entity. We annotated 1000 reports to train and assess the performance of the annotation and NLP model. The reports were separated into two sets: one for training the algorithm ($n = 700$) and one for testing ($n = 300$). The initial 100 samples used for the annotation model development were used as training data. We trained the model for 1000 iterations with a dropout rate of 0.2 at each iteration. The dropout rate represents the proportion of features that are dropped in each iteration of training, which makes the model harder to memorize the training data. We then applied the trained NER model to the training and testing documents.

To assess the model performance, we compared every entity from the NER output with the gold standard generated from the annotation for all labels and the six desired categories, respectively. The model performance was assessed in the training and testing samples separately, and results from the testing samples were reported. We calculated true positive, false positive, and false negative, which were used to calculate the positive predictive value (PPV, also known as precision), sensitivity (also known as recall), and F_1 score (a combined measure of PPV and sensitivity). We used an exact definition and a lenient definition for model performance assessment (-Figure S2). True positive by the exact definition required that the labeled positions from NLP output and manual labeling must reach an absolute match. The lenient definition allowed labeled positions from NLP output that were overlapping the manual labeling to count as a match. Due to the presence of repeated mentioning of events in these reports, we further assessed the model performance after labeled items were classified into distinct events and repeated mentioning of the same events were eliminated.

$$\text{PPV} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}},$$

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}},$$

$$\text{F}_1\ \text{score} = 2 \times \frac{\text{PPV} \times \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}}.$$

2.4 | Statistical analysis

We applied the NLP to all reports and used a matcher to detect reports that aggregately reported events for multiple women. These reports were excluded from the final analysis due to their lack of clarity. We also used a matcher to identify reports that specified the use of other female intrauterine contraceptive devices (IUD). Because reported events may have been related to the IUD, we performed a sensitivity analysis to assess the impact of these reports on final summarized statistics by excluding them.

We determined the reporting source and reporter (see Figure S3 for examples) based on the extracted variable and used the reporter occupation collected in the MAUDE database to fill in missing values. If the reporter's occupation was not specified and the report was recorded as voluntary, the reporter was determined to be a consumer/patient. We further examined the numbers and proportions of reports that were confirmed by health providers and involved legal processes based on the reporting source and directly extracted variables. The report was considered confirmed by providers if the NLP detected the mentioning of medical confirmation or if the reporter was a healthcare provider. The report was determined to have involved legal process if the NLP detected the mentioning of litigation or if the reporter was a lawyer.

We summarized the presence of prespecified and other patient and device events, additional sterilizations and other procedures performed, and time from implantation to removal using descriptive analysis. We classified all predicted output from NLP into subtypes for further analysis based on the IMDRF adverse event terminology, adapted to the current context. We reviewed output that cannot be directly matched to these terminologies to make the classification. The number and proportion of cases reporting these subtypes of events were summarized to assess their prevalence among women who had a device removal. We determined the additional sterilization procedure based on the most invasive procedure for each patient. The invasiveness of the procedure was ranked in the following order (most to least invasive): hysterectomy, salpingectomy, tubal ligation, and re-implantation. The NER method has a probabilistic nature with some degree of errors. Therefore, we calculated the estimated numbers of patient and device events and additional sterilization procedures when adjusting for sensitivity and PPV. The NLP was performed using Python 3.5 and the statistical analysis was performed using SAS 9.3 (Cary, NC).

TABLE 1 Named entity recognition output from natural language processing compared with entities from the manual annotation in testing documents

	Labeled events			Labeled events					
	Exact definition			Lenient definition			Distinct events		
	PPV	Sensitivity	f score	PPV	Sensitivity	*f* score	PPV	Sensitivity	*f* score
All	0.887	0.876	0.881	0.920	0.910	0.915	0.960	0.918	0.939
By categories									
Source	0.978	0.954	0.966	0.978	0.954	0.966	0.984	0.950	0.967
Process	0.873	0.949	0.909	0.879	0.956	0.916	0.947	0.991	0.969
Patient events	0.875	0.856	0.866	0.927	0.910	0.918	0.959	0.921	0.940
Device events	0.794	0.834	0.813	0.810	0.849	0.829	0.936	0.842	0.886
Procedures	0.906	0.889	0.897	0.909	0.892	0.900	0.942	0.836	0.886
Timing	0.941	0.931	0.936	0.944	0.934	0.939	0.966	0.933	0.949

Abbreviations: PPV, positive predictive value.

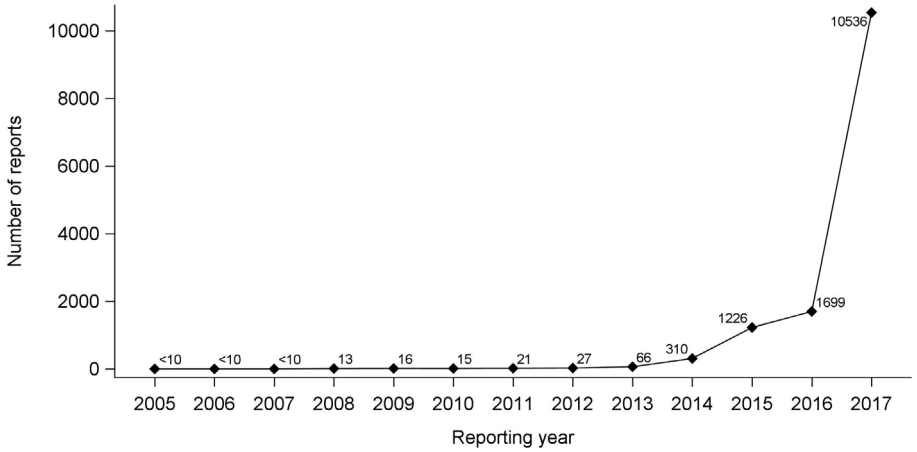


FIGURE 2 Trends in the number of hysteroscopic sterilization device removal reports over time

3 | RESULTS

3.1 | Model performance

Table 1 shows the performance of the model on the testing data. For all labels combined, the PPV and sensitivity of the NER in the testing documents were 87.8% and 87.5% by the exact definition and 90.8% and 90.3% by the lenient definition. When using the lenient definition, which allowed labeled positions from NLP output that overlapped the manual labeling to count as a match, the F_1 scores for label categories of reporting source, process, patient events, device events, procedures, and timing were 96.6%, 91.6%, 91.8%, 82.9%, 90.0%, and 93.9%, respectively. The performance of the NLP to capture patient events was noticeably higher when using the lenient definition, mainly because there were often many descriptive elements of patient events, such as the severity or frequency of events. When assessing the model performance by counting distinct events captured after eliminating repeated events, the overall PPV and sensitivity were 96.0% and 91.8%. The F_1 scores for label categories of reporting source, process, patient events, device events, procedures, and timing were 96.7%, 96.9%, 94.0%, 88.6%, 88.6%, and 94.9%, respectively. The performance of the NLP to capture device events was affected by the repeated mentioning of these events in one report.

3.2 | Descriptive analysis of the reports

Of the 16 783 sterilization device removal reports identified between January 2005 and June 2018, 94 reports were found by the NLP and confirmed by manual review to have no removal procedure (removal was negated, i.e., device not removed). We also found 155 aggregated reports. After excluding these reports, 16 535 reports were included in the final analysis. The number of reports received every year increased from <10 in 2005 to 10 536 in 2017 (Figure 2).

Of the 16 535 reports included, 16 140 were submitted by the manufacturer or user facilities (97.6%) and 395 were submitted by voluntary reporters (2.4%). Of the 16 140 manufacturer and user-facility reports, 2367 (14.7%) were submitted through the regulator, identified from the FDA docket website, or identified through media or social media. The numbers of reports submitted by different reporters within

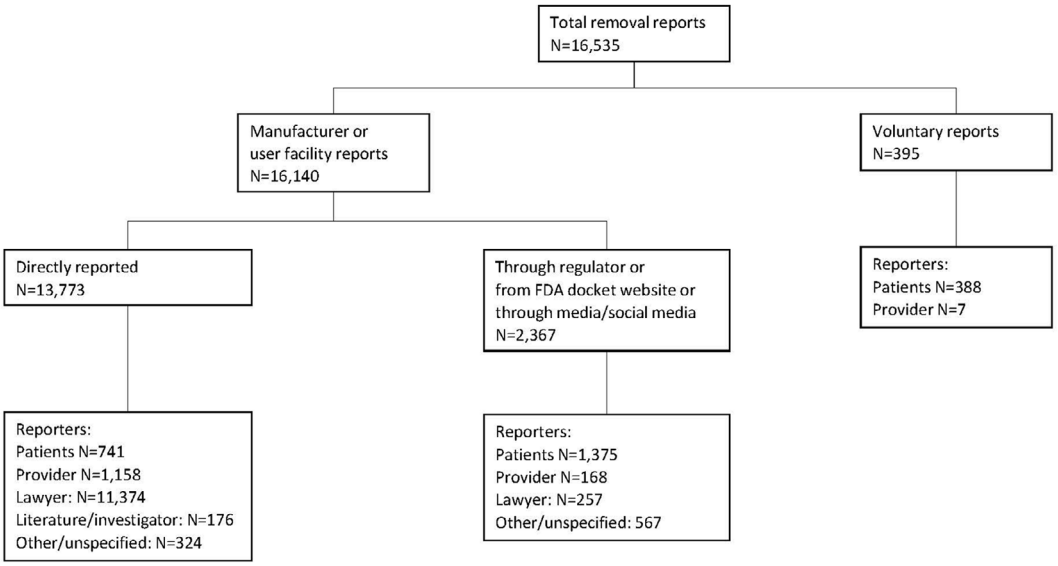


FIGURE 3 Sources and reporters of analyzed device adverse event reports submitted between Jan 2005 and Jun 2018

TABLE 2 Numbers and percentage of reports mentioning prespecified patient and device events and additionally found high-frequency events

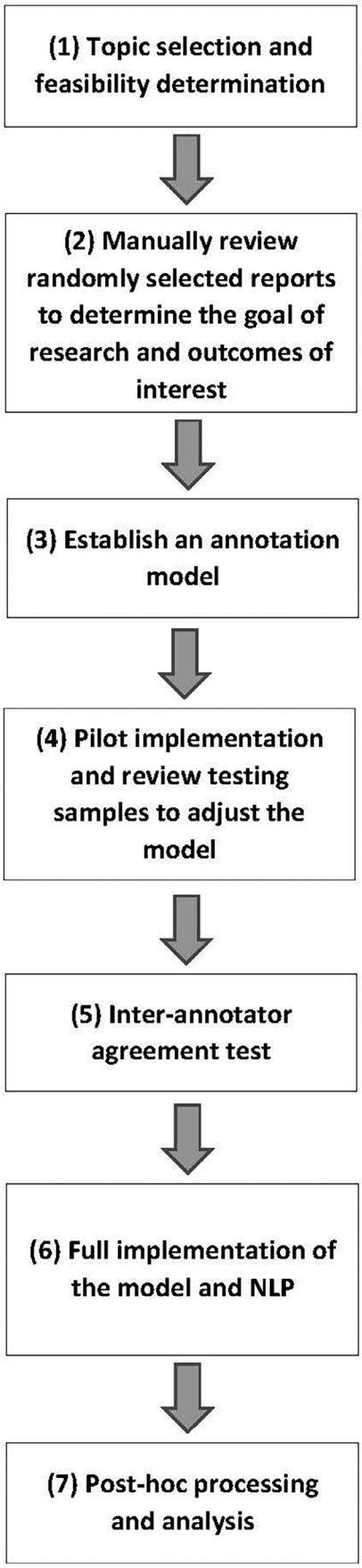
Event category	Reported by
Prespecified patient events	
Abdominal/pelvic/genital pain	13 166 (79.6%)
Bleeding	5790 (35.0%)
Menstrual disorder	6207 (37.5%)
Allergy to metal	1361 (8.2%)
Pregnancy	1165 (7.0%)
Sexual complaints	3931 (23.8%)
Additionally found high-frequency patient events	
Unspecified pain	6773 (41.0%)
Migraine/headache	4099 (24.8%)
Fatigue	3603 (21.8%)
Back pain	3487 (21.1%)
Weight changes	2624 (15.9%)
Hair loss/abnormal growth	2520 (15.2%)
Anxiety/depression	2402 (14.5%)
Bone/muscle/joint pain	2276 (13.8%)
Abdominal distention (swelling/bloating)	2012 (12.2%)
Hive/rash	1630 (9.9%)
Prespecified device events	
Device dislocation	3180 (19.2%)
Organ perforation	1865 (11.3%)
Device breakage	832 (5.0%)
Embedded device	405 (2.4%)
Device shape alteration	78 (0.5%)
Additionally found device events	
Device deployment problems	16 (0.1%)
Device retention	13 (0.1%)

each reporting source category are shown in Figure 3. Overall, the most frequent reporters were lawyers ($N = 11\,631$, 70.3%), patients or consumers ($N = 2504$, 15.1%), and providers ($N = 1333$, 8.1%). Of all reports, 1451 (8.8%) were confirmed by healthcare providers and 11 840 (71.6%) involved legal process.

The average number of patient or device events was 7 per report (range: 0–75). The majority of the device removal cases reported abdominal, pelvic, or genital pain ($N = 13\,166$, 79.6%) (Table 2). A substantial amount of reports mentioned bleeding ($N = 5790$, 35.0%), menstrual disorder ($N = 6207$, 37.5%), and sexual problems ($N = 3931$, 23.8%). Allergy to metal and pregnancy were reported by 8.2% ($N = 1361$) and 7.0% ($N = 1165$) patients. In addition to these prespecified patient events, other frequently reported events were unspecified pain ($N = 6773$, 41.0%), migraine or headache ($N = 4099$, 24.8%), fatigue ($N = 3603$, 21.8%), back pain ($N = 3487$, 21.1%), and weight changes ($N = 2624$, 15.9%). A comprehensive list of reported events was listed in Table S2.

The most frequently reported device events among cases that reported removal were device dislocation or migration ($N = 3180$, 19.2%) and organ perforation ($N = 1865$, 11.3%) (Figure 2C). Device breakage, embedment, and shape alteration were reported by 832 (5.0%), 405 (2.4%), and 78 (0.5%) patients, respectively. The NLP also found 16 device deployment and 13 device retention cases, which were not prespecified. Overall, 24.5% ($N = 4063$) cases reported both patient and device events, 2.2% ($N = 364$) reported device events only, 67.1% ($N = 11\,090$) reported patient events only, and 6.2% ($N = 1018$) reported neither. In the sensitivity analysis excluding reports that specified other IUD use ($N = 331$), the findings of the most commonly reported patient and device events were similar to those from the main analysis (Table S3).

Of the 8118 cases which reported a sterilization procedure in addition to removal, 5814 (71.6%) had a hysterectomy, 2118 (26.1%) had a salpingectomy, 174 (2.1%) had a tubal ligation, and 12 (0.2%)



Topic selection considerations:

1. Keyword search the database to roughly assess whether there are sufficient reports to be analyzed.
2. Literature search and clinical expertise to help determine the relevance of research

Manual review suggestions:

1. Sample across years if multiple years' reports involved
2. Manual review can stop when reviewer feels "saturated" - no new information is gained by reviewing additional reports.
3. Use these reports as training samples in later steps

Outcomes to consider: Device events, patient symptoms and health impact, reporter, additional procedures. Investigators can refer to IMDRF adverse event terminologies.

Create a dictionary and annotation guideline:

1. Clearly define each item and labeling rules.
2. If multiple annotators are involved, the dictionary and rules should be reviewed together.

Pilot testing suggestions:

1. Use ~100 samples (Train/test split: 6:4, 7:3, or 8:2)
2. This step can repeat if multiple rounds of adjustments are needed
3. All samples labeled at this stage to be used as training samples at the full implementation stage. Because when model is adjusted based on these samples, they will overfit the model

This step is only needed when more than one annotator is involved:

1. An IAA>80% is recommended.
2. Interim review and discussion between annotators are very helpful to improving consistency.

Annotation sample size considerations:

1. The performance of the model in pilot testing
2. Variation in reports over time: if large, a meaningful number of training sample (e.g. 100) is needed from each time window.
3. Train/test split: 6:4, 7:3, or 8:2. Make sure to have representation of reports from each time window in both training and testing. May consider using a smaller fraction of samples for testing when having a large labeled sample.

FIGURE 4 A general guide to developing and implementing annotation and natural language processing (NLP) to analyze device adverse event reports and similar data sources

had a device re-implantation. Some cases reported other procedures, such as a procedure on the fallopian tube, uterus, ovary, or vagina ($N = 532$, 3.2%), a urologic or intestinal procedure ($N = 60$, 0.4%), and laparotomy or an unspecified procedure ($N = 698$, 4.2%). In total, 11 293 (68.3%) reports specified implantation and removal years or the time from implantation to removal. Forty percent of these cases ($N = 4546$) reported a removal after 5 years following the initial sterilization procedure. Twenty-two percent of them ($N = 2444$) reported a removal after 7 years following the initial procedure. Event rates adjusted after considering the sensitivity and PPV of the NER model are presented in Table S4.

4 | DISCUSSION

This study presents a roadmap to develop an annotation model for NLP to analyze device adverse event reports. We demonstrate the feasibility and the value of this method in device safety evaluation and share useful strategies adopted during the development of the annotation model and the training of NER algorithms.

Applying the developed annotation model and NLP to analyzing device adverse event reports, we found that the most frequently reported patient events were pain and bleeding, and the most frequently reported device complications were device dislocation and organ perforation. The most prevalent events shown in our study align with the summary of events submitted to the FDA by the manufacturer. However, the prevalence of events obtained in our study reflected the reported events more accurately as our method captured all events, while the coded event summaries submitted to the FDA mostly only noted one or two main events. Moreover, we found that over half of the cases reporting device removal after initial hysteroscopic sterilizations only presented patient events and symptoms (e.g., pain and bleeding) without any device complications. A recent study found that patients seeking device removal after hysteroscopic sterilization had impaired quality-of-life measures.[30] Our findings and these results indicate that patients' experience after device implantation, in addition to device complications, may be important contributors to seeking additional care and procedures. These results stress the need to incorporate patient-reported outcomes in clinical trials and device registries to comprehensively evaluate device safety and help providers assess patients' conditions that warrant medical attention. We also found that 40% had the device removed after 5 years following the initial implantation and 20% after 7 years. Our previous study demonstrated the increased risk of reoperation following hysteroscopic sterilizations during a follow-up of up to 7 years.[8] The current findings suggest that reoperation may happen years after initial sterilization device implantation. Therefore, continuous evaluations of outcomes after hysteroscopic sterilization are essential despite the withdrawal of the device from the market.

These findings demonstrated that it was possible to use the annotation model and NLP to determine the occurrence of events among cases that reported a device removal. Findings from analyzing adverse event reports complement results from studies using administrative or claims data and shed light on the nature of events associated with device removal. These events, such as pain and device dislocation, are difficult to assess based on administrative or claims data. In addition, MAUDE reports contain some unique information, such as the reporting source and the invocation of litigation. These data could be used to understand the mechanism of adverse events reporting by different stakeholders. Research efforts to develop NLP systems to identify adverse events mostly focused on pharmacovigilance based on EHR data.[19–21] Compared with clinical notes or discharge summaries, MAUDE reports often use phrases that were less scientific and more varied since they were recorded by nonmedical personnel. Particularly, we noticed that the variation in reporting language was bigger for those from earlier time periods. Therefore, customized NLP systems based on annotated examples are often needed to parse these reports.

Our study adds to current literature and demonstrates the iterative process involved in developing an annotation model to train a NER algorithm. Two recent studies used NLP to analyze MAUDE reports.[23,24] One study examining safety events after robotic surgeries used rule-based methods that required specifying a dictionary.[23] The other study evaluating safety events after transcatheter valve replacement procedures used an NLP based on predictive methods but did not specify the tagging rules.[24] Neither study provided details for these methods or reported the model performance. With the increasing use of medical devices, there is a growing need for device safety investigations using various methods and data sources. While this study is based on a particular device for female sterilization, the underlying framework and strategies of the established model can serve as a guide (Figure 4) for future research and safety surveillance of other devices using similar sources of text data. This method is particularly useful and efficient when there are increased reports of device adverse events and before significant human resources are committed to summarizing them.

Lastly, it is worth discussing the challenges and limitations in analyzing adverse event reports. First, as noted above, MAUDE reports have large variations in describing events, which has often been a challenge in clinical NLP.[31,32] To improve model performance, we heavily sampled reports from the earlier time period for training to account for variation in the reporting language at that time. We also annotated all variants of events to take advantage of the repetitive reporting format (e.g., one event reported more than once in different phrases). Second, we found that the biggest challenge in establishing consistency among multiple annotators was to determine the candidate text for annotation. MAUDE reports often include comments from reporters or recorders, which may come from individual speculations about events. For scientific purposes, researchers usually desire information related to objective events. Discussions between annotators proved helpful to increase the consistency in the annotation. Third, the NER method we used had a probabilistic nature with some degree of errors. The estimated numbers of patient and device events and additional sterilization procedures after adjusting for sensitivity and PPV indicated that we have likely underestimated the number of events. However, it is unlikely that our conclusion will change based

on model performance. Fourth, the findings of frequent events after device implantation need to be interpreted in its context. These reports reflect patients' complaints and are not scientifically vetted. Device adverse events are collected through passive surveillance and may come from a group of patients who have experienced worse outcomes than the majority of patients. The proportion of patients reporting having a hysterectomy in the MAUDE reports was higher than that found in our previous study based on administrative data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS STATEMENT

The Weill Cornell Medicine Institutional Review Board approved this study as an exempt study and waived informed consent.

DATA AVAILABILITY STATEMENT

Samples of labeled texts and codes to implement NLP will be available at GitHub (https://github.com/MDEpiNet-WCM/Annotation-NLP-for-MAUDE.git) after the publication of the study.

ORCID

Jialin Mao https://orcid.org/0000-0001-6671-3474

REFERENCES

1. Kerin JF, Cooper JM, Price T, et al. Hysteroscopic sterilization using a micro-insert device: results of a multicentre phase II study. *Hum Reprod*. 2003;18(6):1223-1230.
2. Chudnoff SG, Nichols JE Jr, Levie M. Hysteroscopic Essure inserts for permanent contraception: extended follow-up results of a phase III multicenter international study. *J Minim Invasive Gynecol*. 2015;22(6):951-960.
3. U.S. Food and Drug Administration. Essure benefits and risks; 2020. Accessed Jun 8, 2020. https://www.fda.gov/medical-devices/essure-permanent-birth-control/essure-benefits-and-risks
4. U.S. Food and Drug Administration. FDA activities related to essure; 2020. Accessed June 8, 2020. https://www.fda.gov/medical-devices/essure-permanent-birth-control/fda-activities-related-essure
5. McGinley L. Sales of Essure birth control implant to be halted by Bayer; U.S. last to sell controversial device. Accessed June 8, 2020. https://www.washingtonpost.com/news/to-your-health/wp/2018/07/20/sales-of-essure-birth-control-implant-halted-by-bayer-u-s-was-last-to-sell-controversial-device/
6. U.S. Food and Drug Administration. Statement from FDA Commissioner Scott Gottlieb, M.D., on new steps to strengthen the long-term safety oversight of the Essure device following discontinuation of its U.S. Sales; 2018. Accessed Jun 8, 2020. https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-new-steps-strengthen-long-term-safety-oversight-essure
7. Mao J, Pfeifer S, Schlegel P, Sedrakyan A. Safety and efficacy of hysteroscopic sterilization compared with laparoscopic sterilization: an observational cohort study. *BMJ*. 2015;351:h5162.
8. Mao J, Guiahi M, Chudnoff S, Schlegel P, Pfeifer S, Sedrakyan A. Seven-year outcomes after Hysteroscopic and laparoscopic sterilizations. *Obstet Gynecol*. 2019;133(2):323-331.
9. Antoun L, Smith P, Gupta JK, Clark TJ. The feasibility, safety, and effectiveness of hysteroscopic sterilization compared with laparoscopic sterilization. *Am J Obstet Gynecol*. 2017;217(5):570.e1-570.e6.
10. Bouillon K, Bertrand M, Bader G, Lucot JP, Dray-Spira R, Zureik M. Association of Hysteroscopic vs laparoscopic sterilization with procedural, gynecological, and medical outcomes. *JAMA*. 2018;319(4):375-387.
11. U.S. Food and Drug Administration. MAUDE - Manufacturer and User Facility Device Experience; 2018. Accessed January 4, 2019. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm
12. Bahk CY, Goshgarian M, Donahue K, et al. Increasing patient engagement in Pharmacovigilance through online community outreach and Mobile reporting applications: an analysis of adverse event reporting for the Essure device in the US. *Pharmaceut Med*. 2015;29(6):331-340.
13. Gurtcheff SE, Sharp HT. Complications associated with global endometrial ablation: the utility of the MAUDE database. *Obstet Gynecol*. 2003;102(6):1278-1282.
14. Andonian S, Okeke Z, Okeke DA, et al. Device failures associated with patient injuries during robot-assisted laparoscopic surgeries: a comprehensive review of FDA MAUDE database. *Can J Urol*. 2008;15(1):3912-3916.
15. Delaney JW, Li JS, Rhodes JF. Major complications associated with transcatheter atrial septal occluder implantation: a review of the medical literature and the manufacturer and user facility device experience (MAUDE) database. *Congenit Heart Dis*. 2007;2(4):256-264.
16. Deng DY, Rutman M, Raz S, Rodriguez LV. Presentation and management of major complications of midurethral slings: are complications under-reported? *NeurourolUrodyn*. 2007;26(1):46-52.
17. DiBardino DJ, McElhinney DB, Kaza AK, Mayer JE Jr. Analysis of the US Food and Drug Administration manufacturer and user facility device experience database for adverse events involving Amplatzer septal occluder devices and comparison with the Society of Thoracic Surgery congenital cardiac surgery database. *J Thorac Cardiovasc Surg*. 2009;137(6):1334-1341.
18. Hauser RG, Kallinen LM, Almquist AK, Gornick CC, Katsiyiannis WT. Early failure of a small-diameter high-voltage implantable cardioverter-defibrillator lead. *Heart Rhythm*. 2007;4(7):892-896.
19. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc*. 2009;16(3):328-337.
20. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin Pharmacol Ther*. 2012;92(2):228-234.
21. Jagannatha A, Liu F, Liu W, Yu H. 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 Saf*. 2019;42(1):99-111.
22. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011;306(8):848-855.
23. Alemzadeh H, Raman J, Leveson N, Kalbarczyk Z, Iyer RK. Adverse events in robotic surgery: a retrospective study of 14 years of FDA data. *PLoS One*. 2016;11(4):e0151470.
24. Galper BZ, Beery DE, Leighton G, Englander LL. Comparison of adverse event and device problem rates for transcatheter aortic valve replacement and Mitraclip procedures as reported by the Transcatheter valve therapy registry and the Food and Drug Administration postmarket surveillance data. *Am Heart J*. 2018;198:64-74.
25. Device events; 2019. Accessed June 10, 2019. https://www.deviceevents.com/
26. U.S. Food and Drug Administration. Post-approval studies (PAS); 2020. Accessed June 8, 2020. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpma/pma_pas.cfm?t_id=531556&c_id=3675

27. International Medical Device Regulators Forum. Adverse event terminology; 2020. Accessed June 11, 2021. http://imdrf.org/workitems/wi-aet.asp
28. Rim K. Mae2: Portable annotation tool for general natural language use. Paper presented at: Proceedings of 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation; 2016.
29. Honnibal M. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing; 2020. https://spacy.io/
30. Francini S, Duraes M, Charavil A, et al. Essure removal for device-attributed symptoms: quality of life evaluation before and after surgical removal. *J Gynecol Obstet Hum Reprod*. 2020;50:101772.
31. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. *J Biomed Inform*. 2015;57:28-37.
32. Carrell DS, Schoen RE, Leffler DA, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc*. 2017;24(5):986-991.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Mao J, Sedrakyan A, Sun T, et al. Assessing adverse event reports of hysteroscopic sterilization device removal using natural language processing. *Pharmacoepidemiol Drug Saf*. 2022;31(4):442-451. doi:10.1002/pds.5402