frontiers | Frontiers in Drug Safety and Regulation

TYPE Brief Research Report
PUBLISHED 14 November 2022
DOI 10.3389/fdsfr.2022.1020943

OPEN ACCESS

EDITED BY
G. Niklas Norén,
Uppsala Monitoring Centre, Sweden

REVIEWED BY
Luis Pinheiro,
European Medicines Agency, Netherlands
Xiaofei Ye,
Second Military Medical University, China

*CORRESPONDENCE
Vivian Dang,
Vivian.Dang@fda.hhs.gov

SPECIALTY SECTION
This article was submitted to Advanced Methods in Pharmacovigilance and Pharmacoepidemiology, a section of the journal Frontiers in Drug Safety and Regulation

RECEIVED 16 August 2022
ACCEPTED 26 October 2022
PUBLISHED 14 November 2022

CITATION
Dang V, Wu E, Kortepeter CM, Phan M, Zhang R, Ma Y and Muñoz MA (2022), Evaluation of a natural language processing tool for extracting gender, weight, ethnicity, and race in the US food and drug administration adverse event reporting system.
Front. Drug. Saf. Regul. 2:1020943.
doi: 10.3389/fdsfr.2022.1020943

Evaluation of a natural language processing tool for extracting gender, weight, ethnicity, and race in the US food and drug administration adverse event reporting system

Vivian Dang [1]*, Eileen Wu[1], Cindy M. Kortepeter [1], Michael Phan[1], Rongmei Zhang[2], Yong Ma[2] and Monica A. Muñoz [1]

[1]Division of Pharmacovigilance, Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States, [2]Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States

The US Food and Drug Administration Adverse Event Reporting System (FAERS) contains over 24 million individual case safety reports (ICSRs). In this research project, we evaluated a natural language processing (NLP) tool's ability to extract four demographic variables (gender, weight, ethnicity, race) from ICSR narratives. Specificity of the NLP algorithm was over 94% for all demographics, while sensitivity varied between the demographics: 98.6% (gender), 45.5% (weight), 100% (ethnicity), and 85.3% (race). Among ICSRs missing weight, ethnicity, and race in the structured field, few cases had this information in the narrative (>95% missing); consequently, the positive predictive value (PPV) for these three demographics had wide 95% confidence intervals. After NLP implementation, the total number of ICSRs missing gender was reduced by 33% (i.e., NLP identified 472 thousand reports having a gender value in the narrative that was not in the structured field), while the total number of ICSRs missing weight, ethnicity, or race was reduced by less than 4%. This study demonstrated that the implementation of an NLP tool can provide meaningful improvements in the availability of gender information for pharmacovigilance activities conducted with FAERS data. In contrast, NLP tools targeting the extraction of weight, ethnicity, or race from free-text fields have minimal impact largely because the information was infrequently provided by the reporter. Further gains in completeness of these fields must originate from increases in provision of demographic information from the reporter rather than informatic solutions.

KEYWORDS

drug safety, FAERS, individual case safety reports, natural language procesing, text mining, pharmacovigilance

1 Introduction

The U.S. Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) is a continuously growing database with over 24 million individual case safety reports (ICSRs), designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products (FDA, 2022c). ICSRs contain structured and unstructured fields that provide patient information and information regarding adverse events, medication errors, or product quality issues (FDA, 2018). Patient demographic information such as age, gender, weight, ethnicity, and race can be found in the structured fields, which are readily analyzable. Demographic information that is missing from the structured fields may be extracted from the ICSR's free-text narrative, an unstructured field, if the information is present. This process can be labor intensive and time consuming. Therefore, to enhance this process, we investigated a previously created natural language processing (NLP) tool that uses rule-based algorithms to scan the narratives and extract information regarding patients' age, gender, weight, ethnicity, and race. Our NLP algorithm for age has been validated and evaluated in FAERS (Pham et al., 2021); however, its performance should not be generalized to the other demographic variables because each variable has its own rule-based algorithm. Therefore, the objectives of this study are to characterize the presence of data for the remaining four demographic variables in the FAERS structured fields, validate our NLP tool's performance in extracting these demographic variables from the free-text narrative field, and evaluate the NLP tool's ability to address missing demographic variables in the FAERS structured fields.

Demographic information regarding gender, weight, ethnicity, and race can provide new insight to a drug's adverse event profile that may lead to new labeling or clinical considerations. For example, studies have shown that female patients may have a higher risk of liver injury than male patients with nonsteroidal anti-inflammatory drugs (Lacroix et al., 2004; LiverTox, 2020; Schmeltzer et al., 2016). Furthermore, there may be a higher risk of angioedema in black patients and cough in Chinese patients who are treated with an angiotensin-converting enzyme inhibitor compared to non-black and non-Chinese patients, respectively (McDowell et al., 2006; Tseng et al., 2010; Shi et al., 2018). Weight has always been an important factor in weight-based dosing products such as low-molecular-weight heparins and some anesthetics where the therapeutic efficacy and safety depend on the volume of distribution and total blood volume (Ingrande and Lemmens, 2010; Gerlach et al., 2013; Barras and Legg, 2017).

Furthermore, detecting and extracting demographic information from the free-text narrative using decision support tools is important because it will facilitate safety reviewer practices by reducing manual labor and search time (Pandey et al., 2019; Spiker et al., 2020). The additional information extracted will help FDA characterize FAERS data more accurately, provide safety reviewers with more information about their case series, and allow for more timely completion of safety data assessments.

2 Methods

2.1 Data source

FAERS is an electronic database that contains over 24 million ICSRs, including follow-up versions from 1968 to the present (FDA, 2022c). Since 2018, FAERS has received more than 2 million reports per year (FDA, 2022c). The majority of reports are from the United States; however, FAERS also receives reports from foreign countries. Currently, there are three types of reports in FAERS: direct, expedited, and non-expedited. Direct reports are voluntarily submitted to FAERS by healthcare professionals and consumers through the MedWatch program on 3500 and 3500B forms, respectively (FDA, 2020). Expedited and non-expedited reports are mandatory reports submitted by manufacturers either through the MedWatch program on the 3500A form or by an electronic submission system using the International Council for Harmonisation (ICH) E2B (R2) format (FDA, 2022a). An expedited report contains at least 1 unlabeled adverse event along with a serious reported outcome and must be submitted to FAERS within 15 days of the initial receipt of the report by manufacturers (eCFR, 2004). Reported outcomes are classified as serious and non-serious. Serious reported outcomes include death, hospitalization, life threatening, disability, congenital anomaly, required intervention, and other serious outcomes (eCFR, 2004). Non-expedited reports are all other reports that do not meet the criteria for expedited reporting (eCFR, 2004).

Before 2016, the MedWatch forms (3500, 3500B, 3500A) did not have a separate field for ethnicity or race. Reporters would have to include this information in the free text section if available. In 2019, the MedWatch forms (3500, 3500B, 3500A) replaced the data field "Sex" with "Gender" and included more selections such as transgender, intersex, and prefer not to disclose in addition to the previous female and male options. In 2022, the MedWatch forms were updated again to have separate sex and gender fields to align with the definition provided by the Centers for Disease Control and Prevention (GovInfo, 2022); the forms now include options for cisgender, transgender, and a free text section for reporters to specify other gender categories. Currently, the electronic submission system does not have data fields to collect information regarding ethnicity, race, and other gender terms (e.g., transgender, cisgender, undifferentiated).

For this study, we only retrieved the latest version of an ICSR to avoid duplicate reports that were older follow-up versions submitted by the same reporter.

2.2 Natural language processing tool

NLP has been widely adapted to accelerate processing time for large data sets such as pharmacovigilance data, electronic health records, and social media data (Wong et al., 2018). A previous study found that rule-based approaches are superior to machine learning approaches for the extraction of demographic variables from FAERS data and suggested using rules that are based on raw text strings over rules that are based on Part-Of-Speech tags of individual tokens for higher performance (Wunnava et al., 2017). Our NLP tool has four algorithms that use rules based on raw text strings; each algorithm is created to extract a demographic variable of interest from the free-text narrative. For example, the algorithm for weight identifies numeric values before a term that describes a weight unit (e.g., pound, lb), while the algorithm for gender primarily assesses counts of free-text narrative terms reflecting a male or female gender (e.g., female/male, his/her, boy/girl). These rule-based algorithms are written in the Python Programming Language with the incorporation of regular expressions (Supplementary Table S1).

2.3 Characterization of the food and drug administration adverse event reporting system database

We extracted the latest version of ICSRs submitted to the FAERS database from 1 January 1968 to 31 December 2020. We calculated the total number of ICSRs and the annual proportion of ICSRs with missing demographic information in the structured field for each demographic variable.

2.4 Random sample of individual case safety reports to create reference standards

Due to the large differences in the proportion of ICSRs with missing demographic variables in different years, we selected a separate random sample of 750 reports for each demographic variable. Although the NLP tool was designed to work regardless of missingness in the structured field, we only sampled reports with missing demographic information because we are interested in applying NLP to these reports in the FAERS database. The study period for the validation of the NLP algorithm for gender and weight was from 2000 to 2020 (time frame A) because there was a high proportion of ICSRs missing weight (>99%) for most years before 2000, and the proportion of ICSRs missing gender was generally greater than 17% before 1992 but subsequently started decreasing to less than 10%, which continued throughout most years in this study period. The study period for ethnicity and race was from 2016 to 2020 (time frame B) because the MedWatch forms did not have a separate data field to collect ethnicity or race information until 2016; consequently, ethnicity and race were always missing for most years before 2016. Each random sample had 750 reports (a round number for 742) with the assumption that the true Positive Predictive Value (PPV) is 0.80; this number of reports will provide more than 90% power to rule out a PPV of 0.75. We chose to use PPV to determine the sample size because PPV is the most important metric of the three metrics used to evaluate the performance of the NLP tool for our use case; a general rule of minimum PPV at 80% is widely used in regulatory science. In addition, we chose to rule out a PPV of 0.75 because a PPV lower than 0.75 would indicate an NLP tool with poor performance. The four random samples were created using Pandas, a software library written for Python programming language. To form the reference standards, two blinded reviewers manually extracted each demographic variable from the free-text narratives of the random samples. Any disagreements between reviewers were adjudicated by the study team.

2.5 Characterization of all individual case safety reports in study period

We extracted the latest version of all ICSRs for the two study periods. For each demographic variable, ICSRs were stratified by reports with and without the demographic value in the structured field for comparison of the following data elements: report type, reporter country, and reported outcomes.

2.6 Natural language processing tool validation

NLP outputs were compared against the reference standards to obtain confusion matrices, performance scores (sensitivity, specificity, PPV), and 95% confidence intervals (CI). For false positive and negative results, we read the narratives to identify the cause of mismatches. Report type distributions were compared between our samples and FAERS reports in the study periods to ensure that our sample was representative of the reports in FAERS (Supplementary Table S2). We also characterized the four samples by report type, reporter country, and reported outcomes (Supplementary Table S3).

TABLE 1 Characterization of FAERS ICSRs in the two study periods: time frame A (2000–2020) and time frame B (2016–2020), by report type, reporter country (country of the event or country of the reporter if country of the event is missing), and reported outcomes (an ICSR can have more than one outcome). *BSR*, Biologic Safety Report; FAERS, Food and Drug Administration Adverse Event Reporting System; ICSRs, individual case safety reports.

Time frame A: 1 January 2000-31 December 2020 **Time frame B: 1 January 2016-31 December 2020**	**All ICSRs in time frame A *n***	**Reports missing gender in time frame A *n* (%)**	**Reports missing weight in time frame A *n* (%)**	**All ICSRs in time frame B *n***	**Reports missing ethnicity in time frame B *n* (%)**	**Reports missing race in time frame B *n* (%)**
ICSRs	13,486,486	1,426,135 (10.6)	10,407,430 (77.2)	6,722,646	6,630,644 (98.6)	6,629,913 (98.6)
Report type						
Expedited	6,690,590	775,859 (11.6)	5,108,888 (76.4)	3,240,958	3,240,739 (100.0)	3,240,887 (100.0)
Non-expedited	5,980,715	614,818 (10.3)	4,921,251 (82.3)	3,092,842	3,092,841 (100.0)	3,092,841 (100.0)
Direct	810,547	33,876 (4.2)	367,134 (45.3)	384,517	292,735 (76.1)	291,856 (75.9)
Other (30-Day, 5-Day, BSR)	4,634	1,582 (34.1)	3,637 (78.5)	4,329	4,329 (100.0)	4,329 (100.0)
Reporter country						
United States	9,673,176	1,013,147 (10.5)	7,575,422 (78.3)	4,786,380	4,696,954 (98.1)	4,696,351 (98.1)
Foreign	3,661,535	401,383 (11.0)	2,722,493 (74.4)	1,855,458	1,853,842 (99.9)	1,853,797 (99.9)
Not Reported	151,775	11,605 (7.6)	102,995 (67.9)	80,808	79,848 (98.8)	79,765 (98.7)
Reported outcomes						
Death	1,314,311	182,175 (13.9)	1,095,717 (83.4)	594,550	590,410 (99.3)	590,795 (99.4)
Hospitalization	3,091,975	173,000 (5.6)	2,083,954 (67.4)	1,358,127	1,339,289 (98.6)	1,338,758 (98.6)
Life threatening	394,666	23,091 (5.9)	214,678 (54.4)	159,893	150,470 (94.1)	149,642 (93.6)
Disability	289,044	18,086 (6.3)	153,696 (53.2)	99,329	82,791 (83.4)	80,824 (81.4)
Congenital anomaly	49,422	18,567 (37.6)	34,939 (70.7)	19,000	18,747 (98.7)	18,726 (98.6)
Required intervention	116,410	6,772 (5.8)	52,429 (45.0)	7,023	5,129 (73.0)	4,818 (68.6)
Other serious	4,587,949	577,649 (12.6)	3,401,202 (74.1)	2,308,625	2,273,150 (98.5)	2,271,774 (98.4)
Non-serious	5,678,822	590,900 (10.4)	4,705,872 (82.9)	3,152,263	3,126,664 (99.2)	3,128,433 (99.2)

2.7 Evaluation of natural language processing tool's application in the food and drug administration adverse event reporting system

The impact of the NLP tool was evaluated by identifying reports with NLP extractable demographic information among reports missing the data in the structured fields; we compared the proportion of reports missing demographic information before and after NLP implementation. In addition, we conducted a secondary analysis to further explore the impact of the algorithm in addressing missing gender for products with a high proportion (≥60%) of reports missing gender information in the structured field and ≥1000 reports in FAERS. A total of 21 products were identified. We applied the algorithm to all ICSRs collected for these products and compared the proportion of ICSRs missing gender in the structured field before and after implementation for each product.

3 Results

3.1 Characterization of individual case safety reports

Of the 15,321,967 reports in FAERS (latest versions only), from 1968 to 2020, the proportion of reports with missing gender, weight, ethnicity, and race in the structured field was 11.0%, 77.9%, 99.4%, and 99.2%, respectively. Between 2000 and 2015, the proportion of reports missing gender in the structured field was generally less than 10% (Supplementary Figure S1); however, in recent years (2016–2020), after the FDA required manufacturers to electronically submit all ICSRs on 10 June 2015 (FDA, 2022a), the proportion of these reports has been increasing, up to 15% in 2020. Weight was missing from almost all reports before 1993. From 1993 to 2000, the proportion of reports missing weight decreased from 97% to 65% with most reports

TABLE 2 Performance of the four NLP algorithms in their unique reference standards. NLP, natural language processing; PPV, positive predictive value.

NLP algorithms	# Of matches	# Of mismatches	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV % (95% CI)
Gender	717	33	98.6 (95.9, 99.7)	94.4 (92.1, 96.2)	87.5 (82.6, 91.4)
Weight	741	9	45.5 (16.8, 76.6)	99.6 (98.8, 99.9)	62.5 (24.5, 91.5)
Ethnicity	750	0	100.0 (29.2, 100.0)	100.0 (99.5, 100.0)	100.0 (29.2, 100.0)
Race	745	5	85.3 (68.9, 95.1)	100.0 (99.5, 100.0)	100.0 (88.1, 100.0)

submitted through the MedWatch 3500A form. However, the proportion of reports missing weight began increasing after 2006, up to 82% in 2020, as electronic submission increases. Ethnicity and race are two variables with high proportions of missingness. Ethnicity was always missing before 2015. Race was missing almost 100% before 2015 unless it was reported in the free-text field of the 3500A form.

In time frame A, there was a total of 13,486,486 ICSRs; 10.6% and 77.2% reports had missing gender and weight in the structured field, respectively. And in time frame B, there was a total of 6,722,646 ICSRs; 98.6% reports had missing ethnicity and race in the structured field. For both study periods, direct reports had the lowest proportion of ICSRs missing gender, weight, ethnicity, and race (Table 1). For reporter country, there was less than a 4% difference between the proportion of ICSRs with missing gender, weight, ethnicity, and race from the U.S. compared to foreign countries. A higher proportion of reports with missing gender was seen in reports coded with an outcome of congenital anomaly (37.6%) than all other outcomes (5.6%–13.9%). Among ICSRs with missing weight, ethnicity, and race, a lower proportion was found in reports with a serious outcome of "required intervention" when compared to other serious outcomes.

3.2 Natural language processing tool validation

Distribution of ICSRs, by report type, did not vary by more than 5% for all four samples when compared to their corresponding datasets in FAERS (Supplementary Table S2).

3.2.1 Gender natural language processing algorithm

The gender NLP algorithm had a sensitivity of 98.6% (95% CI 95.9%–99.7%), specificity of 94.4% (95% CI 92.1%–96.2%), and PPV of 87.5% (95% CI 82.6%–91.4%) (Table 2). There were 717 matches and 33 mismatches. Of the 717 matches, 210 were true positives and 507 were true negatives. Of the 33 mismatches, 30 were false positives and 3 were false negatives.

In 28 of the 30 reports with false positives, the algorithm incorrectly detected gender term(s) that was not describing the patient (e.g., gender of the patient's parent, healthcare provider, or reporter). In the other 2 reports with false positive results, 1 report had multiple patients, so the algorithm output the gender with the higher term count, but our reviewers recorded unknown. The other report output female because the algorithm detected "HER-2," a protein receptor, as female because it has the term "her" in it. In 2 of the 3 reports with false negatives, the algorithm did not detect "herself" and "f" as female, but reviewers were able to detect female as the patient's gender from interpreting the content in the narratives. In the other report, the algorithm output unknown because the term count was the same for both genders, although the report concerned a male patient.

3.2.2 Weight natural language processing algorithm

The weight NLP algorithm had a sensitivity of 45.5% (95% CI 16.8%–76.6%), specificity of 99.6% (95% CI 98.8%–99.9%), and PPV of 62.5% (95% CI 24.5%–91.5%) (Table 2). There were 741 matches and 9 mismatches. Of the 741 matches, there were 5 true positives, 736 true negatives. Of the 9 mismatches, 3 were false positives and 6 were false negatives. Of the 3 reports with false positives, the algorithm detected the patient's weight loss value as the patient's current weight in 2 reports and detected a drug dose as the patient's weight in 1 report. In the 6 reports with false negatives, the algorithm did not detect a weight value for the patient because the numeric weight value was listed with an unknown weight unit or a weight unit other than pound or kilogram. Of the 750 reports, two reports described weight values in grams that were not captured by the NLP tool; reviewers were able to identify the weight values.

3.2.3 Ethnicity natural language processing algorithm

The ethnicity NLP algorithm had a sensitivity of 100.0% (95% CI 29.2%–100.0%), specificity of 100.0% (95% CI 99.5%–100.0%), and PPV of 100.0% (95% CI 29.2%–100.0%) (Table 2). There were 750 matches and 0 mismatches. Of the 750 matches, there were 3 true positives and 747 true negatives.

FIGURE 1
Percentage of ICSRs missing gender in the structured field before and after NLP implementation from 2000 to 2020 in FAERS. The proportion of ICSRs missing gender reduced by the NLP implementation increased over the years: 0.8 in 2000, 3.4 in 2010, and 6.1 in 2020. ICSRs, individual case safety reports; NLP, natural language processing; FAERS, Food and Drug Administration Adverse Event Reporting System.

3.2.4 Race natural language processing algorithm

The race NLP algorithm had a sensitivity of 85.3% (95% CI 68.9%–95.1%), specificity of 100.0% (95% CI 99.5%–100.0%), and PPV of 100.0% (95% CI 88.1%–100.0%) (Table 2). There were 745 matches and 5 mismatches. From the 745 matches, there were 29 true positives and 716 true negatives. From the 5 mismatches, there were 5 false negatives and no false positives. One report had a value of "white" that was not followed by a prespecified term (e.g., male, female, man, woman, patient). In the other 4 reports, the algorithm did not detect other phrasing of race such as "African-American", "African descent", and "Chinese ethnic origin" as a race value.

3.3 Natural language processing tool application

After NLP implementation, the total number of ICSRs missing gender decreased from 1,426,135 to 954,102 (a 33.1% reduction) during the study period. The impact of the gender NLP algorithm had on ICSRs missing gender in the structured field increased over the years (Figure 1). The proportion of ICSRs with missing gender in the structured field was reduced by 0.8% in 2000, 3.4% in 2010, and 6.1% in 2020, a 762% increase from 2000. In 2019 and 2020, the number of ICSRs missing gender reduced by over 40% after NLP implementation.

In our secondary analysis, the gender NLP algorithm reduced the proportion of ICSRs missing gender in the structured field by a large range (0.0%–95.0%), depending on the product. For example, for burosumab-twza, the proportion of reports with missing gender decreased from 99.1% to 4.1% after NLP implementation. However, for opioid products, the application of the gender NLP algorithm was limited because these reports did not have gender information in the narrative. For example, the proportion of reports missing gender for 4 opioid products (oxymorphone hydrochloride, aspirin/oxycodone hydrochloride, hydrocodone bitartrate, hydromorphone) decreased <1% after NLP implementation (Figure 2).

Supplementary Figures S2–S4 show the proportion of ICSRs with missing weight, ethnicity, race, respectively, in the structured field before and after NLP implementation. Implementation of the NLP algorithms did not meaningfully reduce the number of ICSRs (<4%) missing these demographic information in the structured fields. After NLP implementation, ICSRs with missing weight in the structured field decreased from 10,400,910 to 10,233,758 (1.6% reduction), ICSRs with missing ethnicity in the structured field decreased from 6,629,913 to 6,605,674 (0.4% reduction), and ICSRs with missing race in the structured field decreased from 6,629,913 to 6,380,165 (3.8% reduction). During the study period, the proportion of ICSRs

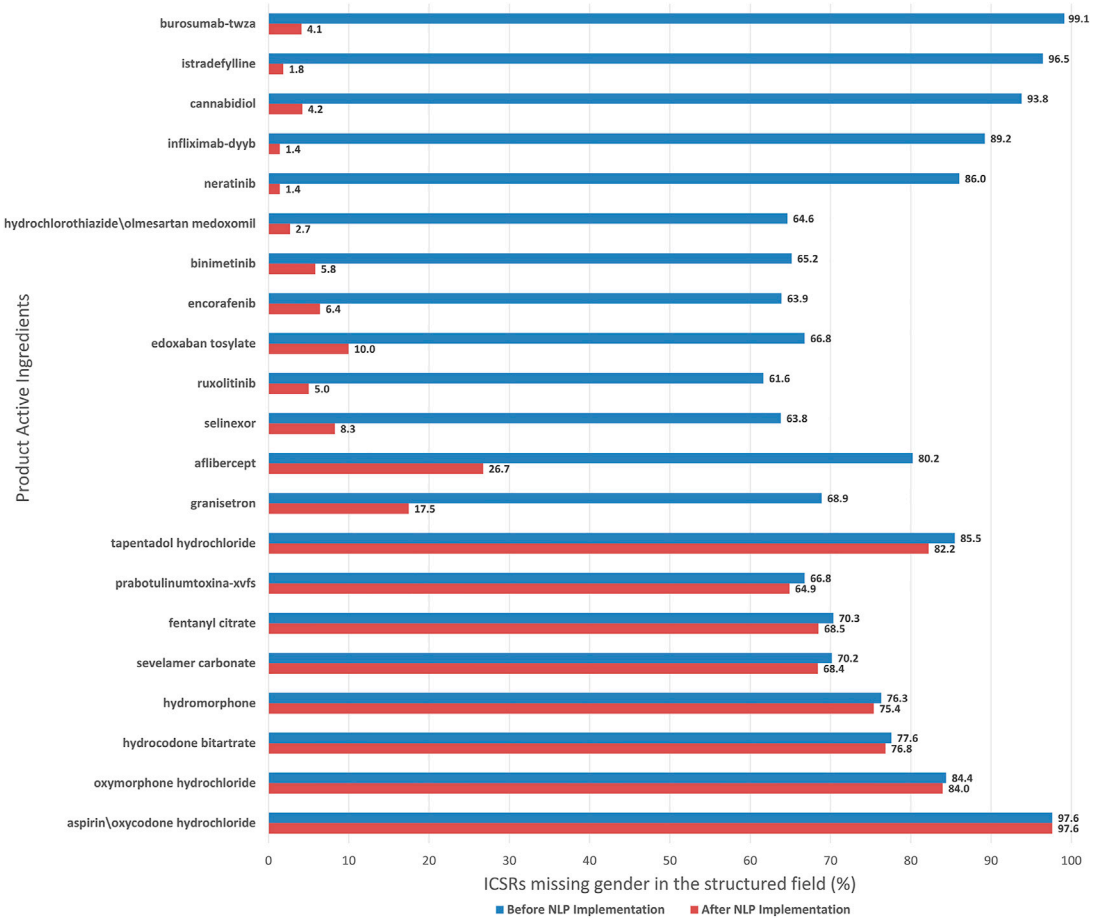

FIGURE 2
Percentage of ICSRs missing gender in the structured field before and after NLP implementation, by decreasing impact of NLP, for products in FAERS with high report counts (≥1,000) and high proportion of reports missing gender (≥60%). ICSRs, individual case safety reports; NLP, natural language processing; FAERS, Food and Drug Administration Adverse Event Reporting System.

missing race in the structured field had the largest reduction (5.4%) after NLP implementation in 2016, but this reduction decreased to 3.6% in 2018 then further decreased to 2.2% in 2020.

4 Discussion

During the study period (time frame A), the proportion of FAERS ICSRs missing gender and weight in the structured field increased 133.7% and 27.4%, respectively. The overall number of reports received annually in FAERS has also increased from 199,799 in 2000 to over 2 million in 2020 and has been continuing to rise, largely due to electronic submission of expedited and non-expedited reports from manufactures (FDA, 2022c). This increase has been attributed, at least in part, to industry sponsored programs (e.g., patient support programs, market research program) and social media (Jokinen et al., 2019; Marwitz et al., 2020). Correspondingly, some of these report sources have been found to have less complete information in structured fields like gender, but missingness is also highly variable within sources (Harinstein et al., 2019; Jokinen et al., 2019). In addition, the proportion of FAERS ICSRs missing race and ethnicity in the structured field were both above 98%. In 2015, FDA began requiring manufacturers to submit all ICSRs and periodic reports electronically using the ICH E2B (R2) format (FDA, 2022a) which does not have separate data fields to collect patients' race and ethnicity information. Until the electronic submission system updates to the FDA regional implementation of E2B (R3) standards (FDA, 2022b), manufacturers are not required to report patients' race and ethnicity information unless they choose to include the information in the free-text fields (ICH, 2001). The incorporation of NLP solutions can help extract clinically relevant information to reduce variability and missingness of demographic information if the information exists in the

narratives. Furthermore, NLP solutions can facilitate safety reviewer practices by reducing search time and manual labor in addition to aiding reviewers needing to curate data on specific demographic characteristics for sub-analyses. In our study, the gender NLP algorithm extracted four correct gender values that were missed by two reviewers during the creation of the reference standard which further illustrates the importance of decision support tools (i.e., even well-trained assessors can make an error). Consequently, reviewers agreed to change their extracted gender values to match the NLP output; therefore, the initial incorrect gender values did not affect the assessment of the algorithm's performance scores.

The gender NLP algorithm demonstrated promising results in this study. The majority of the mismatches were false positives (30/33) which we could reduce by updating the gender NLP algorithm to ignore gender term(s) that exist in sentences with a term that describes a person other than the patient (e.g., nurse, pharmacist, physician, reporter, mother); this will prevent the algorithm from identifying a gender term that does not belong to the patient. During the creation of the gender reference standard, our reviewers searched for gender terms included in the updated version of the MedWatch form (FDA, 2020); however, the gender NLP algorithm only outputs female or male. Although our reviewers did not detect gender terms other than female and male, future research should consider updating the algorithm to include other gender terms such as transgender. However, any updates to the algorithm would need further validation because new errors could arise and influence the performance scores. Besides enhancing our rule-based algorithm, we can also explore other NLP tools (e.g., machine learning) to extract gender from free-text.

Furthermore, the gender NLP algorithm extracted a gender value in more than 472,000 narratives of ICSRs missing gender in the structured field during the study period. As more reports are submitted to the FAERS database, reporters are reporting patients' gender less in the structured field (Supplementary Figure S1) but more in the free-text narratives. Figure 1 shows the proportion of ICSRs with missing gender in the structured field had a higher reduction in 2020 (6.1%) compared to 2000 (0.8%).

Moreover, in our secondary analysis, we found that after NLP implementation, the proportion of ICSRs with missing gender could be reduced up to 95.0%. Burosumab-twza, istradefylline, infliximab-dyyb, cannabidiol, and neratinib are examples of products that had >85% of ICSRs missing gender in the structured field before NLP implementation and <5% after NLP implementation. The implementation of the gender NLP algorithm was unable to reduce the proportion of ICSRs missing gender in the structured field more than 4% for 8 products (Figure 2). ICSRs of these 8 products were serious reports (90.9%), submitted as either expedited or non-expedited reports (99.5%), had an outcome of death (70.0%), and pertained to an opioid product (95.6%). Moore et al. (2016) reviewed serious adverse event reports received in FAERS in 2014 and found that report completeness from drug manufacturers was lower than direct reports, and report completeness was the lowest for the subset of reports with an outcome of death; gender was a component in their completeness measurement. Furthermore, 15.6% of these reports on the 8 products only listed "death" as an adverse event. A previous study found that reports with only death listed as an adverse event were more likely to have incomplete structured data fields and less information in the narrative (Marwitz et al., 2020). Other factors that may lead to reports missing gender in the structured field and narrative are changes in manufacturer operating procedures and reporting practices (Harinstein et al., 2019). Further research is needed to better understand why some of these reports have missing information in both the structured and narrative fields.

Although the weight, race, and ethnicity NLP algorithms have high specificity scores (≥99%), the 95% confidence intervals for sensitivity and PPV were wide (Table 2). Furthermore, after NLP implementation, the proportion of ICSRs missing weight and ethnicity were both reduced by less than 2%. This is largely due to the information not being present in the narratives. The proportion of ICSRs missing race reduced after NLP implementation was 5.4% in 2016 but has decreased over time: 3.6% in 2018, and 2.2% in 2020. This shows that there is an increase in the number of reports missing race in the structure field and narrative. The performance of the NLP algorithms depends on the relevant demographic information being present in the narratives and is restricted if the prevalence is low (Pham et al., 2021).

Our study has some limitations. Although we collected the latest version of the ICSRs, which removed the duplicates of previous versions, we did not further assess our dataset for duplicates because, similar to the NLP age study, we wanted to validate the NLP tool for a random sample of reports that is representative of the FAERS database (Pham et al., 2021). In addition, we did not review report attachments such as literature articles or laboratory documents, therefore, it is possible that we have missed demographic information contained in the attachments and miscalculated the number of reports with missing demographic information. Reports missing information in the structured field were used for validation because they aligned with our pharmacovigilance use case (i.e., a safety reviewer would use this tool during their case retrieval and review only if the structured field is null). A different use case (e.g., quality assurance of reports in FAERS) may warrant further evaluation of the NLP tool's performance among reports with the corresponding field populated. The implementation of this tool in other spontaneous reporting systems that record case narratives in languages other than English would require additional modifications and revalidation. Lastly, we did not know the true PPV, so we assumed it was 0.8 when determining the validation sample size.

5 Conclusion

Currently, FAERS has an uptick of ICSRs missing weight and gender in the structured field and a high proportion of ICSRs are missing weight, race, and ethnicity in the structured field as well as the unstructured narrative. Our study demonstrated that the implementation of an NLP tool can facilitate the extraction of gender information from unstructured text with high performance and tangible impact. Information regarding weight, ethnicity, and race were infrequently provided in the free-text narrative; therefore, the NLP tool had a minimal impact on these variables. Improvement in the availability of this information will need to originate from increases in demographic information provided by the reporter. The use of the NLP tool at FDA is currently under evaluation, among other technologies being studied. Further research and improvements could be made to capture more demographic information from the free-text fields among other helpful information (e.g., product names, adverse events, temporality) to enhance postmarket surveillance.

Data availability statement

The datasets presented in this article are not readily available because the datasets generated during and/or analyzed during the current study are only available from the corresponding author on reasonable request. Requests to access the datasets should be directed to vivian.dang@fda.hhs.gov.

Author contributions

The first draft of the manuscript was written by VD and all authors commented on previous versions of the manuscript. All authors contributed to the study conception, design, and data collection. All authors read and approved the final manuscript.

Funding

This project was supported in part by an appointment to the ORISE Research Participation Program at the CDER administered by the Oak Ridge Institute for Science and Education through an agreement between the US Department of Energy and the US FDA. VD and MP conducted this research as ORISE fellows in the Office of Surveillance and Epidemiology, Center of Drug Evaluation and Research, FDA.

Acknowledgments

We would like to acknowledge the Regulatory Science Staff within FDA's Office of Surveillance and Epidemiology, who worked with students and staff at the Worcester Polytechnic Institute to develop the NLP algorithm.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

This article reflects the views of the authors and should not necessarily be construed to represent FDA's views or policies.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdsfr.2022.1020943/full#supplementary-material

References

Barras, M., and Legg, A. (2017). Drug dosing in obese adults. *Aust. Prescr.* 40 (5), 189–193. doi:10.18773/austprescr.2017.053

eCFR (2004). *21 CFR §314.80 Postmarketing reporting of adverse drug experiences*. Available at: https://www.ecfr.gov/current/title-21/chapter-I/subchapter-D/part-314/subpart-B/section-314.80 (Accessed March 24, 2022).

FDA (2022b). *FDA regional implementation guide for E2B(R3) electronic transmission of individual case safety reports for drug and biological products*. Available at: https://www.fda.gov/media/98536/download (Accessed March 24, 2022).

FDA (2022a). *FDA adverse event reporting system (FAERS) electronic submissions*. Available at: https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-electronic-submissions#:~:text=*FDA%20issued%20a%20final%20rule,electronic%20format%20(FDA%20Archive) (Accessed March 24, 2022).

FDA (2022c). *FDA's adverse event reporting system (FAERS) public dashboard*. Available at: https://fis.fda.gov/sense/app/95239e26-e0be-42d9-a960-9a5f7f1c25ee/sheet/7a47a261-d58b-4203-a8aa-6d3021737452/state/analysis (Accessed April 24, 2022).

FDA (2020). *MedWatch: The FDA safety information and adverse event reporting program*. Available at: https://www.fda.gov/safety/medical-product-safety-information/medwatch-forms-fda-safety-reporting (Accessed March 24, 2022).

FDA (2018). *Questions and answers on FDA's adverse event reporting system (FAERS)*. Available at: https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers (Accessed January 24, 2022).

Gerlach, A. T., Folino, J., Morris, B. N., Murphy, C. V., Stawicki, S. P., and Cook, C. H. (2013). Comparison of heparin dosing based on actual body weight in non-obese, obese and morbidly obese critically ill patients. *Int. J. Crit. Illn. Inj. Sci.* 3 (3), 195–199. doi:10.4103/2229-5151.119200

GovInfo (2022). Federal register/vol. 87, No. 51/wednesday, march 16, 2022/notices. Available at: https://www.govinfo.gov/content/pkg/FR-2022-03-16/pdf/2022-05514.pdf (Accessed August 7, 2022).

Harinstein, L., Kalra, D., Kortepeter, C. M., Muñoz, M. A., Wu, E., and Dal Pan, G. J. (2019). Evaluation of postmarketing reports from industry-sponsored programs in drug safety surveillance. *Drug Saf.* 42 (5), 649–655. doi:10.1007/s40264-018-0759-7

ICH (2001). *Maintenance of the ICH guideline on clinical safety data management: Data elements for transmission of individual case safety reports E2B(R2)*. Available at: https://admin.ich.org/sites/default/files/inline-files/E2B_R2_Guideline.pdf (Accessed March 24, 2022).

Ingrande, J., and Lemmens, H. J. (2010). Dose adjustment of anaesthetics in the morbidly obese. *Br. J. Anaesth.* 105 (1), i16–i23. doi:10.1093/bja/aeq312

Jokinen, J., Bertin, D., Donzanti, B., Hormbrey, J., Simmons, V., Li, H., et al. (2019). Industry assessment of the contribution of patient support programs, market research programs, and social media to patient safety. *Ther. Innov. Regul. Sci.* 53 (6), 736–745. doi:10.1177/2168479019877384

Lacroix, I., Lapeyre-Mestre, M., Bagheri, H., Pathak, A., Monstastruc, K. L., et al. (2004). Nonsteroidal anti-inflammatory drug-induced liver injury: A case-control study in primary care. *Fundam. Clin. Pharmacol.* 18 (2), 201–206. doi:10.1111/j.1472-8206.2004.00224.x

LiverTox (2020). *Nonsteroidal antiinflammatory drugs (NSAIDs)*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK548614 (Accessed January 24, 2022).

Marwitz, K., Jones, S. C., Kortepeter, C. M., Dal Pan, G. J., and Muñoz, M. A. (2020). An evaluation of postmarketing reports with an outcome of death in the US FDA adverse event reporting system. *Drug Saf.* 43 (5), 457–465. doi:10.1007/s40264-020-00908-5

McDowell, S. E., Coleman, J. J., and Ferner, R. E. (2006). Systematic review and meta-analysis of ethnic differences in risks of adverse reactions to drugs used in cardiovascular medicine. *BMJ* 332 (7551), 1177–1181. doi:10.1136/bmj.38803.528113.55

Moore, T. J., Furberg, C. D., Mattison, D. R., and Cohen, M. R. (2016). Completeness of serious adverse drug event reports received by the US Food and Drug Administration in 2014. *Pharmacoepidemiol. Drug Saf.* 25, 713–718. doi:10.1002/pds.3979

Pandey, A., Kreimeyer, K., Foster, M., Dang, O., Ly, T., Wang, W., et al. (2019). Adverse event extraction from structured product labels using the event-based text-mining of health electronic records (ETHER) system. *Health Inf. J.* 25 (4), 1232–1243. doi:10.1177/1460458217749883

Pham, P., Cheng, C., Wu, E., Kim, I., Zhang, R., Ma, Y., et al. (2021). Leveraging case narratives to enhance patient Age ascertainment from adverse event reports. *Pharm. Med.* 35 (5), 307–316. doi:10.1007/s40290-021-00398-5

Schmeltzer, P. A., Kosinski, A. S., Kleiner, D. E., Hoofnagle, J. H., Stolz, A., Fontana, R. J., et al. (2016). Liver injury from nonsteroidal anti-inflammatory drugs in the United States. *Liver Int.* 36 (4), 603–609. doi:10.1111/liv.13032

Shi, V., Senni, M., Streefkerk, H., Modgill, V., Zhou, W., and Kaplan, A. (2018). Angioedema in heart failure patients treated with sacubitril/valsartan (LCZ696) or enalapril in the PARADIGM-HF study. *Int. J. Cardiol.* 264, 118–123. doi:10.1016/j.ijcard.2018.03.121

Spiker, J., Kreimeyer, K., Dang, O., Boxwell, D., Chan, V., Cheng, C., et al. (2020). Information visualization platform for postmarket surveillance decision support. *Drug Saf.* 43 (9), 905–915. doi:10.1007/s40264-020-00945-0

Tseng, D. S., Kwong, J., Rezvani, F., and Coates, A. O. (2010). Angiotensin-converting enzyme-related cough among Chinese-Americans. *Am. J. Med.* 123 (2), e11–e15. 5. doi:10.1016/j.amjmed.2009.06.032

Wong, A., Plasek, J. M., Montecalvo, S. P., and Zhou, L. (2018). natural language processing and its implications for the future of medication safety: A narrative review of recent advances and challenges. *Pharmacotherapy* 38 (8), 822–841. doi:10.1002/phar.2151

Wunnava, S., Qin, X., Kakar, T., Socrates, V., Wallace, A., and Rundensteiner, E. (2017). "Towards transforming FDA adverse event narratives into actionable structured data for improved pharmacovigilance," in Proceedings of the symposium on applied computing, Marrakech, Morocco, 3-4 April 2017, 777–782. doi:10.1145/3019612.3022875