Pharmaceutical Medicine (2021) 35:307–316
https://doi.org/10.1007/s40290-021-00398-5

ORIGINAL RESEARCH ARTICLE

Leveraging Case Narratives to Enhance Patient Age Ascertainment from Adverse Event Reports

Phuong Pham[1,2] · Carmen Cheng[2] · Eileen Wu[2] · Ivone Kim[2] · Rongmei Zhang[3] · Yong Ma[3] · Cindy M. Kortepeter[2] · Monica A. Muñoz[1,2]

Accepted: 5 August 2021 / Published online: 2 September 2021

Abstract

Introduction Missing age presents a significant challenge when evaluating individual case safety reports (ICSRs) in the FDA Adverse Event Reporting System (FAERS). When age is missing in an ICSR's structured field, it may be in the report's free-text narrative.

Objectives This study aimed to evaluate the performance and assess the potential impact of a rule-based natural language processing (NLP) tool that utilizes a text string search to identify patients' numerical age from unstructured narratives.

Methods Using FAERS ICSRs from 2002 to 2018, we evaluated the annual proportion of ICSRs with age missing in the structured field before and after NLP application. Reviewers manually identified patients' age from ICSR narratives (gold standard) from a random sample of 1500 ICSRs. The gold standard was compared to the NLP-identified age.

Results During the study period, the percentage of ICSRs missing age in the structured field increased from 21.9 to 43.8%. The NLP tool performed well among the random sample: sensitivity 98.5%, specificity 92.9%, positive predictive value (PPV) 94.9%, and *F*-measure 96.7%. It also performed well for the subset of ICSRs missing age in the structured field; when applied to these cases, NLP identified age for an additional one million ICSRs (10% of the total number of ICSRs from 2002 to 2018) and decreased the percentage of ICSRs missing age to 27% overall.

Conclusions NLP has potential utility to extract patients' age from ICSR narratives. Use of this tool would enhance pharmacovigilance and research using FAERS data.

Key Points

Missing age in the structured field of post-marketing individual case safety reports (ICSRs) presents a significant challenge for case assessment during safety surveillance.

The rule-based algorithm evaluated in the current study achieved high performance in extracting patients' age in ICSRs in the FDA Adverse Event Reporting System (FAERS) database and resultingly identified patient age for an additional one million ICSRs between 2002 and 2018.

This tool may be beneficial for implementation in other databases where age may not be consistently available in a structured field but otherwise reported in an unstructured narrative field.

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

✉ Carmen Cheng
Carmen.Cheng@fda.hhs.gov

[1] Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, FL, USA

[2] Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA

[3] Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA

1 Introduction

The US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) is a database designed to support FDA's post-marketing safety surveillance program for drug and therapeutic biological products [1]. The FAERS database contains spontaneous individual case safety reports (ICSRs) of adverse events (AEs) and medication errors submitted by healthcare professionals, consumers, and manufacturers.

FDA conducts ongoing surveillance of FAERS for all marketed products. Evaluation of ICSRs includes consideration for the product, event, and patient factors. One crucial patient factor is age, as it provides physiologic, pharmacologic, and epidemiologic context for surveillance. For example, pediatric patients are more susceptible to adverse drug reactions; robust pediatric data for therapies are often lacking, and many products are used off-label [2, 3].

Despite being an elemental demographic, age is frequently missing in an ICSR's structured field [4–6]. However, the narrative of an ICSR allows reporters to provide information in free-form text format, which may include information not captured in structured fields. Currently, FDA reviewers manually review the narratives or use other customized workarounds (e.g., narrative text string searches) to determine if a patient's age is reported in the narrative when missing from the structured field. Beyond ICSR evaluation, missing age information also has implications on accurate retrieval of relevant cases for review that are dependent on the structured age field and on signal detection algorithms that incorporate age information [7]. These implications are especially important for the identification of cases relevant to pediatric-focused post-marketing safety reviews, including those mandated by US legislation in 2002 [8, 9].

Natural language processing (NLP) has been widely used to facilitate drug safety activities with various data sources including electronic health records, internet-based data such as social media, published medical literature, and spontaneous reporting systems [10, 11]. Wunnava et al. [12] developed and evaluated rule-based and machine learning-based extraction methods using NLP to extract patient demographics, drug, and AE information from unstructured narratives of 60 ICSRs from the FAERS database. Their analyses suggest that the rule-based extraction method using raw text strings performed better than the supervised machine learning-based extraction methods or rules based on grammar or part-of-speech for demographic information in FAERS. Based on the demonstrated performance of the rule-based extraction method, a simple NLP algorithm described by Wunnava et al. to extract patient age, we expanded on their findings to validate this tool using a larger gold standard dataset of ICSRs in FAERS. Furthermore, we evaluated the extent to which NLP can improve the identification of patient age from ICSRs in the FAERS database from 2002 to 2018.

2 Methods

2.1 Data Source

The FAERS database contains more than 21 million ICSRs from 1968 to 2020. Manufacturers submit postmarketing ICSRs as expedited (within 15 days) if the ICSRs contain an AE that is not described in the product labeling and led to a serious outcome; otherwise, the ICSRs are submitted as non-expedited [13]. ICSRs that are voluntarily submitted to the FDA by healthcare professionals or consumers through the MedWatch program are referred to as direct reports [14].

The FAERS structure adheres to the international safety reporting guidance issued by the International Conference on Harmonisation (ICH) [15]. ICSRs contain many data elements for patient characteristics, reaction or event, product, and reporter; structured data are encouraged for electronic submissions [16]. Some data elements, such as case narratives, are unstructured fields in the form of free-text; additionally, submission of ICSRs allows for attachments such as medical records, images, and published literature. FDA guidance on data element transmission of ICSRs allows senders to use different ways of including the same data without being redundant to cope with differing information contents (e.g., age information can be sent as date of birth and date of reaction or event, age at the time of reaction or event, or patient age group according to the available information) [16]. The structured age field is populated by the age provided by the reporter or if missing, a calculated age is populated in FAERS if both the patient's date of birth and event date are available.

2.2 Descriptive Analyses

We identified all ICSRs initially received by the FDA from January 1, 2002, to December 31, 2018. The latest ICSR version was used as the representative report if follow-up information was received from the same reporter. We calculated the overall and annual proportion of ICSRs missing an age in the structured field by report type (i.e., expedited, non-expedited, or direct). ICSRs were then stratified by those with and without an age in the structured field for comparison of the following report characteristics: report type, reporter type (e.g., healthcare provider, consumer), reporter country, and reported outcomes. Serious outcomes are defined by US regulations and include one or

more of the following: death, hospitalization, life-threatening, disability, congenital anomaly, and other serious outcomes [13].

2.3 Sample Selection and Gold Standard Ascertainment

A random sample of 1500 ICSRs from all ICSRs received during the study period was selected for reviewers to manually identify patients' age from the ICSR narratives. The size of the random sample was first determined by the considerations of statistical power calculation and then slightly increased given the reasonably easy process of manual review. Assuming the positive predictive value (PPV) is 0.90, a sample size of 1500 provides more than 90% power to rule out a lower bound of 95% confidence interval (CI) of PPV lower than 85%.

The gold standard for the NLP validation process was the manually-extracted patient's age (in years) at the time of the first AE (if multiple AEs were reported) or the patient's age at the time of reporting (if the age at event onset was not reported). If the patient's age in the narrative was reported in days or months, reviewers converted the age to years (values were rounded to the tenth digit to match the rounding performed by the NLP algorithm). Patient's age reported as an age range, approximate age, or not reported in the narrative were assigned as null age.

Two reviewers independently reviewed each ICSR narrative to identify age. If there was a disagreement, a third reviewer adjudicated. Some ICSRs in the dataset may contain age in the structured field, but reviewers were blinded to this value. Age from the structured field was not taken into consideration during the determination of the gold standard or validation phase of the study.

2.4 NLP Algorithm

The NLP algorithm uses rule-based text mining to extract age from the unstructured narrative field of the ICSRs and outputs the age in years. The algorithm searches the text in each ICSR's narrative field and extracts the first instance of the numerical value preceding variations of the text strings reflecting "years" or "years old." If there are no year terms in the text, then the algorithm extracts the first instance of the numerical value preceding variations of the text strings for "months" or "months old" and converts the value to age in years (rounded to the nearest tenth digit). If no terms are extracted for age, then the algorithm outputs null age. The regular expression code used to implement the algorithm is provided in Supplementary Materials Table 1.

Table 1 Outcome classification of NLP tool

	Gold standard age[a]	
	Relevant	Not relevant
NLP age		
Extracted	True positive[b]	False positive[c]
Not extracted	False negative[d]	True negative[e]

NLP natural language processing

[a]Gold standard age: the manually-extracted patient's age (in years) from the report narrative at the time of the first adverse event (if multiple adverse events were reported) or at the time of reporting (if the age at event onset was not reported)

[b]True positive: NLP tool extracted a value that exactly matched the gold standard

[c]False positive: NLP tool extracted a value that either did not exactly match the gold standard or there was no actual value set as the gold standard

[d]False negative: NLP tool did not extract a value when there was an actual value set as the gold standard

[e]True negative: NLP tool did not extract a value and there was no actual value set as the gold standard

2.5 NLP Validation

To evaluate the performance of the NLP tool, we compared the values from the NLP output with the gold standard for an exact match (Table 1). The two primary metrics used to evaluate information retrieval are PPV (precision) and sensitivity (recall). Sensitivity is the probability that a relevant value is retrieved by the tool. PPV is the probability that a retrieved value is relevant (i.e., matches the gold standard). The two metrics are often combined into a single measure called "*F*-measure," which allows researchers to weigh either PPV or sensitivity more heavily [17]. In our study, PPV and sensitivity were equally important, and *F*-measure was calculated as:

$$F = \frac{2 \times \text{PPV} \times \text{sensitivity}}{\text{PPV} + \text{sensitivity}} \times 100$$

Validity was also measured by specificity, the probability that a non-relevant value is not retrieved by the tool. The percentage of overall matching was calculated as:

$$\frac{\text{True positive} + \text{true negative}}{\text{sum of all reports}} \times 100$$

For each of these metrics, the 95% CI was calculated using the binomial "exact" method. In our post hoc analysis, we further assessed the performance of the NLP tool among the ICSRs with age missing and with age available in the structured field, respectively.

To analyze the NLP errors, we reviewed the ICSR narratives to categorize the reason for the mismatch.

2.6 NLP Application

We applied the NLP tool to all ICSRs received during the study period. The patient age for each ICSR was then determined using a combined approach considering the structured age field and NLP extracted age. If age was missing in the structured field, the NLP extracted age was considered the patient age. If the structured age field and NLP extracted age were null, then age was considered missing. Using this combined approach, we determined the proportion of ICSRs annually and overall missing an age. Additionally, ICSRs were tabulated by age group using the structured age field alone versus the combined approach (i.e., structured age field + NLP output). Percentage change between the groups was calculated as:

$$\frac{\text{Number of reports after NLP} - \text{number of reports before NLP}}{\text{number of reports before NLP}} \times 100$$

3 Results

3.1 Characterization of ICSRs

Of 10,300,594 ICSRs received in the study period, the overall percentage of reports with missing age in the structured field was 37.2%. The annual proportion of ICSRs missing age increased over time, from 21.9% in 2002 to 43.8% in 2018 (Fig. 1). Direct reports had the lowest overall proportion of ICSRs missing age (12%), followed by expedited and non-expedited ICSRs (31.5% and 46.6%, respectively). An increased trend in percent of ICSRs with missing age was observed for expedited and non-expedited ICSRs; however, no clear trend was observed for direct reports.

Characteristics of ICSRs with and without age in the structured field are described in Table 2. Approximately half of all ICSRs were reported by healthcare professionals, among which 68.1% provided an age in the structured field. Reports originating outside of the USA contained age more often than those originating in the USA (73.7% vs 58.8%). Of the serious outcomes reported, congenital anomalies had the lowest proportion (35.6%) and life-threatening had the highest proportion (87.2%) of ICSRs with age.

3.2 Gold Standard Ascertainment

The distribution of ICSRs by year and report type for the random sample of 1500 ICSRs was similar between the study sample and all ICSRs received (Supplementary Materials Figs. 1, 2), which indicated the sample was representative of the total ICSRs. In the random sample, 38% (570/1500) of the ICSRs were missing age in the structured field. After manual review of the 1500 ICSR narratives, we identified a numerical age from the narratives as the gold standard for 868 ICSRs (57.9%). The reviewers determined null age to be the gold standard for the remaining 632 ICSRs (42.1%); note that for ICSRs without age identified in the narrative, age may be available in the structured field.

3.3 NLP Validation

Among the 1500 ICSRs in our random sample, the NLP tool extracted an age for 894 ICSRs: 849 were true positives and

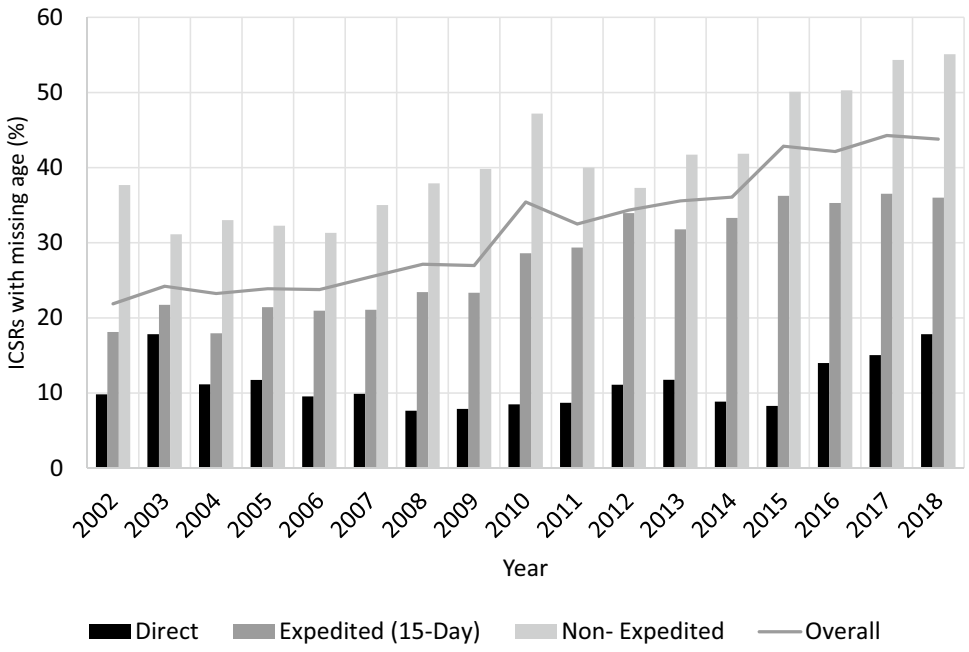

Fig. 1 Percentage of FAERS ICSRs with missing age in the structured field by report type and overall, from 2002 to 2018. *FAERS* FDA Adverse Event Reporting System, *ICSR* individual case safety report

Table 2 Descriptive characteristics of ICSRs with and without age in the structured field from 2002 to 2018

	Reports with age *n* (row%)	Reports missing age *n* (row%)	All reports *n* (column%)
ICSRs	6,472,213 (62.8)	3,828,381 (37.2)	10,300,594
Report type[a]			
Expedited	3,476,845 (68.5)	1,599,289 (31.5)	5,076,134 (49.3)
Non-expedited	2,474,799 (53.4)	2,158,227 (46.6)	4,633,026 (45.0)
Direct	520,569 (88.0)	70,865 (12.0)	591,434 (5.7)
Reporter type			
Consumer	2,808,625 (58.1)	2,028,425 (41.9)	4,837,050 (47.0)
Healthcare professional	3,448,800 (68.1)	1,612,323 (31.9)	5,061,123 (49.1)
Other/unknown	214,788 (53.4)	187,633 (46.6)	402,421 (3.9)
Reporter country			
USA	4,424,274 (58.8)	3,096,017 (41.2)	7,520,291 (73.0)
Other	2,047,939 (73.7)	732,364 (26.3)	2,780,303 (27.0)
Outcome reported[b]			
Congenital anomaly	14,958 (35.6)	27,051 (64.4)	42,009 (0.4)
Death	690,537 (66.4)	350,022 (33.6)	1,040,559 (10.1)
Disability	175,682 (73.8)	62,284 (26.2)	237,966 (2.3)
Hospitalization	1,974,104 (80.2)	487,157 (19.8)	2,461,261 (23.9)
Life-threatening	269,786 (87.2)	39,473 (12.8)	309,259 (3.0)
Other serious	2,275,691 (65.7)	1,190,552 (34.4)	3,466,243 (33.7)
No serious outcomes	2,240,265 (52.8)	2,002,861 (47.2)	4,243,126 (41.2)

ICSR individual case safety report

[a]Expedited and non-expedited ICSRs are reports that manufacturers are required to submit by regulation (US Code of Federal Regulations 314.80)

[b]An ICSR may be associated with more than one outcome

45 were false positives. Among the 606 ICSRs without an age extracted from the narrative, 593 were true negatives and 13 were false negatives. The NLP tool achieved high performance with sensitivity of 98.5% (95% CI 97.4–99.2%), PPV of 94.9% (95% CI 93.3–96.3%), *F*-measure of 96.7% (95% CI 95.3–97.7%), and specificity of 93.0% (95% CI 90.7–94.8%). Overall, 96.1% (95% CI 95.0–97.0%) of the ICSRs had a match between the NLP output and the gold standard dataset.

Our qualitative error analysis identified several categories of false positive and negative errors. Table 3 provides the frequencies of error categories and examples. The absence of a year or month unit was the most frequent reason for false negative errors (30.7%). This occurred because reviewers were able to determine age from the narrative's context despite the absence of terms analogous to "years" or "months." Other causes of false negatives were typographical errors, missed extractions of numbers at the beginning of a paragraph, non-numerical ages, and age unit of weeks and days. Missed extractions at a paragraph's start occurred because the tool required a space to be present before a numerical value. The most common cause of false positive errors was when NLP extracted a value that was a unit of time that was not age-related (60%). Often, these ICSRs described a length of time for the use of a medication or was related to information from the patient's medical history. The second most common cause of false positive errors included non-patient-specific ages extracted by NLP; examples included literature reports that described age characteristics of patients in clinical trials, observational studies, or case series.

Among the 1500 ICSRs in our random sample, there were 930 ICSRs with age available in the structured field and 570 ICSRs with age missing in the structured field (Supplementary Materials Tables 2 and 3). In the post hoc analysis, we evaluated the performance of the NLP tool among the subset of ICSRs with age missing and with age available in the structured field. The sensitivity and specificity of the NLP tool for the reports with age missing in the structured field were 98.5% and 93.3%, respectively, whereas for the reports with age in the structured field, sensitivity and specificity were 98.5% and 92.2%, respectively. However, the PPV and *F*-measure were lower for the reports with age missing in the structured field (82.3% and 89.7%, respectively), compared to those for the reports with age in the structured field (97.8% and 98.3%, respectively). Among the reports with age missing in the structured field, only 24% had age in the narratives.

Table 3 Observed causes for false negative and false positive errors

Error type	Error category	*n* (%)
False Negatives ($n = 13$)	Age reported without a unit *"age at therapy onset: 78,""46 old," "I'm only 65," "A 44 (units not provided) male patient"*	4 (30.7)
	Missed extraction (age reported at beginning of narrative or new paragraph)	3 (23.1)
	Typographical error *"68-years-old," "a68-year-old," "a74-Year old"*	3 (23.1)
	Non-numerical age *"two months old neonate," "one year old"*	2 (15.4)
	Age not reported in years or months *"42 weeks 6 days old"*	1 (7.7)
False Positives ($n = 45$)	Value extracted was not age-related, but another unit of time *"About 2 months ago," "every 6 months," "A 10-Year, Post-Marketing, Observational Study," "on DRUG x 1 year"*	27 (60.0)
	Age extracted was not patient-specific (age information from clinical trial, observational study, or case series)	8 (17.8)
	Estimated age reported *"more than 80-year-old," "age 70 or 71-year-old," "16 or 17 year old," "approximately 66 years old"*	4 (8.9)
	Not patient's age at time of adverse event (age reported in connection with medical history)	3 (6.7)
	Updated age reported in narrative[a]	2 (4.4)
	Age extracted was patient's child's age instead of patient's age *"infant was 3 months old when she started"*	1 (2.2)

[a]Algorithm extracted the first text string reflecting an age value, but unable to review full narrative for an updated age reported later in the narrative

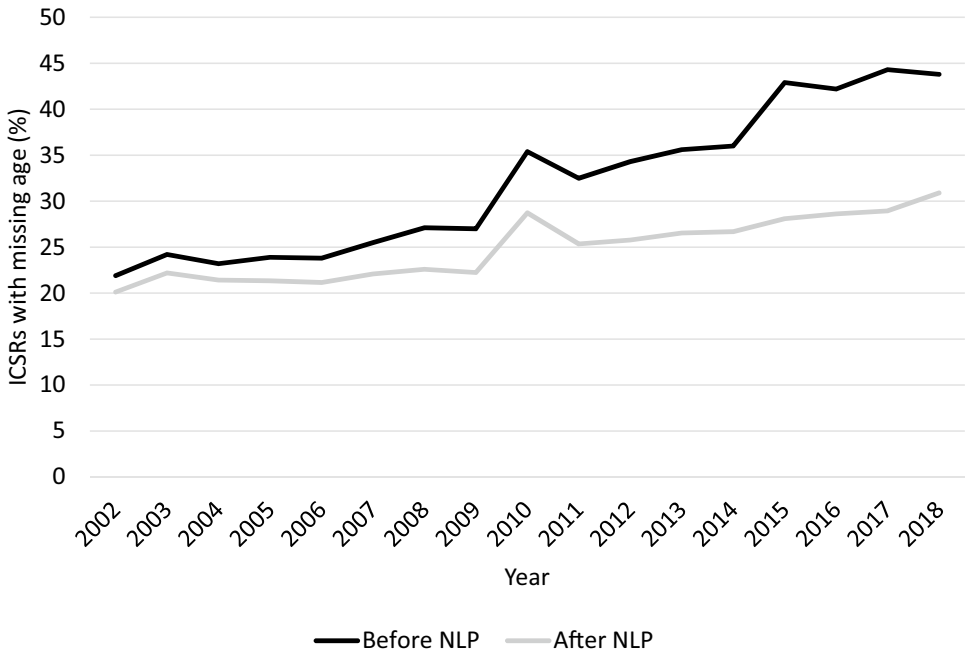

Fig. 2 Percentage of FAERS ICSRs with missing age before and after NLP implementation. *FAERS* FDA Adverse Event Reporting System, *ICSR* individual case safety report, *NLP* natural language processing

In contrast, 78% of the reports with age in the structured field had age in the narratives.

3.4 NLP Application

Among the 3.8 million ICSRs missing age in the structured field, the NLP tool identified age for an additional one million ICSRs. Resultingly, the overall number of ICSRs missing age in the study period decreased 10% (37.2–27%). Figure 2 provides the annual percentage of ICSRs missing age using the structured age field alone and the combined approach (i.e., structured age field + NLP extracted ages).

Table 4 shows the number of ICSRs with age before and after application of the NLP tool, displayed by age groups. In the pediatric age group (≤ 16 years), the number of ICSRs with age in the structured field after NLP application increased by 59%. For the younger pediatric age groups (< 6 years), the number of ICSRs with age in the structured field after NLP application was substantially higher (more than doubled) compared to the number of ICSRs before NLP application. In contrast, for the older age groups (≥ 17 years), the increase in the number of ICSRs with age in the structured field was more moderate (increase by less than 20%).

4 Discussion

As the overall number of FAERS ICSRs increased from 2002 to 2018, the percentage of ICSRs with missing age in the structured field doubled. This significant increase provided the impetus for exploring solutions to support extraction of age from the unstructured case narrative. Improving patient age ascertainment from ICSRs will aid in age-specific safety surveillance activities. We did not identify other published studies that focused on the extraction of age from narratives except the publication by Wunnava et al. This previous study explored different methods for the extraction of age from ICSR narratives and observed that a rule-based text string search performed better than those based on other methods studied; their text string search algorithm for age in 60 ICSRs achieved a sensitivity of 83%, PPV of 89%, and *F*-measure of 86% [12]. Our validation of the algorithm in a dataset of 1500 ICSRs, using the text string search for age (reported in months or years), achieved a sensitivity of 98.5%, PPV of 94.9%, and *F*-measure of 96.7% for all reports. We further compared the algorithm's performance among reports with age missing/not missing in the structured field. The performance of the NLP tool depends on the prevalence of relevant age in narratives and is restricted if the prevalence is low.

While advances have been made in the use of NLP to extract biomedical information from various data sources, our analysis demonstrated that a straightforward text string-searching algorithm could extract a numerical age from an ICSR's narrative free-text with high performance. This finding is significant because this rule-based extraction method is easy to implement. Unlike more sophisticated machine learning-based methods, the use of rule-based extraction does not require the need to train datasets. Furthermore, the text string search algorithm is interoperable across different databases. In addition to FAERS, this tool may be beneficial for implementation in other databases containing patient information where age may not be consistently available in a structured field but otherwise reported in an unstructured narrative or free-text field. The high performance of this tool supports its implementation in pharmacovigilance practices.

A new data field that considers both the structured age field and narrative-extracted age information can reduce the manual curation performed on ICSRs missing age in the structured field. For the application of NLP to ICSRs, we determined the patient age using a combined approach considering the structured age field first if a value was present, then applied the NLP extracted age only if a value was missing in the structured age field. Application of NLP to ICSRs from 2002 to 2018 using the combined approach decreased the number of ICSRs missing an age from 3.8 to 2.8 million (10% of the total number of ICSRs). We observed more than 50% increase in the number of ICSRs pertaining to the pediatric age group. Postmarketing safety information for pediatrics is especially important because of the limited number of pediatric patients in clinical trials and off-label usage [18]. Additionally, it is not possible to identify all risks of a product from premarketing clinical trials. Therefore,

Table 4 Number of ICSRs with patient age before and after NLP implementation from 2002 to 2018 across age groups (n = 10,300,594)

Age in years	Before NLP n (%)	After NLP[a] n (%)	Percent change in number of reports[b]
< 2	67,411 (0.7)	145,538 (1.4)	+ 115.9
2–5	54,253 (0.5)	107,225 (1.0)	+ 97.6
6–11	92,330 (0.9)	129,870 (1.3)	+ 40.7
12–16	118,679 (1.2)	145,754 (1.4)	+ 22.8
17–30	577,653 (5.6)	670,949 (6.5)	+ 16.2
31–59	2,641,417 (25.6)	2,978,684 (28.9)	+ 12.8
60–79	2,358,797 (22.9)	2,694,814 (26.2)	+ 14.2
> 80	561,673 (5.5)	649,960 (6.3)	+ 15.7
Missing	3,828,381 (37.2)	2,777,800 (27.0)	− 27.4

ICSR individual case safety report, *NLP* natural language processing

[a]If age was missing in the structured field, the age extracted from the narrative by the NLP tool was considered the patient age.

[b]Percent change calculated as: (number of reports after NLP − number of reports before NLP) × 100 / number of reports before NLP

postmarketing safety surveillance furthers our understanding of a product's safety in pediatric patients. However, pediatric ICSRs account for a very small percentage of the total ICSRs in the FAERS database (3% during the study period). In consideration of the relatively low number of pediatric ICSRs in FAERS, as well as the known limitations of spontaneous databases, such as underreporting, it is critical to increase the ability to accurately identify and retrieve additional pediatric ICSRs in the database. Further research is needed to determine the implications of using this derived age field for signal identification (e.g., incorporation into signaling algorithms).

We have two hypotheses for the observed larger percentage increase of ICSRs for pediatrics relative to adult age groups following NLP. First, there may be differential inclusion of age information by reporters in an ICSR's narrative due to the perceived greater importance of providing age information for pediatrics relative to other age groups. Second, the disproportional change may also be attributed to false positive errors. For 30 of 45 (67%) observed false positive errors, NLP incorrectly extracted a value for age that was less than 17 years. This is not surprising given the most common false positive error was the extraction of a value associated with other time frames, such as treatment duration and event onset. These time frames are commonly described with small numbers (e.g., 1 year ago). The small sample of pediatric cases and false positives limits our ability to draw more definitive conclusions, but the performance of the tool in the pediatric ICSR subset may benefit from additional evaluation.

Our qualitative error analysis identified several areas for improvement. Notably, although the NLP algorithm extracts the first instance of an age value preceding terms related to "years" (if none detected, then "months"), the number of false positive errors introduced by this method was low ($n = 3$). The most common cause of false positive errors was the extraction of a numerical value preceding the terms related to "years" or "months" but unrelated to age. Modifications to the algorithm could be made to reduce the number of erroneous extractions for such values unrelated to age. To reduce false negative errors and increase sensitivity, potential improvements include modifying the algorithm to capture missed extractions due to age reported with variations in spacing, paragraph breaks, or other non-numerical characters, as non-numerical values, with a decimal, or in weeks or days. However, any modifications may introduce new errors and would require validation.

Despite the use of the NLP tool, we were unable to determine age for 27% of the ICSRs in the study period. Refinements to the tool may result in incremental improvements, but it is unlikely they would substantially reduce missingness given the already high sensitivity. Major improvements in missingness will need to come from improvements in the reported data. Notably, age was less frequently missing in ICSRs sent directly to the Agency via the MedWatch program. An ICSR may be missing age because the information may be unknown to the reporter or protected for privacy reasons [19].

While beyond the objective of the study to investigate reasons for the trend in missingness, we explored missing patterns in data fields relevant to age post hoc. Currently, if age is not provided in the structured age field, FAERS calculates age using the date of birth and event date when both are available. We noted that the availability of date of birth data was stable over time, but the annual proportion of cases with an event date has decreased. It is unclear why this occurs, but it may be due to the increase in reports associated with industry-sponsored programs [6]. The completeness of data fields from these types of programs is highly variable [20, 21]. Jokinen et al. noted that certain types of programs (e.g., patient assistance programs) have been associated with lower vigiGrade scores (a measure of data completeness that includes patient age) [21, 22]. Future work should explore the reasons for the data trends.

There are some limitations to our study. We likely underestimated the availability of age in FAERS as some ICSRs contain attachments that may include relevant age information (e.g., literature article). The NLP tool is currently unable to evaluate text contained in attachments, but optical character recognition technologies would make this possible [23]. Additionally, we did not deduplicate ICSRs in the study dataset; therefore, we are unable to account for the number of unique ICSRs when we applied the NLP tool to ICSRs in the study period. Although we performed the FAERS search to retrieve the latest version of an ICSR, duplicate ICSRs describing the same patient and adverse event may still occur as a result of submission of ICSRs by different manufacturers or reporters. We did not deduplicate reports in the study set of 1500 reports because we wanted to validate the NLP tool for a random sample of reports that is representative of the overall database during the study period (2002–2018). We think validating the tool in a representative sample with duplicates would reflect the reality that many researchers are not able to deduplicate those reports when conducting their research using FAERS. Finally, our gold standard discarded estimated ages, although an estimated age or age group would be more useful than no age information.

5 Conclusions

This study demonstrated the potential for a rule-based text mining NLP algorithm to extract patients' age from the narratives of post-marketing ICSRs thereby increasing the number of ICSRs with age for analysis. The use of this tool would facilitate pharmacovigilance practice and research

using the FAERS data. The tool demonstrated good overall performance; however, further improvements may be considered. Further research to expand the use of NLP to extract information beyond age from unstructured data fields will improve postmarketing surveillance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40290-021-00398-5.

Acknowledgements 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. 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 U.S. Department of Energy and the U.S. FDA. Phuong Pham conducted this research while she was an ORISE fellow in the Office of Surveillance and Epidemiology, Center of Drug Evaluation and Research, FDA.

Declarations

Funding No sources of funding were used to assist in the preparation of this study.

Conflict of interest Phuong Pham, Carmen Cheng, Eileen Wu, Ivone Kim, Rongmei Zhang, Yong Ma, Cindy M. Kortepeter, and Monica A. Muñoz have no conflicts of interest.

Availability of data and material The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability Not applicable.

Authors' contributions All authors contributed to the study conception, design, and data collection. Data analysis was performed by Phuong Pham, Carmen Cheng, and Monica Munoz. The first draft of the manuscript was written by Phuong Pham and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval This study was granted an exemption for review by the U.S. Food and Drug Administration Institutional Review Board.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. US Food and Drug Administration. Questions and answers on FDA's Adverse Event Reporting System (FAERS). https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers. Accessed 25 Nov 2019
2. De Salas CSRL. Pharmacovigilance in pediatric population. In: Pharmacovigilance. IntechOpen; 2019.
3. Hoon D, Taylor MT, Kapadia P, Gerhard T, Strom BL, Horton DB. Trends in off-label drug use in ambulatory settings: 2006–2015. Pediatrics. 2019;144(4): e20190896.
4. Moore TJ, Furberg CD, Mattison DR, Cohen MR. Completeness of serious adverse drug event reports received by the US Food and Drug Administration in 2014. Pharmacoepidemiol Drug Saf. 2016;25(6):713–8.
5. Misu T, Kortepeter CM, Munoz MA, Wu E, Dal Pan GJ. An evaluation of "Drug Ineffective" postmarketing reports in drug safety surveillance. Drugs Real World Outcomes. 2018;5(2):91–9.
6. Marwitz K, Jones SC, Kortepeter CM, Dal Pan GJ, Munoz MA. An evaluation of postmarketing reports with an outcome of death in the US FDA adverse event reporting system. Drug Saf. 2020;43(5):457–65.
7. Duggirala, HJ, Tonning JM, Smith E, Bright RA, Baker JD, Ball R, et al. Data mining at FDA. http://www.fda.gov/downloads/ScienceResearch/DataMiningatFDA/UCM443675.pdf. Accessed 28 Nov 2019
8. Johann-Liang R, Wyeth J, Chen M, Cope JU. Pediatric drug surveillance and the Food and Drug Administration's adverse event reporting system: an overview of reports, 2003–2007. Pharmacoepidemiol Drug Saf. 2009;18(1):24–7.
9. US Food and Drug Administration. Best Pharmaceuticals for Children Act and Pediatric Research Equity Act. https://www.fda.gov/science-research/pediatrics/best-pharmaceuticals-children-act-and-pediatric-research-equity-act. Accessed 3 May 2020
10. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform. 2017;73:14–29.
11. Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy. 2018;38(8):822–41.
12. Wunnava S, Qin X, Kakar T, Socrates V, Wallace A, Rundensteiner E. Towards transforming FDA adverse event narratives into actionable structured data for improved pharmacovigilance. In: 2017 Proceedings of the symposium on applied computing, pp 777–82. https://doi.org/10.1145/3019612.3022875
13. Electronic Code of Federal Regulations. Postmarketing reporting of adverse drug experiences. 21 CFR §314.80. https://www.ecfr.gov/cgi-bin/text-idx?SID=394813c1cb14662de9713c5f7b783043&mc=true&tpl=/ecfrbrowse/Title21/21cfr314_main_02.tpl. Accessed 14 May 2020
14. US Food and Drug Administration. MedWatch: The FDA Safety Information and Adverse Event Reporting Program. Last updated: May 12, 2020. https://www.fda.gov/safety/medwatch-fda-safety-information-and-adverse-event-reporting-program. Accessed 14 May 2020
15. International Council for Harmonisation (ICH). Efficacy guidelines. https://www.ich.org/page/efficacy-guidelines. Accessed 30 Apr 2020
16. US Department of Health and Human Services. E2BM data elements for transmission of individual case safety reports. 2002. https://www.fda.gov/media/71208/download. Accessed 30 Apr 2020
17. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296–8.
18. US Food and Drug Administration. Pediatric safety. Last updated: March 22, 2018. https://www.fda.gov/science-research/pediatrics/pediatric-safety. Accessed 14 May 2020
19. US Food and Drug Administration. Guidance for Industry: E2B(R3) Electronic Transmission of Individual Case Safety Reports (ICSRs) implementation guide—data elements and message specification. February 2014. https://www.fda.gov/media/81904/download. Accessed 30 Apr 2020

20. Harinstein L, Kalra D, Kortepeter CM, Munoz MA, Wu E, Dal Pan GJ. Evaluation of postmarketing reports from industry-sponsored programs in drug safety surveillance. Drug Saf. 2019;42:649–55.
21. Jokinen J, Bertin D, Donzanti B, Hormbrey J, Simmons V, Li H, et al. Industry assessment of the contribution of patient support programs, market research programs, and social media to patient safety. Ther Innov Regul Sci. 2019;53(6):736–45.
22. Bergvall T, Noren GN, Lindquist M. vigiGrade: a tool to identify well-documented individual case reports and highlight systematic data quality issues. Drug Saf. 2014;37(1):65–77.
23. Schmider J, Kumar K, LaForest C, Swankoski B, Naim K, Caubel PM. Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing. Clin Pharmacol Ther. 2019;105(4):954–61.