A comparison of two user-friendly methods to identify and support correction of misspelled medications

Christopher R. Dasaro [*], Ahmad Sabra, Yunho Jeon, Tankeesha A. Williams, Nancy L. Sloan, Andrew C. Todd, Susan L. Teitelbaum

World Trade Center Health Program General Responder Data Center, Department of Environmental Medicine and Climate Science, Icahn School of Medicine at Mount Sinai, 17 East 102 Street, 2nd Floor, New York, NY 10029, United States of America

ABSTRACT

Objective: To identify and support correction of misspelled medication names recorded as free text, we compared the relative effectiveness of two user-friendly methods, used without reliance on clinical knowledge.
Methods: Leveraging the SAS® COMPGED function, fuzzy string search programs examined 1.8 million medication records from 183,600 World Trade Center General Responder Cohort monitoring visits conducted in New York and New Jersey between 7/16/2002 and 3/31/2021, producing replicable generalized edit distance scores between the reported and correct spelling. Scores < 120 were selected as optimal and compared to Stedman's 2020 Plus Medical/Pharmaceutical Spell Checker first suggested word, used as the comparative standard because it employs both spelling and phonetic similarities to suggest matching words. We coded each methods' results as identifying or not identifying the medications within each visit.
Results: Most types of medications (94.4 % anxiety, 98.4 % asthma and 94.6 % ulcer/gastroesophageal reflux disease) were correctly spelled. Cross tabulations assessed the agreement (anxiety 99.9 %, asthma 99.6 % and 98.4 % ulcer/ gastroesophageal reflux disease), false positive (respectively 0.02 %, 0.03 % and 2.0 %) and false negative (respectively 1.9 %, 0.5 % and 1.0 %) values. Scores < 120 occasionally correctly identified medications missed by the spell checker. We observed no difference in medication misspellings across socio-economically and culturally diverse patient characteristics.
Conclusions: Both methods efficiently identified most misspelled medications, greatly minimizing the review and rectification needed. The fuzzy method is more universally applicable for condition-specific medications identification, but requires more programming skills. The spell checker is inexpensive, but benefits from modest programming skills and is only available in some languages.

1. Introduction

The importance of accurate spelling in reporting medications has long been recognized (Blair, 1960). Misspelled medications can lead to erroneous medical records and patient care errors in clinical settings (Hussain and Qamar, 2016; Lai et al., 2015; Wittich et al., 2014; Gates et al., 2021; Srinivasamurthy et al., 2021) and can produce or obscure associations between administered medications and their health effects in research (Gamble et al., 2012). Despite the advent of electronic health records (EHR) and drop-down lists for medication names, they remain

frequently recorded as free text (Uzuner et al., 2010; Jensen et al., 2017). Clinicians may not find the medications they seek in electronic lists for various reasons, including being unsure of the spelling and/or discomfort with electronic reporting (Gardner et al., 2019). Misunderstanding patients due to language or cultural differences may also play a role in recording misspelled medications, which could particularly affect diverse patient populations.

Numerous approaches based in natural language processing have been suggested to correct electronically recorded free text, including contextual and non-contextual mechanisms for text mining and program

functions to identify text that looks or sounds like a correctly spelled word (Uzuner et al., 2010; Hajihashemi and Pancoast, 2012; Lambert, 1997; Wang et al., 2018; Gueddah et al., 2012). Medical and pharmaceutical spell checker programs have become a simpler, popular user-friendly choice to correct misspelled medications in EHRs, including supplemental programs to improve them (Lai et al., 2015; Crowell et al., 2004; Tolentino et al., 2007).

This study aimed to compare and describe the utility of two user-friendly (not requiring natural language processing skills) methods, a fuzzy string search procedure and a medical spell checker to identify medications that clinic staff directly electronically recorded into a database as free text. To assess their use without clinical training, three study authors with SAS (SAS Institute Inc.) and SPSS (IBM Corp.) programming skills implemented the methods to identify medications patients verbally reported at their health monitoring visits in a diverse patient population.

2. Methods

2.1. Study population

The Centers for Disease Control and Prevention/National Institute for Occupational Safety and Health has supported World Trade Center (WTC) related health screening starting July 1, 2002. The WTC Health Program provides medical monitoring and treatment for a socio-demographically and culturally diverse group of people who participated in the rescue, recovery, debris clearing and related services in response to the September 11, 2001 attacks on the World Trade Center. The General Responder Cohort (GRC), currently served by five clinical centers in New York and New Jersey, included 46,268 members who provided written voluntary consent for aggregation and use of their data for research through March 31, 2021. (Dasaro et al., 2017; CfDCa, 2022) This research, approved on 7/9/2020 by the Icahn School of Medicine at Mount Sinai (previously Mount Sinai School of Medicine) Institutional Review Board (approval number 15–1266), was conducted according to the World Medical Association Declaration of Helsinki (1975, revised 2013) and in accordance with national and institutional committees' standards regarding human studies. All study participants included in the analysis provided written informed consent.

2.2. Measures

We report on 38,788 consenting GRC members with 1,048,597 medications records between July 1, 2002 and March 31, 2021. Clinic staff directly recorded self-reported medication use in two, free-text data fields in an electronic database at monitoring visits scheduled every 12–18 months. As is common in health research, the WTC Program supports research to identify condition-specific outcomes. To examine a user-friendly, efficient way to identify condition-specific medications and produce generalizable results, we sought medications for the treatment of three conditions related to WTC responder exposure, anxiety including post-traumatic stress disorder, asthma and ulcers/gastroesophageal reflux disease (GERD). These conditions have a wide range of prevalence in the GRC (10 % general anxiety disorder, 18 % post-traumatic stress disorder, 31 % asthma and 48 % GERD, through March 31, 2021) and number (17 to 68) of Program Pharmacy Benefits Manager (PBM June 4, 2021 formulary file) approved medications (Table S1). (Farris et al., 2016; Wisnivesky et al., 2011).

We created fuzzy programs leveraging the SAS version 9.4 COMPGED function to estimate the 'Generalized Edit Distance' (GED) between recorded free text strings when compared to the list of correctly spelled target medications (Table S1). The fuzzy program removed extraneous elements such as punctuation marks then extracted the first three strings from each of the two free text (brand and generic name) medications fields, separating them into a maximum of six single word strings, while retaining the two original medications fields. An annotated modifiable generic version of our fuzzy program that describes its functions (to compare the six strings to the condition-specific medications list, produce a GED score and export the data to Excel) is provided in Table S2 to benefit others' use. For example, if an original field included 'Umeclidinium-Vilanterol inhaler', the program created three strings, the first being 'Umeclidinium', the second 'Vilanterol' and the third 'inhaler'. In this manner, we created 1,848,562 single string records. The fuzzy program conducted three searches, one for each condition, on all of the strings. To be comprehensive, the fuzzy program used both forwards and backwards searches to identify their possible distinct letter sequence recognition, aided by string length comparisons (particularly useful with unequal distances); the forward- and backwards matching scores were identical for each individual string. We tested using various scores on all of the medication strings and selected a GED score of < 120 as a threshold point with minimal misclassification (fewest false positives and false negatives) to identify the targeted medications (Table S1) for each condition. A GED score of zero indicated a perfect match, while a score of ≥ 120 indicated a non-matching word (Table 1). For example, if the Umeclidinium or Vilanterol spelling had a GED score < 120 it would be identified as a match, however the word Chloride from "Methacholine Chloride" would not be considered a match as it is solely a form of methacholine (per Google). As we postulated the condition prevalence and number of medications might affect the results, the fuzzy program produced condition-specific GED score Excel data files for each condition.

Stedman's Plus version 2020 Medical/Pharmaceutical spell checker (Wolters Kluwer) includes a comprehensive list of current and discontinued medications, periodically updated. Per the manufacturer's formulation, the medical spell checker is automatically integrated into the Microsoft Office general spell checker function. The medical spell checker program identifies misspelled words using phonetic similarity in addition to spelling distance algorithms to suggest a list of correctly spelled words it finds similar to the misspelled word. The spell checker does not provide distance estimates. For time-efficiency, a frequency list of all 26,738 unique word records identified from the 1,848,562 single word strings created by the fuzzy search method was spell-checked and the first word suggested by the spell checker was accepted. In this manner, neither method relied on clinical knowledge. An author proficient in SPSS programming composed and used SPSS for Windows version 24.2 syntax to assign the original medication name if it was correctly spelled and to assign the first spell checker suggested name when not correctly spelled, to create three new variables, asthma, anxiety or ulcer/GERD medication name.

Table 1
Smallest Generalized Edit Distance (GED) Matching Scores < 120 Observed within a Monitoring Visit 7/16/2002 and 3/31/2021 (1,848,562 single word records; visit records = 183,600).

GED Fuzzy < 120 Score[a]	Number of GED matching scores < 120					
	Anxiety		Asthma		Ulcer	
	n	%[b]	n	%[b]	n	%[b]
0	10,078	94.4	63,756	98.4	72,056	94.6
10	13	0.1	27	0	74	0.1
20	30	0.3	57	0.1	343	0.4
30	0	0	0	0	4	0
40	0	0	2	0	2	0
50	2	0	81	0.1	14	0
60	0	0	0	0	0	0
70	7	0.1	6	0	14	0
100	544	5.1	875	1.4	3666	4.8
Total Identified (GED <120)	10,674		64,804		76,173	
Total Not Identified (GED ≥ 120)	172,926		118,796		107,427	

[a] Excludes scores ≥ 120.
[b] Percent of total identified with a GED score < 120.

We transferred the entire fuzzy Excel data files into SPSS, the program used to conduct the statistical analysis, then merged them by using an anonymized participant number and visit date. As some study participants had multiple medication records for each of their monitoring visits, the data were restructured (long to wide) by their visit date, thus creating 183,600 single records, to identify whether a condition-specific medication was used at each visit, thus portraying their medication use as would be normal in a clinic visit. Using SPSS, we coded the fuzzy-produced GED scores < 120 as a condition-specific medication. For example, if the visit record included an anxiety medication with a fuzzy-produced GED score < 120, the anxiety medication variable was coded one, otherwise the variable was coded zero to indicate the method did not identify a target medication. If the SPSS-defined new anxiety variable indicated the spell checker identified a listed anxiety medication (Table S1), we created a dichotomous variable that coded the spell checker-identified anxiety medication variable as one, otherwise it was coded zero. The fuzzy-produced and spell checker identified asthma and ulcer/GERD medications' were dichotomously coded in the same manner.

2.3. Statistical analysis

Frequencies were produced to assess the percentage of records that were correctly spelled (with a GED score of zero) at the outset for each monitoring visit. Cross-tabulations between the two methods' dichotomous variables assessed the agreement, the false positive rate and the false negative rate for each condition. Agreement was the percentage of all records identified via both methods. In this study, the visit-specific records identified by both methods as containing a sought medication were considered the true positive (TP) records, and, regardless of whether the record was blank or contained words other than a target medication, the visit records where neither method identified a sought medication were the true negative (TN) records. As the fuzzy process used only spelling distance, we compared its results to the spell checker results because the spell checker used both spelling and phonetic word similarity to identify the sought medication names. Using the spell checker as the comparative standard, the visit records identified by the GED score but not by the spell checker as containing the sought medications were considered false positive (FP) records. The visit records identified by the GED score as not containing the sought medications that the spell checker identified as containing the sought medications were considered the false negative (FN) records. The false positive rate FP/(FP + TN) represents the percent of visit records that the GED score misclassified as including the sought medication when the spell checker did not, e.g., over-identification. Similarly, the false negative rate FN/(TP + FN), represents under-identification, the percent of visit records that the GED score misclassified as not including a sought medication when the spell checker identified the visit as including a sought medication. For comparison, sensitivity analyses were conducted using GED scores of < 100 and < 140 for each of the three conditions to determine how neighboring distinct scores compared with the chosen < 120 GED score. However, as the GED scores occasionally correctly identified sought medications not identified by the spell checker method, statistics are also presented regarding the frequency of medications correctly identified by one method but not the other (Table S3).

3. Results

Using a threshold score < 120, exact matches (GED score of zero) were identified in visits reporting: anxiety medications (10,078 of 10,674 [94.4 %]); asthma medications (63,756 of 64,804 [98.4 %]); and ulcer/GERD medications, (72,056 of 76,173 [94.6 %], Table 1).

3.1. Anxiety

The agreement between the anxiety medication GED score and the

spell checker was nearly universal (Table 2). The false positive (over-identification) rate was almost zero, however the false negative rate (missed identification) was nearly 2 %. A score < 140 yielded the same results. A score < 100 yielded 99.5 % agreement, a 6.7 % FP rate and a near-zero FN rate. Some visits had more than one condition-specific medication that one method identified but the other did not. Including all medications identified by one method but not the other, the < 120 score did not identify 219 visits with anxiety medications identified by the spell checker and the spell checker did not identify 27 visits with reported anxiety medications with a GED score < 120 (Table S3).

3.2. Asthma

For asthma, the GED score < 120 yielded almost universal agreement, with FP and FN rates below half a percent (Table 2). An asthma medication GED score < 140 yielded similar results. A GED score < 100 yielded similar agreement and FN rates, but the FP rate was 1.4 %. The < 120 score did not identify 573 visits with reported asthma medications identified by the spell checker and the spell checker did not identify 341 visits with asthma medications with a GED score < 120 (Table S3).

3.3. Ulcer/GERD

The agreement for ulcer/GERD medications, the GED score < 120 was also very high, but the FP and FN numbers were high even though the rates were 1 %-2%. (Table 2). A < 140 score yielded 93.3 % agreement, and FP and FN rates identical to the < 120 score. A < 100 score yielded a 98.8 % agreement, a 3 % FP rate and a near zero FN rate. The GED score < 120 did not identify 843 reported medications identified by the spell checker and the spell checker did not identify 101 reported medications with a GED score < 120 (Table S3). Conversely, in all but 28 visits in which ulcer/GERD medications with a GED score < 120 were not identified by spell checker, the fuzzy program misclassified acid from the term 'folic acid' (n = 2067) as 'AID' (n = 73) and 'Xai' (n = 1) as the medication 'Axid' (n = 2,141).

3.4. Percent of medications correctly spelled by participants stratified by participants' language and social determinants of health

Stratification of the correctly spelled medications (GED score = 0) by GRC members' language, ethnicity, race, sex and education did not identify any differences except for anxiety medications among the small number of participants classified as "other" primary language, although many small differences were statistically significant due to large sample sizes (Table 3). These results may indicate that there was little bias by language or other social determinants of health characteristics that might have influenced receptivity to the patients and/or language comprehension by clinic staff recording the medications.

4. Discussion

This study compared two fairly user-friendly methods, a fuzzy procedure and a medical-pharmaceutical spell checker, that were used without clinical knowledge, to support the identification and correction of misspelled medications electronically recorded by clinic staff as free text. While there were hundreds of visits identified as false negatives for anxiety, asthma and ulcer/GERD medications, and hundreds of false positives for asthma and ulcer/GERD, there was almost universal agreement between the two processes as most of the medications were correctly spelled. Both processes can identify and greatly minimize the need for medication spelling correction and reduce errors in medical histories.

There was also great similarity in the percent of medications recorded as correctly spelled across patients language-related social determinants of health characteristics, indicating that there was little bias by the clinic staff in understanding and recording patient responses. The

Table 2

Comparison of Generalized Edit Distance (GED) Score < 120 and Medical Spell Checker Visits 7/16/2002 and 3/31/2021 with Medications Identification.

		Anxiety Category			Asthma Category			Ulcer/GERD Category		
		Spell Checker Identified								
		Yes	No	Total	Yes	No	Total	Yes	No	Total
Fuzzy identified	Yes	10,647	27	10,674	64,498	306	64,804	74,004	2169	76,173
	No	210	172,716	172,926	343	118,453	118,796	742	106,685	107,427
	Total	10,857	172,743	183,600	64,841	118,759	183,600	74,746	108,854	183,600
Agreement		(10647 + 172716)/183600 = 99.9 %			(64498 + 118453)/183600 = 99.6 %			(74004 + 106685)/183600 = 98.4 %		
Kappa		98.8 %, p ≤ 0.001			99.2 %, p ≤ 0.001			96.7 %, p ≤ 0.001		
False positive[a]		27/172743 = 0.02 %			306/118759 = 0.3 %			2169/108854 = 2.0 %		
False negative[a]		210/10857 = 1.9 %			343/64841 = 0.5 %			742/74746 = 1.0 %		

[a] The spell checker is the standard to which the fuzzy results are compared.

Table 3

Percent of Visits 7/16/2002 and 3/31/2021 where Medications were Recorded with Correct Spelling (as Identified by Generalized Edit Distance (GED) Score <120) by Language and Other Social Determinants of Health.

	%	Anxiety Category n	p	%	Asthma Category n	p	%	Ulcer/GERD Category n	p
Primary Language			0.01			≤0.001			≤0.001
English	94.5	9652		98.3	56,647		94.2	66,184	
Spanish	94.4	558		99.2	4956		97.7	6497	
Polish	92.4	223		99.3	1924		97.7	1949	
Other	83.7	43		98.8	241		97.6	253	
Not reported	96.5	198		98.1	1036		91.8	1290	
Total	94.4	10,674		98.4	64,804		94.6	76,173	
Hispanic Ethnicity			0.19			≤0.001			≤0.001
Non-Hispanic	94.6	7138		98.3	42,135		94.4	48,975	
Hispanic	93.5	1588		98.9	14,206		96.3	16,015	
Not reported	94.7	1948		98.2	8463		93.2	11,183	
Total	94.4	10,674		98.4	64,804		94.6	76,173	
Race: 4 Categories			≤0.001			≤0.001			≤0.001
White	95.0	7510		98.1	38,650		94.6	47,559	
Black	91.2	455		99.2	6484		92.6	5464	
Other	93.7	1160		99.1	11,585		96.1	12,490	
Unknown	93.2	1549		98.2	8085		93.8	10,660	
Total	94.4	10,674		98.4	64,804		94.6	76,173	
Sex			0.92			0.65			0.93
Male	94.4	8702		98.4	53,516		94.6	64,757	
Female	94.4	1972		98.4	11,288		94.6	11,416	
Total	94.4	10,674		98.4	64,804		94.6	76,173	
Education			0.04			≤0.001			0.007
<HS Graduate	96.2	844		99.1	5072		95.3	5848	
HS Graduate	93.4	2373		98.6	13,808		94.9	16,246	
<BA/BS	94.5	4065		98.3	25,689		94.4	30,073	
BA/BS/Graduate	94.4	2849		98.0	17,225		94.4	20,630	
Not reported	95.2	543		98.6	3010		94.9	3376	
Total	94.4	10,674		98.4	64,804		94.6	76,173	

five clinical centers conducting the program monitoring visits all operate in the New York / New Jersey metropolitan area, and likely have staff who are themselves culturally diverse and experienced working with patients from diverse ethnic backgrounds and accents.

Over 1.8 million records and nearly 27,000 reported words in the medication free text fields were assessed. To explore the potential range of misclassification and to produce results that might be generalizable to other health conditions and medications, analyses included three health conditions that have a broad range of prevalence and medications for which they are used. As posited, the results were distinct across the three conditions. Asthma, which has the largest number of covered medications, had a small false positive rate and the lowest false negative rate of the three conditions. Ulcer/GERD, which had an intermediate number of covered medications, had the highest false positive rate and an intermediate false negative rate compared to anxiety and asthma. The lowest false positive rate and the highest false negative rate occurred with anxiety, which had the lowest prevalence and fewest covered medications.

Some of this study's methods may have limited our observed misclassification. Clinic staff were responsible for direct electronic data entry. This may have maximized the percentage of correctly spelled

medications. We found that in the worst of circumstances, only 3 % of visits had incorrectly spelled medications. As is true with many EHRs, in our data capture system, clinicians can check a box to indicate whether program members reported no medication changes since their last visit. For those situations, the subsequent visit medications pre-populate from the previous visit, which may have exacerbated the frequency of misspelled words. However, in our analysis, misspelling individual medications within a visit was more common than across visits. By analyzing medication use within each monitoring visit, our analysis minimized misclassification associated with misspelling because both generic and brand names were commonly recorded at each visit, and there is less likelihood of both being misspelled.

Both methods can search for medication names and standard medications abbreviations. The spell checker benefits from modest computer program coding ability. The fuzzy method may require more program coding ability. Additional methods such as correcting erroneously concatenated (missing white space between) words and using context-recognition that makes corrections according to the expected word within its surrounding words may produce slightly better correction. (Lai et al., 2015; Wang et al., 2018; Kukich, 1992; Hladek et al., 2020) The most frequent misclassifications were due to our singling out words

(de-contextualizing them). Using both methods can efficiently identify discordancies that can easily be contextually corrected by anyone who reviews the original visit records. Using staff without clinical knowledge to identify misspelled medications could limit the costs of correcting clinical records.

To search for and identify medication names, the fuzzy string search procedure may be more transparent and universally applicable than the spell checker, and used alone greatly minimizes the number of records that would benefit from subsequent review and correction. One limitation of the fuzzy string search procedure is determining the best threshold point for words (or, in other efforts that may search for phrases). For the GED score < 120, the 100 score identified and misidentified substantially more medication names than the lower scores. By excluding fewer similarly spelled words, scores below 100 excluded most of the misclassifications, but also missed more visits with sought medications identified by the spell checker. Identifying a GED score that maximizes correct identification and minimizes misclassification of the words sought requires testing the efficiency of distinct scores and choosing the best threshold score, which may require some knowledge of medications in order to identify false negatives and false positives. Across all three of our selected health conditions, accepting GED scores over 10 missed the correct spelling of many more reported medications than did the medical spell checker. Especially with a condition like asthma, for which many medications are used, any GED score except zero, will incur a number of errors. However, as most medications were correctly spelled initially, the fuzzy string search procedure efficiently produced a limited list of misspelled words, most of which could be easily corrected, by someone with or without contextual knowledge of the patient's condition. Other SAS functions can correct the most common spelling errors (an extraneous or excluded character, swapped letter position or incorrectly duplicated letters) can be added to the fuzzy program (Gueddah et al., 2012; Gueddah and Yousfi, 2013). As creating a universal list of medications is a burden already accomplished by medical spell checker programs, given a limited number of target medications the fuzzy method may be best used to seek condition-specific medications.

The spell checker is a simple, pragmatic and inexpensive approach (<$100 USD but with annual update fees) that identified words having similar spelling and phonetics to produce a list of likely spelling corrections. Thus, the spell checker can efficiently improve both prospective and retrospective electronic medications reporting, particularly when those conducting data entry are appropriately trained and knowledgeable about medication names. The spell checker method is also more practical when a universal search is required for medications names corrections. However, unlike the fuzzy method, spell checker software may not be available in various countries' primary language and may not include locally available medication names, thus limiting its use. In such circumstances, including medication names in document spell checkers is possible, however the greatest limitation of our spell checker accuracy was its manufacturer's integration with the generic word processor spell checker. A separate medical spell checker function not integrated into the generic document spell checker function could overcome this limitation. Using the spell checker to retrospectively review and correct each record is time consuming for any large database because its 'change all' option does not simultaneously correct all records, but rather prompts the user to implement the correction for each record. Therefore, writing and running a program, as we did using modest programming skills, can efficiently correct the misspelled words. While the spell checker's first suggested spelling correction was mostly accurate, further corrections would require clinical knowledge.

As is common with many diagnostic tests, the spell checker is a better but not truly a 'gold' standard. Regardless, the evolution of medicine renders most methodologic standards more 'silver' than 'gold'. For example, microscopic instruments, tissue sampling and clinician and computer identification of tumors has improved over time, rendering older techniques outdated. Nor does an infallible gold standard spelling denominator exist, when words are so misspelled that their correct spelling cannot be determined, at least without contextual review.

Alphabetized lists of free text self-reported medications produce neighbors-to-the-correct terms, which likely supported fuzzy and spell checker medications identification as most medications names were relatively long strings with few close alternatives. Visual inspection of such lists can simplify spelling corrections. In our data, this would have captured and supported correction of nearly all the misspelled medications.

5. Conclusion

The fuzzy string search procedure to identify and support the correction of misspelled medications involves moderate programmer skills, while the more time consuming spell checker method benefits from modest programmer skills. When used retrospectively, both methods identify almost all the misspelled medications reported as free text, and greatly minimize the amount of review and rectification needed.

Ethics approval

Funding

8. Role of the funding source

9. Disclaimer

The contents of this report are the sole responsibility of the authors and do not necessarily represent the official views of, nor an endorsement, by the National Institute for Occupational Safety and Health (NIOSH), the Centers for Disease Control and Prevention of the U.S. Department of Health and Human Services (CDC/HHS), or the U.S. Government.

Authors contributions

Christopher R Dasaro, Nancy L Sloan, Susan L Teitelbaum and Andrew C Todd contributed to the study conception and design; Material preparation and analysis were performed by Nancy L Sloan, Ahmad Sabra, Yunho Jeon and Christopher R Dasaro. The first draft of the manuscript was composed by CRD, NLS and SLT and all authors reviewed and commented on previous versions of the manuscript. All authors read and approved the final manuscript.

CRediT authorship contribution statement

Christopher R. Dasaro: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Ahmad Sabra:** Writing – review & editing, Software, Resources, Formal analysis. **Yunho Jeon:** Writing – review & editing, Software, Resources, Formal analysis. **Tankeesha A. Williams:** Writing – review & editing, Resources, Methodology. **Nancy L. Sloan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Andrew C. Todd:** Writing – review & editing, Visualization, Project administration, Funding acquisition, Conceptualization. **Susan L. Teitelbaum:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The relevant data are available within the manuscript. De-identified datasets for analysis of the current study may be requested of the corresponding author made available upon submission of the World Trade Center Data Center Data Use Agreement and Data Request Form (including an attestation) and requisite IRB approval.

Acknowledgments

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmedr.2024.102765.

References

Blair, C.R., 1960. A program for correcting spelling errors. Inf. Control 3 (1), 60–67.

Prevention CfDCa. World Trade Center Health Program 2022 [Available from: https://www.cdc.gov/wtc/ataglance.html#enrollmentWTC.

Crowell, J., Zeng, Q., Ngo, L., Lacroix, E.M., 2004. A frequency-based technique to improve the spelling suggestion rank in medical queries. J. Am. Med. Inform. Assoc. 11 (3), 179–185.

Dasaro, C.R., Holden, W.L., Berman, K.D., Crane, M.A., Kaplan, J.R., Lucchini, R.G., et al., 2017. Cohort Profile: World Trade Center Health Program General Responder Cohort. Int. J. Epidemiol. 46 (2), e9.

Farris, S.G., Paulus, D.J., Gonzalez, A., Mahaffey, B.L., Bromet, E.J., Luft, B.J., et al., 2016. Posttraumatic stress symptoms and body mass index among World Trade Center disaster-exposed smokers: A preliminary examination of the role of anxiety sensitivity. Psychiatry Res. 241, 135–140.

Gamble, J.M., McAlister, F.A., Johnson, J.A., Eurich, D.T., 2012. Quantifying the impact of drug exposure misclassification due to restrictive drug coverage in administrative databases: a simulation cohort study. Value Health 15 (1), 191–197.

Gardner, R.L., Cooper, E., Haskell, J., Harris, D.A., Poplau, S., Kroth, P.J., et al., 2019. Physician stress and burnout: the impact of health information technology. J. Am. Med. Inform. Assoc. 26 (2), 106–114.

Gates, P.J., Hardie, R.A., Raban, M.Z., Li, L., Westbrook, J.I., 2021. How effective are electronic medication systems in reducing medication error rates and associated harm among hospital inpatients? A systematic review and meta-analysis. J. Am. Med. Inform. Assoc. 28 (1), 167–176.

Gueddah, H., Yousfi, A., Belkasmi, M., 2012. Introduction of the weight edition errors in the Levenshtein distance. Internat. J. Adv. Res. Artif. Intell. 1 (5).

Gueddah, H., Yousfi, A., 2013. The impact of Arabic inter-character proximity and similarity on spell-checking. In: SITA. 2013(8th International Conference on Intelligent Systems: Theories and Applications), pp. 1–4.

Hajihashemi, Z., Pancoast, P., 2012. Reducing free-text communication orders placed by providers using association rule mining. AMIA Annu. Symp. Proc. 2012, 1254–1259.

Hladek, D., Stas, J., Pleva, M., 2020. Survey of automatic spelling correction. Electronics-Switz. 9, 10.

Hussain, F., Qamar, U., 2016. Identification and Correction of Misspelled Drugs' Names in Electronic Medical Records (EMR). Vol 2 (Iceis). In: Proceedings of the 18th International Conference on Enterprise Information Systems, pp. 333–338.

Jensen, K., Soguero-Ruiz, C., Mikalsen, K.O., Lindsetmo, R.O., Kouskoumvekaki, I., Girolami, M., et al., 2017. Analysis of free text in electronic health records for identification of cancer patient trajectories. Sci Rep-Uk. 7.

Kukich, K., 1992. Techniques for automatically correcting words in text. Comput Surv. 24 (4), 377–439.

Lai, K.H., Topaz, M., Goss, F.R., Zhou, L., 2015. Automated misspelling detection and correction in clinical free-text records. J. Biomed. Inform. 55, 188–195.

Lambert, B.L., 1997. Predicting look-alike and sound-alike medication errors. Am J Health-Syst Ph. 54 (10), 1161–1171.

Srinivasamurthy, S.K., Ashokkumar, R., Kodidela, S., Howard, S.C., Samer, C.F., Chakradhara Rao, U.S., 2021. Impact of computerised physician order entry (CPOE) on the incidence of chemotherapy-related medication errors: a systematic review. Eur. J. Clin. Pharmacol. 77 (8), 1123–1131.

Tolentino, H.D., Matters, M.D., Walop, W., Law, B., Tong, W., Liu, F., et al., 2007. A UMLS-based spell checker for natural language processing in vaccine safety. BMC Med. Inf. Decis. Making 7, 3.

Uzuner, O., Solti, I., Cadag, E., 2010. Extracting medication information from clinical text. J Am Med Inform Assn. 17 (5), 514–518.

Wang, Y.S., Wang, L.W., Rastegar-Mojarad, M., Moon, S., Shen, F.C., Afzal, N., et al., 2018. Clinical information extraction applications: A literature review. J. Biomed. Inform. 77, 34–49.

Wisnivesky, J.P., Teitelbaum, S.L., Todd, A.C., Boffetta, P., Crane, M., Crowley, L., et al., 2011. Persistence of multiple illnesses in World Trade Center rescue and recovery workers: a cohort study. Lancet 378 (9794), 888–897.

Wittich, C.M., Burkle, C.M., Lanier, W.L., 2014. Medication Errors: An Overview for Clinicians. Mayo Clin. Proc. 89 (8), 1116–1125.