Federal Aviation Administration

DOT/FAA/AM-21/23
Office of Aerospace Medicine
Washington, DC 20591

Drug Name Correction of Medication Records from Aeromedical Certification Exams

Haibiao Ding[1]

Kyle Copeland[1]

Richard Greenhaw[1]

Bill Mills[1]

Christy Hileman[1]

Vinh Kieu[2]

[1]Civil Aerospace Medical Institute, Federal Aviation Administration, Oklahoma City, OK 73125

[2]Medical Specialties Division, Federal Aviation Administration, Washington, DC 20591

June 2021

Final Report

NOTICE

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-21/23		

4. Title and Subtitle	5. Report Date
Drug Name Correction of Medication Records from Aeromedical Certification Exams	June 2021
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Haibiao Ding[1], Kyle Copeland[1], Richard Greenhaw[1], Bill Mills[1], Christy Hileman[1], Vinh Kieu[2]	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
[1]Civil Aerospace Medical Institute, Federal Aviation Administration, Oklahoma City, OK 73125 [2]Medical Specialties Division, Federal Aviation Administration, Washington, DC 20591	11. Contract or Grant No.

12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	14. Sponsoring Agency Code

15. Supplemental Notes

16. Abstract

Misspelled drug names have been introduced in Federal Aviation Administration medication records from aeromedical certification exams because of typographical errors and from transcriptions of hand-written medical prescriptions. The correction of misspelled drug names contributes to aviation safety by improving the accuracy of the records and database quality needed for aeromedical research and operational queries. This report describes development of an automatic spell correction system to correct misspelled drug names. This study compares candidate algorithms using linguistic and context metrics as well as reference dictionaries. A selection confidence index is introduced to rank the fuzzy matching results. The best correction rates are over 96% for the validation data set and approximately 93% for the much larger data set of records. To further improve the correction, more effort might be made in correcting cases in which one correct drug name or English word is misused as another drug name, refining the reference dictionary, interpreting abbreviations and strings with multiple words squeezed together, and using machine learning algorithms.

17. Key Words	18. Distribution Statement
Spelling Correction, Fuzzy Matching, Drug Name Correction, Aeromedical Certification Exam, Medication Record	Document is available to the public through the Internet: http://www.faa.gov/go/oamtechreports/

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	18	

ACKNOWLEDGEMENTS

This study was sponsored by the Office of Aerospace Medicine (AAM-1). The authors especially thank Stacey Zinke (manager of AAM-630) and Dr. Charles DeJohn (medical officer) for their support and guidance. The authors also wish to thank Civil Aerospace Medical Institute librarian Roni Anderson, Risk-Based Decision Support System database administrator Rudy Lin, and all others who supported the research.

CONTENTS

Drug Name Correction of Medication Records from Aeromedical Certification Exams

1. INTRODUCTION

Drug names are frequently misspelled in medication records. Misspelling in records has many causes, including phonetic variations, typographical mistakes (typos), reordered terms, prefixes and suffixes, abbreviations, truncated letters and missing or extra spaces, and misreading of handwritten records (Senger et al., 2010; Ferner & Aronson, 2016; Workman et al., 2019). Manual correction of misspelled words is tedious, consumes considerable effort for a large number of medical records, and can introduce new typos. Automating spelling corrections is one way to reduce the workload of researchers using these records.

Automatic spelling corrections are typically made based on how close the misspelled words are to the correct spelling. Common methods of automating spelling correction are based on concepts of edit distance, phonetic similarity, and context-sensitivity. Edit distance scores indicate how many changes or character replacements must be made to transform the examined text into a target text. Phonetic coding scores indicate similarities in expected pronunciation based on linguistic knowledge. Context-based correction methods take advantage of statistical measures of word frequencies and co-occurrences.

Often, a combination of different approaches is used for fuzzy matching. Flor and Futagi (2012) reported that edit distance methods could be augmented by phonetic information. One example from clinical records is the composite score for misspelling matching of Lai et al. (2015), which incorporates orthographic edit distance, phonetic code, and occurrence frequency together.

Algorithms have been developed for use in word correction in various types of medical records, such as clinical records, electronic medical records (Hussain & Qamar, 2016), and even social media, to extract drug or medication information using phonemes of correct spelling (Pimpalkhute et al., 2014), or by an association of a medicine with its effects (Jiang et al., 2018). However, there has been little research and algorithm development related to drug name correction specific to the medication records collected by Aviation Medical Examiners for the Federal Aviation Administration (FAA). This study examines means of improving the FAA's medication data quality in its electronic records by comparing several proposed methods using orthographic, phonetic, and context metrics in the correction of misspelled drug names found in these records in the context of the development of a feasible and operational system in an Oracle database that may be used to correct misspelled drug names.

2. METHODS

2.1. The Corpus

Medication records used in this study come from FAA Form 8500-8 as recorded in FAA's Document Imaging Workflow System (DIWS). The medications are reported to Aviation Medical Examiners by airmen seeking aeromedical certification from the FAA. Examiners submit them to the FAA on Form 8500-8 as part of the certification process. They will hereafter be referred to collectively as "the Corpus" and more individually as "Corpus records," where each corpus record is the entry from a particular 8500-8 and could thus be a single word, a phrase, a list, etc.

2.2. Misspelling Detection, Fuzzy Matching, and Software Implementation

Dictionaries are used to detect misspellings. Any word not in the dictionaries is considered misspelled. This study scope is limited to those with at least four characters and not more than 30 characters for individual words. For misspelled words, the closest words are selected as their correct forms from reference dictionaries based on the highest scores of similarity metrics. Two edit distance similarity metrics and three phonetic similarity metrics are considered individually and in weighted and unweighted combinations. In weighted combinations, the weighting is a measure of context-sensitivity based on word frequency. These elements in the empirical selection score formulae of fuzzy matching are derived from the research of Crowell et al. (2004) and Lai et al. (2015). Modifications include use of edit distance similarity instead of edit distance and use of the natural logarithm of frequency in some normalization schemes. The normalized edit distances aided the evaluation of fuzzy matching results from multiple approaches.

Although various drug name and English word dictionaries are built as references, not all words in the dictionaries are used during fuzzy matching. Dziadek et al. (2017) reported that their best method used a domain-specific corpus-based dictionary. The corpus-based dictionary is more relevant to the object domain. Also, its smaller size reduces computing complexity during fuzzy matching. Thus, we chose those words existing in both dictionaries and the Corpus as references.

Crowell et al. (2004) found that re-sorting by frequency of word occurrence in the medical domain could significantly improve spelling correction for medical queries. The sorted list also helps candidate selection when selection scores are equal. In this study, the algorithms take advantage of the sorted reference word list based on their frequencies in the domain.

N-gram language models have been used in many fields of natural language processing (Hirst, 2008) to incorporate context information into text correction. When $N > 1$, the model takes into account previous (N-1) words (Aouragh et al., 2015). Dziadek et al. (2017) found that trigram frequencies performed best in their context-sensitive spelling correction of clinical text. In this study, the single word count table served as a 1-gram model table providing frequencies of the single words in the domain, while a co-occurrence table is used to provide information similar to an N-gram model.

The server hardware and software used in the study were as follows:

Operation system: Microsoft Windows x86 64-bit

Number of processors: 4

CPU speed: 2.93GHz

RAM: 16 GB

Oracle database version: 11.2.0.4.0.

The most time-costly operations are computing spelling scores during fuzzy matching. The program stores the misspelled and corrected pairs in a table to reduce repetitive calculation of fuzzy matching for the same misspelled words or phrases. Because the Corpus is stored in Oracle databases, all algorithms and metric calculations are coded in PL/SQL in Oracle (11g version 11.2.0.4.0).

2.3. Dictionaries

For fuzzy matching, six reference dictionaries (five dictionaries of drug names and a dictionary of common English words) were used as references to construct five subject-matter-focused dictionaries used in the algorithm trials.

Reference dictionaries

The reference dictionary FDA_FAA_DrugNames contains 8,393 unique records. It consists of drug names from three U.S. government sources:

- A list of trade and generic names of U.S. Food and Drug Administration (FDA) approved drugs, downloaded at *https://clincalc.com/Pharmacy/DrugSpell.aspx*. The list consists of 7,369 records.
- A list of medication names (brand and generic) within each drug class from the Federal Air Surgeon MedList database used for pilots and Air Traffic Controller Specialist of FAA. It has 2,537 drug names.
- A list of drug names from an FAA autopsy database (Medical Analysis Tracking Registry [MANTRA]) at the Civil Aerospace Medical Institute. This table has 1,440 drug names.

Two reference dictionaries are compiled from the Drugs.com Drug Information Database (2020). In this database, individual drug (or drug-class) content compiled by Wolters Kluwer Health, American Society of Health-System Pharmacists, Cerner Multum, and IBM Watson Micromedex is peer-reviewed and delivered to users by Drugs.com. More information about the Drugs.com Drug Information Database can be found at *https://www.drugs.com/support/about.html*.

The reference dictionary DrugsComUSA contains 21,839 records. It consists of drug names used in the USA and is available at *https://www.drugs.com/drug_information.html*.

The reference dictionary DrugsComINT has 211,310 entries. It consists of drug names used internationally and is available at *https://www.drugs.com/international/*.

The reference dictionary English82K is a list of common English words from Github and is used for a program named SymSpell *https://github.com/wolfgarbe/SymSpell* (Garbe, 2019). This consists of 82,765 English words and their frequencies.

Trial dictionaries

The trial dictionaries are built from combinations of the reference dictionaries, with extra words and repeated words removed. These are:

- Dictionary 1: all drug names in FDA_FAA_DrugNames in descending order of frequency (8,393 drug names).
- Dictionary 2: drug names in Dictionary 1 used in the Corpus (4,066 drug names).
- Dictionary 3: drug names in Dictionary 2 and English words in English82K used in the Corpus (9,535 words).
- Dictionary 4: Dictionary 3 and drug names from DrugsComUSA and DrugsCom INT used in the Corpus. It also adds drug names with a dash and/or a space that match single words in the Corpus (15,414 words).
- Dictionary 5: Dictionary 4 without the words from DrugsComINT used in the Corpus (10,632 words).

2.4. Word and Phrase Tables

Prebuilt tables are used to increase the speed of fuzzy matching. These tables collect word frequency and co-occurrence, or correct words and phrases used in the Corpus for the fuzzy matching.

Single-word count table

The single word count table is a list of unique single words or strings and their frequencies in the Corpus. Corpus records are tokenized first. All non-alphabet characters are replaced with spaces, and the spaces are used as token boundaries, then all letters are changed to upper case. All unique tokens are counted. The table consists of 102,501 words or strings from approximately 5.3 million Corpus records.

Co-occurrence word table

The co-occurrence table contains context information measuring how close words appear together in individual Corpus records. It has three fields: word, wordAfter, and frequency. The frequency field value is a weighted co-occurrence frequency based on relative position: the weight is 1/N where the "wordAfter" is the N^{th} word after the "word."

Correct single word table

The correct single word table has those words shared by the single word count table and reference dictionaries. The purpose of this table is to improve performance in fuzzy matching in both computing speed and result.

Correct phrase table

The correct phrase table counts those Corpus records that match any entry in the drug name dictionaries after changing all to upper case and trimming spaces at both ends. It contains both single drug names and medication phrases since some Corpus records have only single drug names. This table is built for phrase fuzzy matching. It serves as part of the co-occurrence context feature

and, thus, helps overall correction. It also helps match those words with spaces in the middle, such as correcting “C IALIS” to “CIALIS.”

2.5. Validation Data Set of Drug Names

The validation data set was used to evaluate correction algorithms. It is a collection of reviewed pairs of misspelled and corrected drug names from FAA medication databases in DIWS and MANTRA. Thanks to previous DIWS-related reviewing efforts, a table named “Drug_Misspelled” in DIWS was available, which provided over 1,000 pairs of “Bad_Spelling” and “Corrected” drug names/phrases. The other source was a mapping of corrected drug names in MANTRA to the corresponding misspelled names in Corpus records in DIWS at the same medical examination identity. The validation data set has 955 unique pairs of misspelled and corrected drug names.

2.6. Sampling

Sample sets were used to test algorithm combinations after the validation set to explore the influences of various score weighting functions related to word and phrase frequencies (i.e., context-sensitivity). For each sample set, Oracle random package “DBMS_RANDOM” was used to randomly select 2,500 samples from a total of 5.3 million Corpus records. Each sample set was subdivided into five subsets of 500 samples.

2.7. Statistics

Differences of result metrics were tested with a Student’s t-test (paired two samples for means) for correction on same sample sets and single-factor analysis of variance for the metrics on different sample sets. The software used to perform the tests was the data analysis package within Microsoft Excel (2016).

2.8. Base Similarity Metrics

The base similarity metrics used in fuzzy matching included both orthographic and phonetic metrics. Two methods were used to calculate edit distance (orthographic) similarity: Levenshtein distance similarity and Jaro-Winkler distance similarity. Three metrics were used for phonetic coding: Soundex, Metaphone, and Double Metaphone.

Levenshtein edit distance similarity

One measure of the closeness is edit distance, also known as Levenshtein Distance, named after the Russian scientist Vladimir Levenshtein, who devised the algorithm in 1965 (Levenshtein, 1966). It is a measure of similarity between two strings: s1 and s2. The distance is the minimum number of insertions, deletions, or substitutions required to transform s1 to s2. For example, the edit distance between strings “shackleford” and “shackelford” is 2. Oracle (11g version 11.2.0.4.0) built-in package UTL_MATCH contains a function EDIT_DISTANCE_SIMILARITY. This function calculates the Levenshtein Distance between two strings by counting the number of character changes (inserts, updates, deletes) required to transform the first string into the second. It returns a normalized result ranging from 0 (no match) to 100 (complete match).

Jaro-Winkler edit distance similarity

The Jaro-Winkler algorithm is another way of calculating edit distance between two strings. This method, developed by the U.S. Census Bureau, is a string comparator measure that gives values of partial agreement between two strings (Williams, 2017). Jaro–Winkler distances are calculated based on the matching characters in two strings and exploit the idea that differences near the start of the string are more significant than differences near the end of the string (Winkler, 1990). The algorithm is suited for name entity matching (Christen, 2006; Cohen et al., 2003). Oracle (11g version 11.2.0.4.0) built-in package UTL_MATCH function Jaro_Winkler_Similarity was used to obtain Jaro-Winkler distances between strings. It returns a normalized result ranging from 0 (no match) to 100 (complete match).

Soundex phonetic similarity

Phonetic matching algorithms started with the Soundex algorithm, which was developed in the mid-20th century (Odell, 1956). The Soundex code consists of a letter followed by three numerical digits. Oracle (11g version 11.2.0.4.0) built-in function Soundex was used to obtain Soundex codes.

Metaphone phonetic similarity

In 1990, Lawrence Phillips published the Metaphone algorithm in *Computer Language* magazine (Philips, 1990). When compared to the Soundex algorithm, Metaphone yields a more reliable phonetic key by considering several English pronunciation rules. For this study, a modified PL/SQL function based on PL/SQL code from *http://www.geocities.ws/oracletricks/plsql/metaphone.txt*, following the procedure of Metaphone at *https://en.wikipedia.org/wiki/Metaphone* was written to calculate Metaphone codes (Philips, 1990).

Double Metaphone phonetic similarity

Metaphone failed to match some obviously homophonic words, such as “Bryan” and “Brian.” In 2000, Phillips published the Double Metaphone phonetic matching algorithm. This algorithm produces a second key representing the native pronunciation (Philips, 2000). Strong matches occur between primary keys of two strings, and weak matches occur between their alternate keys. Moderate matches are between a primary and an alternate key. This study uses an online PL/SQL package implementing the algorithm (*https://github.com/AliArdaOrhan/Double_Metaphone*), and primary keys are used for strong matching where Double Metaphone is mentioned.

2.9. Context Sensitivity

Context sensitivity can also help with spelling correction and text cleaning (Schierle et al., 2008). Context-based correction methods take advantage of two statistical measures: word frequency (F) and co-occurrence. Word frequency is the count of a unique word or token using the raw text corpus. Co-occurrence refers to how frequently words appear together in a similar context. For any two randomly chosen words, w_n and w_m, the expected probability for two statistically independent words w_n and w_m appearing in that order is (Eq. 1),

$$P(w_n w_m) = P(w_n)P(w_m). \quad (1)$$

They can be regarded as co-occurrent if the common appearance frequency of the two words is significantly higher than expected.

Frequencies are often normalized to the total number of occurrences in the corpus or occurrences per ten- or hundred-thousand. For this study, if normalized, frequencies were normalized to vary from 1 (least common) to 10,000 (most common) using equation 2,

$$F_{norm} = (F - F_{min}) \times (R_{max} - R_{min}) / (F_{max} - F_{min}) + R_{min}, \quad (2)$$

where F_{norm} is the normalized frequency, F is the un-normalized frequency, F_{min} is the minimum un-normalized frequency, F_{max} is the maximum un-normalized frequency, R_{max} is the maximum allowed value of F_{norm} (i.e., 10,000), and R_{min} is the lowest allowed value of F_{norm} (i.e., 1).

For calculations, words in the dictionary that do not occur with correct spelling in the Corpus are assigned an unnormalized frequency of 1 (instead of 0). The unnormalized frequency-based selection score weighting factors, B, used in this report are:

$$B_{word} = 1 + \ln(F), \quad (3)$$

and

$$B_{co\text{-}occurence} = 1 + \ln(Z) \quad (4)$$

where Z is the sum of F for the word pairs from the current Corpus record in the co-occurrence word table if there is at least one pair, else $B_{co\text{-}occurrence} = 1$.

2.10. Composite Similarity Metrics

The five base similarity metrics were combined with frequency-based weights to generate three composite similarity metrics: a spelling score (C_{spell}), a simple frequency-weighted composite selection score (C_1), and a normalized composite selection score (C_2) (Eqs. 5-7),

$$C_{spell}(w_i,w_j) = 2 \times D_{spell}(w_i,w_j) + D_{phone}(w_i,w_j), \quad (5)$$

$$C_1(w_i,w_j) = C_{spell}(w_i,w_j) \times B_{word}, \text{ and} \quad (6)$$

$$C_2(w_i,w_j) = (C_{spell}(w_i,w_j) + k \times \ln(F_{norm}(w_i))) \times 100 / (300 + k \times \ln(R_{max})), \quad (7)$$

where w_i is the word (or word group) in question from the corpus, w_j is a candidate correct word (or word group) in the dictionary (or fuzzing matching table), D_{spell} is the edit distance similarity of their spellings, D_{phone} is the edit distance similarity of their phonetic codes, and k is a weighting factor with a positive integer value.

When using normalized scoring, C_2, the selection confidence index (I) (Eqs. 8-10) is also calculated.

$$I = Q_{abs} + Q_{rel}. \quad (8)$$

$$Q_{abs} = C_{max}. \quad (9)$$

$$Q_{rel} = C_{max} - (C_{2nd} + C_{3rd}) / 2. \quad (10)$$

Here, Q_{abs} is the selection score C for the best match, C_{max}, and Q_{rel} quantifies the superiority of the best match when compared with the two next highest scoring candidates, C_{2nd} and C_{3rd}, respectively. The index is used in two ways. First, it is used to decide between the alternative correction sources when using normalized selection scores, words, or phrases from the Correct Phrase Table. For this use, the index is calculated for both possibilities (word and phrase), and correction with the highest index is selected as the final correction. Secondly, the index is used to match correction rates with the index, which allows algorithm effectiveness to be estimated for a database based on evaluations of samples.

2.11. Evaluation Metrics

Three metrics were used to evaluate results: misspelled rate, correction rate, and negative correction rate. The misspelled rate was defined as the percentage of the number of records misspelled (i.e., unmatched in the dictionary) relative to the total number of records. The correction rate was defined as the percentage of the number of rightly corrected records relative to the total number of misspelled records. The negative correction rate was defined as the percentage of false positives in the records counted as misspelled (i.e., cases of the drug name or word being correct but not in the reference dictionary).

3. RESULTS

3.1. Algorithm Development I: Evaluation of Base Similarity Metrics and Reference Dictionaries with the Validation Data Set of Drug Names.

Table 1 shows correction rates using various similarity metrics, both alone and in various combinations. Candidates with the highest selection scores were chosen as the correct form of the misspelled drug names. When more than one candidate shared the same selection score, the candidate with the highest frequency in the Corpus was chosen.

The Levenshtein edit distance similarity performed better than Jaro-Winkler similarity. Of the three phonetic coding metrics, Metaphone coding achieved the best result, and Soundex was the worst. Orthographic metrics (Levenshtein distance similarity and Jaro-Winkler similarity) resulted in higher correction percentages than phonetic metrics. Combining phonetic and orthographic metrics almost always improves correction when compared with single metric approaches.

A comparison of results of methods 3A and 3B with 2A and 2B indicates that the value of using a frequency factor is dictionary dependent. It was helpful in drug name correction when common English words were added to the reference dictionary (Dictionary 3). Otherwise, adding the

frequency factor to the selection formula resulted in lower correction rates. This is likely because common English words are not in the validation data set, and the extra words serve only to provide incorrect options for correction from the dictionary.

Table 1. *Correction of Drug Names in the Validation Data Set Using Various Dictionaries and Methods*

Methods		Correction Rate (%)		
		Dictionary 1	Dictionary 2	Dictionary 3
Single Metric	Soundex	58.85	69.01	68.17
	Double Metaphone	81.26	85.86	82.93
	Metaphone	88.38	90.68	88.90
	Jaro-Winkler distance similarity	86.07	88.06	82.93
	Levenshtein distance similarity	93.72	94.35	92.77
Composite Metrics	Method 1A	91.83	92.46	90.26
	Method 1B	92.15	92.98	91.20
	Method 1C	83.04	85.45	81.99
	Method 2A	93.40	93.93	92.04
	Method 2B	94.03	94.45	92.67
	Method 2C	91.62	92.77	90.89
	Method 3A	93.19	93.09	92.77
	Method 3B	93.30	93.19	92.98

Note:

Method 1A: $C = C_{spell}$ using Jaro-Winkler distance similarity and Double Metaphone.

Method 1B: $C = C_{spell}$ using Jaro-Winkler distance similarity and Metaphone.

Method 1C: $C = C_{spell}$ using Jaro-Winkler distance similarity and Soundex.

Method 2A: $C = C_{spell}$ using Levenshtein distance similarity and Double Metaphone.

Method 2B: $C = C_{spell}$ using Levenshtein distance similarity and Metaphone.

Method 2C: $C = C_{spell}$ using Levenshtein distance similarity and Soundex.

Method 3A: $C = C_1$ using Levenshtein distance similarity and Double Metaphone.

Method 3B: $C = C_1$ using Levenshtein distance similarity and Metaphone.

Restricting reference drug names to those only used in the Corpus improved the correction performances, especially for the single metric algorithms, when comparing results of Dictionary 1 and Dictionary 2. Using Dictionary 3, with its added common English words as the reference dictionary, reduces correction rates since it becomes harder to select from similar candidates.

In summary, metrics of Levenshtein distance similarity and Metaphone codes are the best combination and, thus, chosen for use in further algorithm development trials. Using drug names that exist in the Corpus as a reference dictionary improves the correction rate. Frequency weighting shifts from harmful to helpful as the dictionary size increases. Using a smaller and more relevant reference dictionary can result in a higher correction rate.

3.2. Algorithm Development II: Frequency Factor Selection

Unlike the validation data set, in actual records, a misspelled word could be a drug name or a common English word, and the reference dictionary should include both drug names and common English words. The validation set and literature review indicate a frequency factor could be beneficial when attempting to correct these records. Thus, we performed trials using various dictionaries and frequency factors to correct samples of the Corpus. Frequency factors investigated included single word count frequency, co-occurrence frequency when there is more than one word in a Corpus record, or both.

Table 2 reveals the correction results for samples of Corpus records. Two sets of 2,500 samples were randomly selected, one for Dictionary 3 and another for Dictionary 4. The three algorithms, using different frequency factors, were applied to each sample set. As expected, the additional drug names from Drugs.com in Dictionary 4 significantly reduced negative correction rates ($P < 0.01$). The much larger dictionary also lowered the misspelled rate since more words were recognized as correct, but the difference was not significant ($P = 0.057$). Both the misspelled rate and the negative correction rate were independent of correction algorithm selection because these two quantities depend only on the dictionary-corpus comparison. Another interesting result was the significantly improved correction rates ($P < 0.01$) when using Dictionary 4 instead of Dictionary 3, contrary to the general rule of smaller dictionaries generally giving better correction rates. A possible explanation for the improved correction rates when using the larger dictionary is that the added forms of drug names from Drugs.com in Dictionary 4 added better matches. As physicians do not submit aeromedical records using only U.S. government-approved names, this would not be surprising—they can use their personally preferred names or common drug name abbreviations.

Table 2. *Impact of Frequency Factor and Reference Dictionary on Correction of Medication Records*

	Misspelled Rate		Correction Rate		Negative Correction Rate	
	Dictionary 3	Dictionary 4	Dictionary 3	Dictionary 4	Dictionary 3	Dictionary 4
Algorithm 1	10.40	8.28	77.73	91.13	14.14	1.91
Algorithm 2	10.40	8.28	79.96	90.05	14.14	1.91
Algorithm 3	10.40	8.28	78.90	90.05	14.14	1.91

Note:

Algorithm 1: $C = C_{Spell} \times B_{word} \times B_{co\text{-}occurrence}$.

Algorithm 2: $C = C_{Spell} \times B_{word}$ if the record has only one word, else $S \times B_{co\text{-}occurrence}$.

Algorithm 3: $C = C_{Spell}$.

When examining some cases of wrong correction (as opposed to negative correction), the frequency factor had more impact than expected on selection scores when used as a multiplier in the formula. The frequency factors used did not consistently affect the correction rates, and no significant impacts on correction rates were observed ($P > 0.05$). A better approach was to better control the impact of the frequency factor in the selection scoring by normalization. This used Eq. 7 (i.e., score C_2), which normalized the composite selection score while retaining the ability to vary the impact of the frequency factor by adding a constant multiplying factor (k). Over 85% of Corpus records are single words. To simplify selection score calculation, the co-occurrence frequency factor was replaced with whole phrase matching using a table of correct phrases built as a reference (see above). This allowed two options for record correction: breaking down the record into individual words, correcting the single words using a reference dictionary of correct single words, and fuzzy matching the whole record with the correct phrase table. The correction process with the higher I (Eq. 8) score was selected as the final result for each Corpus record.

Table 3 provides results of this correction algorithm where k = 1, 5, 9, 13 and 17, respectively. Misspelled rates and negative correction rates were constant because they used the same set of 2,500 random samples and the same dictionary (and are different from Table 2 Dictionary 4 results because of a different sample). When comparing the best results in Table 3 (k = 13) with the best results in Table 2 (Algorithm 1 with Dictionary 4), the new algorithm significantly increases correction rates ($P = 0.03$).

Table 3. *Correction of Aviation Medication Records with C_2 and Selection Confidence Index*

k	Misspelled Rate	Correction Rate	Negative Correction Rate
1	9.48	89.48	0.40
5	9.48	94.04	0.40
9	9.48	95.72	0.40
13	9.48	96.34	0.40
17	9.48	94.59	0.40

Note: C_2 = normalized composite selections score.

For low k values, the correction rate increases with the value of k. For the chosen set of k values, the peak correction performance is observed at k = 13. Thus, 13 is chosen as the value of k for the best correction performance.

3.3. Algorithm Development III: Removal of International Drug Names from Reference Dictionary

The correction rates depend not only on the selection formula but also on the reference dictionary. A misspelled name in one field, country, or language can be a correct name in another. A good way to check a correction algorithm is to apply it to a validation data set. Table 4 displays the impact of international drug names in the reference dictionary on correction rates in the validation data set. The international drug names in Dictionary 4 but not in Dictionary 5 result in much lower correction rates: more than 100 misspelled drug names in the validation data set were correct international drug names.

This suggested that more accurate correction to U.S. names in the Corpus would be possible if international drug names were removed from the reference dictionary. To test the impact of Dictionary 5, the final algorithm was used to correct a new sample set of Corpus records. When k = 13, the new correction rate was 93.04, the misspelled rate was 9.68, and the negative correction rate was 1.32. When compared with the best result (k = 13) in Table 3, where the international drug names were included in the reference dictionary, the negative correction rate increased approximately 1%, and the correction rate went down approximately 3%. However, the change of correction rate was not significant ($P > 0.05$).

Table 4. *Correction of Drug Names in the Validation Set Using Selection Score C_2 With and Without International Drug Names in Reference Dictionaries*

k	Correction Rate with International Drug Names (Dictionary 4)	Correction Rate without International Drug Names (Dictionary 5)
3	80.42	95.71
7	85.03	96.34
9	85.86	96.44
13	86.18	96.54
17	85.97	95.71

Note: Where k is the multiplier in the score selection formula.

3.4. Correction of Drug Names in All Medication Records

Based on the evaluations of algorithms using the validation set and sample sets with various dictionaries and frequency-related weighting factors, the best correction algorithm is C_2 with k = 13 and Dictionary 5. To select the best frequency factor, results from both single word matching and phrase matching should be calculated, and the choice with a higher I value should be used as the final correction.

Table 5 reveals statistics of the correction of all Corpus records using the selected dictionary. Almost 90% of the records do not require any correction, and there are very small percentages of both empty records and records with extra dashes, spaces, or dots. After removing these extra characters, approximately 10% require attempted spelling correction.

Table 5. *Summary of Correction of Drug Names in All Medication Records (5.35 Million)*

Correction Status	Percentage	Note
Correct	88.67	No correction needed
Empty	0.01	Empty records
Trimmed	1.12	Removed extra space(s), dash(es), or dot(s)
Misspelled	10.20	Misspelling detected then corrected

To estimate the correction rate of all Corpus records, sample correction rates are first calculated at each I level and then applied to all records. For each interval, the number of records, the sample correction rates, and the estimated number of corrected misspelled records are displayed in Table 6. The total correction rate is 92.88%, close to samples when international drug names were removed from the reference dictionary (93.04%). Sample correction rates for I values lower than

55, which did not occur in the samples, are projected from a best-fitting trend curve that is not allowed to be negative.

Table 6. *Estimation of Correction Rate of All Medication Records Based on Sample Correction Rates at Each Selection Confidence Index Level*

Floor of Confidence Index Values	Number of Misspelled Records	Sample Correction Rate (%)	Estimated Corrected
40-44	16	0.00	0
45-49	421	0.00	0
50-54	3,581	0.00	0
55-59	7,230	25.00	1,807.5
60-64	10,025	16.67	1,671.17
65-69	14,611	60.00	8,766.6
70-74	16,208	70.00	11,345.6
75-79	19,465	88.89	17,302.44
80-84	26,270	77.27	20,298.83
85-89	29,537	100.00	29,537
90-94	37,258	93.94	35,000.17
95-99	44,533	100.00	44,533
100-104	52,250	100.00	52,250
105-109	47,766	100.00	47,766
110-114	63,709	100.00	63,709
115-119	69,745	100.00	69,745
120-124	77,014	100.00	77,014
125-129	17,953	100.00	17,953
>130	8,568	100.00	8,568
Subtotal	546,160		507,267.3
Estimated Total Correction Rate		92.88%	

Concerning space and time complexity, the correction costs approximately 36 hours to detect and correct the misspelled drug names and English words in the 5.3 million Corpus records using the

PL/SQL programs. This includes word counting and fuzzy matching of individual words and whole phrases.

4. DISCUSSION

This study does not consider such situations as misusing one correct drug name or English word as another. This includes at least two interesting categories. One is when a correct drug name was misspelled as another correct drug name, and the other is when a misspelled drug name was not corrected because the misspelling was a common English word. For example, there are over 30 Corpus records in which "aspiring" was initially not found as a misspelling of "aspirin." A PL/SQL stored procedure was developed in the study to pick up candidate pairs where a common English word could also be a misspelled drug name. However, it is challenging to automate verification of those pairs without reviewing details for the context of the use. For instance, "Aspiring" is also used as a medical term. Table 7 lists 10 of these pairs. More knowledge of context, including obtaining such metadata as medication type, is required to handle such situations properly.

Table 7. *Ten Common English Words That Sometimes are Misspelled Drug Names or Medical Terms*

Common English words	Drug (Medication) Names
ACCOLADE	ACCOLATE
ASPIRING	ASPIRIN
BOWELL	BOWEL
BOWL	BOWEL
DIVAN	DIOVAN
INTEL	INTAL
LOPED	LOPID
MERIDA	MERIDIA
SINGULAR	SINGULAIR
SINS	SINUS

Overall, the processing of drug name correction is not fast, but it is feasible. Jobs requiring significant computer time for correction can be set to run at night or on weekends. When the first-round corrections are complete, correcting new Corpus records becomes trivial.

Bryan et al. (2015) explored the medication name designation process and compliance with World Health Organization naming guidelines and found that International Nonproprietary Names have

greater potential for confusion. This may explain why adding international drug names to the reference dictionary reduced the correction rate of the validation data set in this study. However, some international drug names exist in the Corpus. It might be beneficial to include popular international drug names in the reference dictionary.

In addition to these international drug names, some Corpus records also contain chemical names, abbreviations, or a long name consisting of a few words squeezed together, among which drug names are sometimes misspelled. Examples are "ACETYLSALICYLACID," "DIPHENYLCYCLOPROPENONE," "DPCP," "IVIG," and "ASPIRINONEPERDAYFORPROFLAXIS." More effort is required to interpret or correct such chemical or pharmaceutical drug information.

The study results show that co-occurrence did not improve the correction rate. This may be due to the syntactic simplicity of most Corpus records. Of the 5.3 million Corpus records studied, over 85% contain single words. When the co-occurrence feature was removed from the correction process, adding phrase matching compensated for its removal. Phrase matching is easier to implement than co-occurrence. It also reduces the time and space complexities compared to using the Unified Medical Language System SPECIALIST lexicon (Lai et al., 2015) or building a database like Medline 5-Grams (Lu et al., 2015).

In this study, elements (spelling score and ln(frequency)) in the empirical selection score formula of fuzzy matching are derived from the research of Crowell et al. (2004) and Lai et al. (2015). However, edit distance similarity has been used instead of edit distance. The normalized edit distance helps in evaluating how well the fuzzy matching works across different approaches. A better modification was to use normalized frequency weighting as an addend (Eq. 7); the best results reached when k = 13 (i.e., the maximum value of the frequency part is 13*ln(10000) = 120). Thus, in the best fuzzy matching of this study, orthographic similarity, phonetic similarity, and word frequency contribute approximately 50%, 25%, and 25%, respectively.

Selection confidence indices are calculated to evaluate the effectiveness of fuzzy matching. The values enable users to separate corrections at various confidence levels and only accept those with satisfactory confidence. Future improvements could focus on those with low confidence.

Results suggest that in future studies, larger sample sizes could aid in selecting algorithms. Algorithms were selected based on shifts in the correction rate that were statistically insignificant and sometimes < 1%.

The final correction rate for all 5.3 million Corpus records was estimated to be approximately 92.88%. This process can be considered efficient, and it is comparable to the results of Workman et al. (2019) in which approximately 90% correction rates of clinical text were reported using Word2Vec, Levenshtein edit distance constraints, a lexical resource, and corpus term frequencies.

REFERENCES

Aouragh, L., Gueddah, H., & Yousfi, A. (2015). Adapting the Levenshtein distance to contextual spelling correction. *International Journal of Computer Science & Applications, 12*(1), 127-133.

Bryan, R., Aronson, J.K., ten Hacken, P., Williams, A., & Jordan, S. (2015). Patient safety in medication nomenclature: orthographic and semantic properties of international nonproprietary names. *PLoS One, 10*(12), e0145431.

Christen P. (2006). *A comparison of personal name matching: techniques and practical issues.* Department of Computer Science, the Australian National University, Canberra, Australia. Retrieved February 14, 2020 from *http://users.cecs.anu.edu.au/~Peter.Christen/publications/tr-cs-06-02.pdf*

Cohen, W. Ravikumar, P., & Fienberg, S. (2003). A Comparison of String Metrics for Matching Names and Records. *Proceedings of the IJCAI '03 Workshop on Information Integration on the Web* (pp. 73–78).

Crowell, J., Zeng, Q., Ngo, L., & Lacroix, E. M. (2004). A frequency-based technique to improve the spelling suggestion rank in medical queries. *Journal of the American Medical Informatics Association: JAMIA, 11*(3), 179–185. https://doi.org/10.1197/jamia.M1474

Dziadek, J., Henriksson, A., & Duneld, M. (2017). Improving Terminology Mapping in Clinical Text with Context-Sensitive Spelling Correction. *Studies in Health Technology and Informatics, 235*, 241–245.

Drugs.com. (2020). Retrieved January 8, 2020, from https://www.drugs.com

Ferner, R. E., & Aronson, J. K. (2016). Nominal ISOMERs (Incorrect Spellings Of Medicines Eluding Researchers)-variants in the spellings of drug names in PubMed: a database review. *BMJ (Clinical research ed.), 355*, i4854. https://doi.org/10.1136/bmj.i4854

Flor, M., & Futagi, Y. (2012). On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp.105–115).

Garbe, W. (2019). *SymSpell*. GitHub. Retrieved November 28, 2019, from https://github.com/wolfgarbe/SymSpell

Hirst, G. (2008). *An evaluation of the contextual spelling checker of Microsoft Office Word 2007*. Department of Computer Science University of Toronto, Toronto, Canada. Retrieved January 8, 2020, from ftp://www.cs.toronto.edu/public_html/public_html/pub/gh/Hirst-2008-Word.pdf

Hussain F., & Qamar, U. (2016). Identification and correction of misspelled drugs' names in electronic medical records (EMR). In *Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016)* (Volume 2, pp. 333-338).

Jiang, K., Chen, T., Huang, L., Calix, R. A., & Bernard, G. R. (2018). A Data-Driven Method of Discovering Misspellings of Medication Names on Twitter. *Studies in Health Technology and Informatics*, *247*, 136–140.

Lai, K. H., Topaz, M., Goss, F. R., & Zhou, L. (2015). Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, *55*, 188–195. https://doi.org/10.1016/j.jbi.2015.04.008

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lu, C. J., Tormey, D., McCreedy, L., &Browne, A. C. (2015). Generating the MEDLINE n-gram set. *AMIA 2015 Annual Symposium*, San Francisco, CA, November 14-18, (pp. 1569).

Microsoft Corporation. (2016). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel

Odell, M. K. (1956). The profit in records management. *Systems (New York)*, *20*, 20.

Philips, L. (1990). Hanging on the metaphone. *Computer Language, 7*(12), 39-44.

Philips, L. (2000). The double metaphone search algorithm. *C/C++ Users Journal, 18*:38-43.

Pimpalkhute, P., Patki, A., Nikfarjam, A., & Gonzalez, G. (2014). Phonetic spelling filter for keyword selection in drug mention mining from social media. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2014*, 90–95.

Schierle, M., Schulz, S., & Ackermann, M. (2008). From spelling correction to text cleaning–using context information. In *Data Analysis, Machine Learning and Applications* (pp. 397-404). Springer.

Senger, C., Kaltschmidt, J., Schmitt, S. P., Pruszydlo, M. G., & Haefeli, W. E. (2010). Misspellings in drug information system queries: characteristics of drug name spelling errors and strategies for their prevention. *International Journal of Medical Informatics*, *79*(12), 832–839. https://doi.org/10.1016/j.ijmedinf.2010.09.005

Williams, R. (2017). *UTL_Match and Soundex for Fuzzy Matching of Data in an Oracle Database*. A-Team Chronicles. https://www.ateam-oracle.com/utl_match-and-soundex-for-fuzzy-matching-of-data-in-an-oracle-database

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 354–359).

Workman, T. E., Shao, Y., Divita, G., & Zeng-Treitler, Q. (2019). An efficient prototype method to identify and correct misspellings in clinical text. *BMC Research Notes*, *12*(1), 42. https://doi.org/10.1186/s13104-019-4073-y