Soft Bigram Similarity to Identify Confusable Drug Names

Chapter · May 2019

DOI: 10.1007/978-3-030-21077-9_40

4 authors:

Christian Eduardo Millán-Hernández
Universidad Tecnológica de la Mixteca
7 PUBLICATIONS **86** CITATIONS
SEE PROFILE

René Arnulfo García Hernández
Universidad Autónoma del Estado de México (UAEM)
85 PUBLICATIONS **657** CITATIONS
SEE PROFILE

Yulia Ledeneva
Universidad Autónoma del Estado de México (UAEM)
75 PUBLICATIONS **529** CITATIONS
SEE PROFILE

Ángel Hernández-Castañeda
Universidad Autónoma del Estado de México (UAEM)
22 PUBLICATIONS **178** CITATIONS
SEE PROFILE

Soft Bigram Similarity to Identify Confusable Drug Names

Christian Eduardo Millán-Hernández[0000-0002-8683-7500], René Arnulfo García-Hernández[0000-0001-7941-377X], Yulia Ledeneva[0000-0003-0766-542X], and Ángel Hernández-Castañeda[0000-0003-0766-542X]

Autonomous University of State of Mexico, Toluca 50000, México
ceduardo.millan@gmail.com, renearnulfo@hotmail.com, yledeneva@yahoo.com, angelhc2305@gmail.com

Abstract. Look-alike and Sound-alike drug names are related to medication errors where doctors, nurses, and pharmacists prescribe and administer the wrong medication. Bisim similarity is reported as the best orthographic measure to identifying confusable drug names, but it lacks from a similarity scale between the bigrams of a drug name. In this paper, we propose a Soft-Bisim similarity measure that extends to the Bisim to soften the comparison scale between the Bigrams of a drug name for improving the detection of confusable drug names. In the experimentation, Soft-Bisim outperforms others 17 similarity measures for 396,900 pairs of drug names. In addition, the average of four measures is outperformed when Bisim is replaced by Soft-Bisim similarity.

Keywords: LASA Drug Names, N-gram Similarity, Orthographic Similarity.

1 Introduction

A medication error that involves confusable drug names occurs as result of weak medication system and human errors-related factors [1–3]. Many human factors are related to the Look-Alike and Sound-Alike drug names (LASA) problem like visual perception error, auditory perception error, short term memory error, and motor control are errors. However, the similarity between confusable drug names is a detectable root-cause. Drug names like *cycloserine* and *cyclosporine* are involved in LASA errors. LASA pairs normally sound similar and have a similar spelling [4]. Sometimes the confusion happens when the names are communicated in prescriptions handwritten, for example, the drugs *Avandia* and *Coumadin* [5]. In other cases, the confusion occurs in verbal communication when the pronunciation sounds similar. For example, *Zantac* and *Xanax* [6].

Nowadays, the Institute for Safe Medication Practice (ISMP) publishes a list that contains LASA pairs that were previously reported [7–10]. Regulatory agencies, including the Food and Drug Administration (FDA), the World Health Organization (WHO), and the Joint Commission are implementing strategies to identify and to prevent a LASA error.

String-matching algorithms are used to measure the distance or the similarity between two drug names and to identify a priori potential confused drug names. For example, *Edit Distance* (ED) measures the minimum of the insertion, elimination and substitution operations to transform a string to another [11]. For example, the LASA pair *cycloserine* and *cyclosporine* has a distance of two because there are needed at least two edit operations (a substitution $p \rightarrow e$ and an elimination of letter *o*) to transform *cycloserine* in *cyclosporine*.

Longest Common Subsequence (LCS) measures the maximum possible length of the longest common subsequences between two drug names. NLCS represents the Normalization of LCS that is obtained by dividing the maximum length of the longest drug name. In the previous example, /*cyclos-rine/* is the LCS and the NLCS is 0.833. NLCS presents a weakness to ignore subsequences that does not represent a similarity between drug names. For example, the no-LASA pair *Benadryl* and *Cardura* have the LCS /adr/ [6].

Ngram similarity represents a drug name as the set of all its contiguous subsequences (grams) of size *n* [12] [13]. For example, the bigrams for the LASA pair *cycloserine* and *cyclosporine* are {*cy*, *yc*, *cl*, *lo*, *os*, *se*, *er*, *ri*, *in*, *ne*} and {*cy*, *yc*, *cl*, *lo*, *os*, *sp*, *po*, *or*, *ri*, *in*, *ne*}, respectively. In this case, eight bigrams are shared, and the number of bigrams is 10 and 11, respectively. Therefore, the similarity is $(2 \times 8)/(10 + 11) = 0.762$. However, the Ngram similarity of the LASA pair *Verelan* and *Virilon* is zero [6].

Nsim similarity [14] extends to NLCS but it manages the *n*-grams of a drug name with a scale of similarity. The predefined scale of similarity between a pair of *n*-grams is computed by counting the identical matching letters in each position and normalized by *n*. *Bisim* is a specific case of *Nsim* with a predefined scale of similarity. For example, the bigrams ***cy*** and ***cy*** have a similarity of 2/2 =1 and the bigrams ***s****e* and ***s****p* of 1/2=0.5. The similarity scale presents a weakness when computes values for bigrams like *sp* and *ps*, or *sp* and *es*; because it misplaces completely the common letters in previous or next positions. This issue is a common root-cause when a visual or auditory perception error happens [15–17]. Even the pairs of bigrams {*aa*}{*aa*} and {*ac*}{*ac*} computes the same similarity, it is clear that in the first example the letter *a* match all the letters of the bigrams showing a higher similarity. In this manner, commonalities characteristics that are presented in LASA pairs [18] needs to be considered to adjust a softened similarity scale.

In this paper, we propose a new softened similarity measure based on Bisim that increase the accuracy to identify LASA pairs. For this, different cases that form the scale of bigrams are identified, and a proposed methodology based on an evolutionary algorithm to soften the scale of the similarity is described. Therefore, this paper is based on the hypothesis that an evolutionary approach can adjust better the weights of the scale of similarity between *n*-grams.

2 Definitions

String matching algorithms recover the common correspondences between the drug names that are used to determinate a similarity or a distance measure. Measures are classified as distance (as closer to zero as more related are the names) or similarity (as greater is the value as more related are the names). A normalized similarity/distance measure keeps a scale between different similarity values.

Similarity and distance measures detect the particular look-alike (orthographic cause) and sound-alike (phonetic case) issue. In this sense, the measures are classified as orthographic or phonetic in relation to the used approach to detect the confusion.

2.1 Orthographic distance measures

Edit distance (ED). Given the drug names *X* and *Y* as sequences of size *n* and *m*, respectively, ED (also called *Levenshtein*) refers to the minimum cost of editing operations (insertion, deletion and substitution) to convert the sequence *X* into *Y* [11, 19–21]. In this paper, all editing operations have a cost of 1. In this case, the edit distance between *X* and *Y* is given by edit (n, m) computed by the following recurrence:

$$edit(i,j) = \begin{cases} max(i,j) & i = 0 \vee \mathrm{j} = 0 \\ edit(i-1,j-1) & x_i = y_j \\ min\begin{cases} edit(i-1,j)+1 \\ edit(i,j-1)+1 \\ edit(i-1,j-1)+cs(x_i,y_i) \end{cases} & x_i \neq y_j \end{cases} \quad (1)$$

A Normalized ED (NED). NED is computed by dividing the ED between the length of the longer sequence [6, 21–25].

2.2 Orthographic similarity measures

Prefix Similarity. Given the drug names *X* and *Y* as sequences of size *m* and *n* respectively, *Prefix* represents the ratio of the longest contiguous common initial letters [6], see Eq. 2. The common prefix for drug names *Accutane* and *Accolate* is *Acc* (|*Acc*| = 3), and the normalized prefix similarity is 0.375.

$$Prefix(X,Y) = \frac{|x_1=y_1, x_2=y_2, \ldots, x_i=y_i|}{\max(X,Y)} \quad (2)$$

N-gram Similarity. Represents a sequence of the set of all its contiguous subsequences (grams) of size *n* [12]. For example, if $|X| = m$ and *n = 2* (bigrams), then $X' = \{x_1x_2, x_2x_3, \ldots, x_{m-1}x_m\}$ [6, 14, 26]. Given the sequences *X* and *Y*, the *n-gram similarity* is defined as the *Dice similarity* [27] between the sets *X'* and *Y'* in the next way:

$$Dice(X',Y') = \frac{2|X' \cap Y'|}{|X'|+|Y'|} \quad (3)$$

N-gram similarity presents a weakness because it is well-known that the prefixes and suffixes of the drug names are involved in their confusion [6, 18]. For increasing the sensitivity of the *N-gram similarity* some variations with respect to initial and final letters area applied. Lambert [14] proposes to add spaces (or a letter not included in the names) (B)efore and (A)fter in both drug names to make that the initial or final letters appear in one or more *n*-grams. Lambert experimented with the variants of Bigram-(1B, 1A, 1B1A and 1A) and Trigram-(1B, 1A, 1B1A, 2B, 2A, 2B2A, 1B2A and 2B1A) [14, 17, 28].

Normalized LCS (NLCS). *NLCS* similarity lets to maintain an order in the common matching letters. Given the sequences X and Y of size n and m, respectively, the *NLCS similarity* is defined as the ratio of the length of the longest common subsequences between X and Y, $NLCS = |lcs(n,m)|/max(m,n)$, where $lcs(n,m)$ can be calculated by the recurrence in Eq. (4) [6, 14, 23–25, 29].

$$lcs(i,j) = \begin{cases} 0, & i = 0 \vee j = 0 \\ lcs(i-1,j-1)+1, & x_i = y_j \\ max\big(lcs(i,j-1), lcs(i-1,j)\big) & x_i \neq y_j \end{cases} \tag{4}$$

Nsim Similarity. It is proposed by Kondrak [6, 23, 24] and it combines features implemented by grams of size β, non-crossing-links constraints and the first letter it is repeated at the begging of the drug name. A specific case of *Nsim* is the measure *Bisim* [6]. Given the sequences (with the first repeated letter) X and Y representing the drug names of size n and m, respectively, *Bisim similarity* is defined as:

$$Bisim(X,Y) = \frac{nsim(n,m)}{max(n,m)}$$

$$nsim(i,j) = \begin{cases} 0, & i = 0 \vee j = 0 \\ max\begin{cases} nsim(i,j-1), \\ nsim(i-1,j), \\ nsim(i-1,i-1) + \begin{matrix} in \\ other \\ case \end{matrix} \\ s\big(x_i x_{i+1}, y_j y_{j+1}\big), \end{cases} \end{cases} \tag{5}$$

$$s\big(x_i x_{i+1}, y_j y_{j+1}\big) = \frac{1}{2}\sum_{k=0}^{1} id(x_{i+k}, y_{j+k}) \tag{6}$$

$$id(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \tag{7}$$

2.3 Related work

Using a list of 1,127 LASA pairs and 1,127 non-LASA pairs, Lambert [14] evaluates 22 measures with ten-fold cross-validation technique and concludes that Trigram2B, NED and Editex [20] are the best measures to identify LASA pairs.

Kondrak [6, 23–25] proposes the orthographic Nsim similarity and the phonetic Aline similarity [30, 31] where the recall metric is used to evaluate the results of 12 measures with the USP LASA list [32] of 360 unique drug names. Kondrak [6] concludes that Bisim is the best orthographic measure. Bisim is used to create automated warning systems to identify potential LASA errors in prescription electronic systems

[4, 33] and in the software POCA by the FDA [6]. Furthermore, the average of Bisim, Aline, Prefix, and NED measures outperform to Bisim [6].

3 Proposed Method

In this paper, a Soften Bigram Similarity measure (Soft-Bisim) is proposed. First, the cases of bigrams involved in the scale of similarity in Soft-Bisim are described. After that, the fitness function used to find the weights in the scale of similarity by a genetic algorithm is described. Our hypothesis is that an evolutionary approach defines better the levels in the scale of similarity compared to the original similarity scale proposed by Kondrak in Bisim (cf. Eq. 7 & 8). In other words, we consider this problem as an evolutionary approach for optimizing the internal parameters of the similarity scale.

3.1 Definition of Soft-Bisim similarity

Given the drug names X and Y as sequences of size n and m, respectively, Soft-Bisim is defined as:

$$Soft - Bisim(X,Y) = \frac{Bisim(n,m)}{max(n,m)} \tag{8}$$

$$Bisim(i,j) = \begin{cases} 0, & i = 0 \vee \mathrm{j} = 0 \\ max \begin{cases} Bisim(i,j-1), \\ Bisim(i-1,j), \\ Bisim(i-1,i-1) + \begin{matrix} in \\ other \\ case \end{matrix} \\ s(x_i\, x_{i+1}, y_j\, y_{j+1}), \end{cases} \end{cases} \tag{9}$$

Where the proposed scale of similarity for Soft-Bisim is defined as:

$$s(a_{\mathrm{i}}a_{\mathrm{i+1}}, b_{\mathrm{j}}b_{\mathrm{j+1}}) = \begin{cases} w_1,\ a_{\mathrm{i+1}} = b_{\mathrm{j+1}} \neq a_{\mathrm{i}} \neq b_{\mathrm{j}} \\ w_2,\ a_{\mathrm{i}} \neq a_{\mathrm{i+1}} \neq b_{\mathrm{j}} \neq b_{\mathrm{j+1}} \\ w_3,\ a_{\mathrm{i}} = a_{\mathrm{i+1}} = b_{\mathrm{j}} \neq b_{\mathrm{j+1}} \vee a_{\mathrm{i}} = b_{\mathrm{j}} = b_{\mathrm{j+1}} \neq a_{\mathrm{i+1}} \\ w_4,\ a_{\mathrm{i}} = b_{\mathrm{j+1}} \wedge a_{\mathrm{i+1}} = b_{\mathrm{j}} \\ w_5,\ a_{\mathrm{i}} = b_{\mathrm{j+1}} \neq a_{\mathrm{i+1}} \neq b_{\mathrm{j}} \vee a_{\mathrm{i+1}} = b_{\mathrm{j}} \neq a_{\mathrm{i}} \neq b_{\mathrm{j+1}} \\ w_6,\ a_{\mathrm{i}} = a_{\mathrm{i+1}} = b_{\mathrm{j+1}} \neq b_{\mathrm{j}} \vee a_{\mathrm{i+1}} = b_{\mathrm{j}} = b_{\mathrm{j+1}} \neq a_{\mathrm{i}} \\ w_7,\ a_{\mathrm{i}} = b_{\mathrm{j}} \neq a_{\mathrm{i+1}} \neq b_{\mathrm{j+1}} \\ w_8,\ a_{\mathrm{i}} = a_{\mathrm{i+1}} = b_{\mathrm{j}} = b_{\mathrm{j+1}} \\ w_9,\ a_{\mathrm{i}} = b_{\mathrm{j}} \wedge a_{\mathrm{i+1}} = b_{\mathrm{j+1}} \end{cases} \tag{10}$$

For increasing accuracy to identify confusable drug names it is needed to find the set of weights $W = \{w_1, w_2, \ldots, w_9\}$ of the scale of similarity of Soft-Bisim. For this, a Genetic Algorithm is used [34–36].

3.2 Finding the Scale of Similarity for Soft-Bisim

The fitness function of the Genetic Algorithm is designed to evaluate each individual in relation to the objective to optimize.

The FDA reviews the similarity of a new drug name with all drug names that were previously registered. Therefore, the f-measure evaluation widely used in the information retrieval is used as the fitness function [37]. Given a LASA pair $(d_i, d_j) \in List\ of\ LASA\ pairs$, the f-measure for the query d_i evaluates the size of the set of retrieved drug names in ranking one (most similar drug names to the query d_i), but if d_j does not appears in the last set, the f-measure add the size of the retrieved drug names in the next ranking, until appears d_j. In this way, f-measure evaluates the ability to find a relevant drug name from a query. The f-measure could be obtained at every ranking (r). In fact, we desire to improve the f-measure in the top four rankings. Therefore, the fitness function computes a macro-averaging f-measure for the queries of all different drug names (set D) based on the sum of the first four rankings, see Eq. 11. In other words, the fitness function gives more relevance to the combination of weights in W (Eq. 10) that, after retrieving the queries of all different drug names with the Soft-Bisim measure, produces the best sum of the first four f-measure evaluation.

$$fitness(D) = \sum_{r=1}^{4} f - measure(D, r) \tag{11}$$

4 Results and discussion

In all the experiments, the ground truth USP-858 collection with 858 LASA pairs is used. The USP-858 contains 630 unique drug names, and it can generate 36,900 pairs of drug names. That means that 0.3% of LASA pairs must be recovered.

4.1 Calculating the Scale of Similarity for Soft-Bisim

Although, the genetic algorithm only optimizes the macro-averaging f-measure for the top four positions, the comparison in Table 1 shows an improvement, with respect to Bisim, in all positions of ranking to retrieve LASA pairs. As it is possible to observe, the weight w_9 for Soft-Bisim maintains a higher relevance than w_8 while in Bisim w_8 and w_9 are the same. On the contrary, the case when all the letters are different the weight is not zero.

Table 1. Comparison of the macro-averaging f-measure evaluation for Bisim and Soft-Bisim with the USP-858 collection where the resulting weights for Soft-Bisim are: $w_1 = 0$, $w_2 = 0.1$, $w_3 = 0.4$, $w_4 = 0$, $w_5 = 0$, $w_6 = 0.2$, $w_7 = 0.4$, $w_8 = 0.6$ and $w_9 = 0.8$; and the implicit weights for Bisim are: $w_{1\ldots3} = 0$, $w_{4\ldots7} = 0.5$, $w_7 = 0$, and $w_8 = w_9 = 1$. *In the last row the ten-fold cross-validation results are showed.

Ranking		1	2	3	4	5	6	7	8	9	10
Bisim	F-Meas.	48.86	39.78	29.74	22.45	17.18	13.44	10.64	8.75	7.24	5.96

Soft-Bisim	ΣF-Meas.	48.86	88.65	118.40	140.86	158.04	171.49	182.14	190.90	198.14	204.11
	F-Meas.	51.07	45.20	39.03	33.45	28.89	25.75	23.08	20.93	19.07	17.52
Soft-Bisim*	ΣF-Meas.	51.07	96.27	135.31	168.76	197.65	223.41	246.49	267.42	286.50	304.03
	F-Meas.	51	44.75	38.79	33.51	29.18	25.97	23.33	21.21	19.38	17.85
	ΣF-Meas.	51	95.75	134.53	168.05	197.23	223.2	246.53	267.74	287.12	304.97

4.2 Evaluation of orthographic measures

In Fig. 1, Soft-Bisim is compared to all orthographic measures presented in section 2.1 and 2.2. In this case, Trigram-2B maintains the relevance indicated by Lambert but Bisim is more relevant that Trigram2B. It is worth to mention that Trigram2B2A and Trigram2B1A are more relevant that Bisim. However, Soft-Bisim obtains the best performance with the adjusted similarity scale.

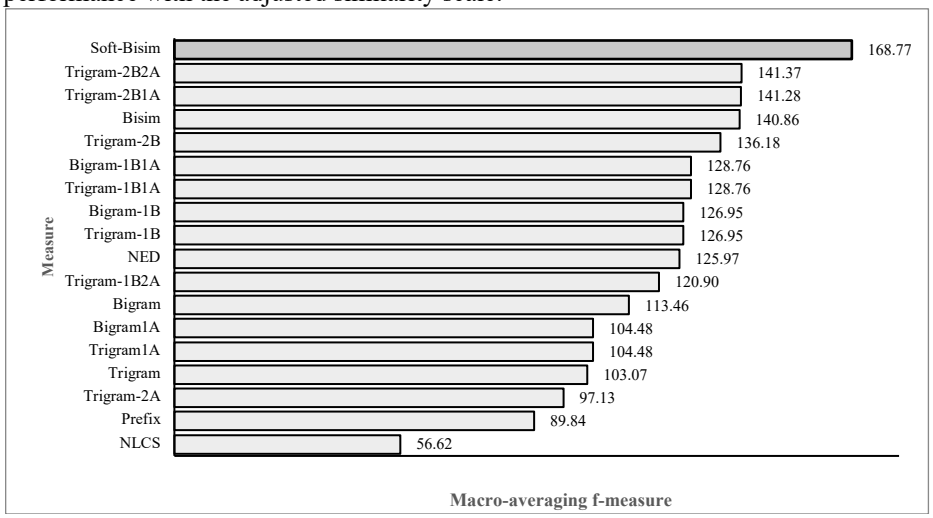

Fig. 1. Ranking obtained for each orthographic measure according to sum of top four positions of macro-averaging f-measure.

4.3 A combined measure with Soft-Bisim

Using the Average of Prefix, NED, Bisim and Aline, Kondrak [6] proposes a combined measure that outperform to Bisim: Avg_{Bisim}(Prefix, NED, Bisim, Aline). In this paper, we propose two combined measures, in the first one, Soft-Bisim is added to the average Avg_{all} (Prefix, NED, Bisim, Aline, Soft-Bisim), and the second one, Bisim is replaced by Soft-Bisim in the average, $Avg_{SoftBisim}$(Prefix, NED, Aline, Soft-Bisim). In table 2, the comparison of original combined proposed by Kondrak and our proposed combined measures are presented.

Table 2. Macro-averaging f-measure evaluation for $Avg_{Bisim,}$ Avg_{All} and $Avg_{SoftBisim}$.

Ranking	Avg_{Bisim}		Avg_{All}		$Avg_{SoftBisim}$	
	F-Measure	ΣF-Measure	F-Measure	ΣF-Measure	F-Measure	ΣF-Measure

1	51.36	51.36	51.63	51.63	51.70	51.70
2	44.69	96.05	45.56	97.20	45.42	97.12
3	39.63	135.68	39.78	136.98	40.01	137.13
4	35.11	**170.80**	34.87	**171.85**	35.13	**172.27**
5	30.79	201.59	30.72	202.58	30.89	203.16
6	27.55	229.15	27.76	230.34	27.85	231.02
7	25.04	254.19	25.05	255.39	25.10	256.12
8	22.81	277.00	22.87	278.27	22.86	278.98
9	20.84	297.85	20.92	299.19	21.06	300.04
10	19.31	317.17	19.45	318.64	19.47	319.52

In Table 3, using Bisim as a Baseline measure the best measures to identify confuse drug names are showed. In this case, all the combined measures outperform to the individual measures. However, the best individual measure is Soft-Bisim that it is involved in the first two combined measures. Moreover, the best performance is achieved when Bisim is replaced by Soft-Bisim.

Table 3. Comparison of Soft-Bisim with the best previous measures.

Rank	1	2	3	4	5	6	7
Measure	**$Avg_{SoftBisim}$**	**Avg_{all}**	Avg_{Bisim}	**Soft-Bisim**	Trigram-2B2A	Trigram-2B1A	Bisim
F-Meas.	172.27	171.85	170.80	168.76	141.37	141.27	140.86
p-value	0.005	0.005	0.005	0.005	0.005	0.005	Baseline

5 Conclusion

The problem of confusion of drug names needs attention because it is still growing. All measures presented in this paper (except by Nsim) are designed or adjusted to different application or domain. In this sense, Nsim takes into consideration characteristics that take part on confusable drug names like the fact that the initial letters are frequently involved in a confused drug name. In this paper we propose to Soft-Bisim measure that it is a new orthographic measure for identifying LASA pairs based on Nsim similarity with an extension to soften the scale of similarity between the bigrams that conforms a drug name. In specific, nine combinations of weights were calculated. For this, the sum the first-four macro-averaging f-measure of the retrieved pairs is proposed as the fitness function in a genetic algorithm.

According to the experimentation, Soft-Bisim increases the accuracy with respect to Bisim in a retrieved list of potential LASA pairs in all the ranking positions. Furthermore, Soft-Bisim outperforms significantly to the others 17 orthographic measures used in this problem. In this paper, we found that the measures Trigram-2B2A and Trigram-2B1A are good measures since outperform to the Bisim measure.

In addition, a new average combination of four measures using Soft-Bisim is proposed. This new average combination outperforms to the previous average that use Bisim measure. Even thought, we only use a list of drug names Soft-Bisim can be used to retrieve other cases of confusions like in proper names or brand names.

References

1. Billstein-Leber, M., Carrillo, C.J.D., Cassano, A.T., Moline, K., Robertson, J.J.: ASHP guidelines on preventing medication errors in hospitals, www.ashp.org/Pharmacy-Practice/Policy-, (2018).
2. Cohen, M.R., Domizio, G.D., Lee, R.E.: The role of drug names in medication errors. Medicat. errors. Wahihgton, DC Am. Pharm. Assoc. 87–110 (2007).
3. Medication without harm.: Medication Without Harm. World Health Organization, Geneva (2017).
4. Rash-Foanio, C., Galanter, W., Bryson, M., Falck, S., Liu, K.L., Schiff, G.D., Vaida, A., Lambert, B.L.: Automated detection of look-alike/sound-alike medication errors. Am. J. Heal. Pharm. 74, 521–527 (2017).
5. Tittemore, L.M.: The name game, https://sunsteinlaw.com/l-tittemore/, (2017).
6. Kondrak, G., Dorr, B.: Automatic identification of confusable drug names. Artif. Intell. Med. 36, 29–42 (2006).
7. FDA: FDA and ISMP Work to Prevent Medication Errors. 2017, (2012).
8. Craigle, V.: MedWatch: The FDA Safety Information and Adverse Event Reporting Program. J. Med. Libr. Assoc. 95, 224–225 (2007).
9. Gershman, J.A., Fass, A.D.: Medication Safety and Pharmacovigilance Resources for the Ambulatory Care Setting: Enhancing Patient Safety. Hosp. Pharm. 49, 363–368 (2014).
10. Getz, K.A., Stergiopoulos, S., Kaitin, K.I.: Evaluating the completeness and accuracy of MedWatch data. Am. J. Ther. 21, 442–446 (2014).
11. Wagner, R.A., Fischer, M.J.: The String-to-String Correction Problem. J. ACM. 21, 168–173 (1974).
12. Pfeifer, U., Poersch, T., Fuhr, N., Vi, L.I.: Searching Proper Names in Databases. In: HIM. pp. 259–275. Citeseer (1995).
13. Pfeifer, U., VI, L.I.: Searching Proper Names in Databases. October. 20, 1–13 (1994).
14. Lambert, B.L., Lin, S.J., Chang, K.Y., Gandhi, S.K.: Similarity as a risk factor in drug-name confusion errors: The look-alike (orthographic) and sound-alike (phonetic) model. Med. Care. 37, 1214–1225 (1999).
15. Schroeder, S.R., Salomon, M.M., Galanter, W.L., Schiff, G.D., Vaida, A.J., Gaunt, M.J., Bryson, M.L., Rash, C., Falck, S., Lambert, B.L.: Cognitive tests predict real-world errors: The relationship between drug name confusion rates in laboratory-based memory and perception tests and corresponding error rates in large pharmacy chains. BMJ Qual. Saf. 26, 395–407 (2017).
16. Lambert, B.L., Dickey, L.W., Fisher, W.M., Gibbons, R.D., Lin, S.-J., Luce, P.A., McLennan, C.T., Senders, J.W., Clement, T.Y.: Listen carefully: the risk of error in spoken medication orders. Soc. Sci. Med. 70, 1599–1608 (2010).
17. Lambert, B.L.: Predicting look-alike and sound-alike medication errors. Am. J. Heal. Pharm. 54, 1161–1171 (1997).
18. Shah, M.B., Merchant, L., Chan, I.Z., Taylor, K.: Characteristics that may help in the identification of potentially confusing proprietary drug names. Ther. Innov. Regul. Sci. 51, 232–236 (2017).
19. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. pp. 707–710 (1966).

20. Zobel, J., Box, G.P.O., Dart, P.: Phonetic String Matching : Lessons from Information Retrieval. Proc. 19th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. 166–172 (1996).
21. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. IEEE Trans. Knowl. Data Eng. 19, 1–16 (2007).
22. Chen, S., Liu, Y., Wei, L., Guan, B.: PS-FW: A Hybrid Algorithm Based on Particle Swarm and Fireworks for Global Optimization. Comput. Intell. Neurosci. 2018, 1–27 (2018).
23. Kondrak, G., Dorr, B.: Identification of confusable drug names: a new approach and evaluation methodology, (2004).
24. Kondrak, G., Dorr, B.J.: A similarity-based approach and evaluation methodology for reduction of drug name confusion. ALBERTA UNIV EDMONTON (2003).
25. Kondrak, G.: N-gram similarity and distance. Lect. Notes Comput. Sci. 3772, 115–126 (2005).
26. Chen, L.-C., Chen, C.-H., Chen, H.-M., Tseng, V.S.: Hybrid data mining approaches for prevention of drug dispensing errors. J. Intell. Inf. Syst. 36, 305–327 (2011).
27. Adamson, G.W., Boreham, J.: The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Inf. storage Retr. 10, 253–260 (1974).
28. Lambert, B.L., Chang, K.-Y., Lin, S.-J.: Effect of orthographic and phonological similarity on false recognition of drug names. Soc. Sci. Med. 52, 1843–1857 (2001).
29. Lambert, B.L., Yu, C., Thirumalai, M.: A system for multiattribute drug product comparison. J. Med. Syst. 28, 31–56 (2004).
30. Kondrak, G.: Phonetic alignment and similarity. Comput. Hum. 37, 273–291 (2003).
31. Kondrak, G.: Algorithms for Language Reconstruction, (2002).
32. USP: USP quality review (76). US Pharmacopeia. (2001).
33. Or, C.K.L., Wang, H.H.L.: Examining text enhancement methods to improve look-alike drug name differentiation accuracy. In: Proceedings of the Human Factors and Ergonomics Society. pp. 645–649 (2013).
34. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley, Reading, Massachusetts (1989).
35. Holland, J.H.: Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press (1992).
36. Mitchell, M.: An introduction to genetic algorithms. Cambridge, Massachusetts London, England, Fifth printing (1999).
37. Croft, B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company (2009).