Similarity As a Risk Factor in Drug-Name Confusion Errors: The Look-Alike (Orthographic) and Sound-Alike (Phonetic) Model

Author(s): Bruce L. Lambert, Swu-Jane Lin, Ken-Yu Chang and Sanjay K. Gandhi

Source: *Medical Care*, Dec., 1999, Vol. 37, No. 12 (Dec., 1999), pp. 1214-1225

Published by: Lippincott Williams & Wilkins

Stable URL: http://www.jstor.com/stable/3766938

MEDICAL CARE
Volume 37, Number 12, pp 1214–1225

Similarity As a Risk Factor in Drug-Name Confusion Errors

The Look-Alike (Orthographic) and Sound-Alike (Phonetic) Model

BRUCE L. LAMBERT, PHD,*† SWU-JANE LIN, BPHARM,* MSA, KEN-YU CHANG, BPHARM, MPH,* AND SANJAY K. GANDHI, BPHARM, PHD*†

BACKGROUND. One of every four medication errors reported in the United States is a name-confusion error. The rate of name-confusion errors might be reduced if new and confusing names were not allowed on the market and if safeguards could be put in place to avoid confusion between existing names.

OBJECTIVES. To evaluate several prognostic tests of drug-name confusion, alone and in combination, with respect to their sensitivity, specificity, and overall accuracy.

RESEARCH DESIGN. Case-control study. Twenty-two different computerized measures of orthographic similarity, orthographic distance, and phonetic similarity were used to compute similarity/distance scores for n = 1,127 cases (ie, pairs of names that appeared in published error reports or national error databases) and n = 1,127 controls.

MAIN OUTCOME MEASURES. Mean similarity/distance scores were compared across cases and controls. The performance of each measure at distinguishing between cases and controls was evaluated by tenfold crossvalidation. Dose-response relationships were examined. Univariate and multivariate logistic regression models were formed and evaluated by 10 fold crossvalidation.

RESULTS. Cases had significantly higher similarity scores than controls. Every measure of similarity proved to be a significant risk factor for error. There was a significant increasing trend in the odds-ratio as a function of similarity. A three-predictor logistic regression model had crossvalidated sensitivity of 93.7%, specificity of 95.9% and accuracy of 94.8%.

CONCLUSIONS. A sensitive and specific test of drug-name confusion potential can be formed using objective measures of orthographic similarity, orthographic distance, and phonetic distance.

Key words: medication errors; linguistics; terminology; drug names; drug approval; orthography; phonetics; psycholinguistics; look-alike; sound-alike. (Med Care 1999;37:1214–1225)

Roughly one of every four errors reported to Medication Error Reporting Program (MERP, administered by the Institute for Safe Medication Practices and the US Pharmacopeia, Inc.) involves

*From the Department of Pharmacy Administration, University of Illinois at Chicago, Chicago, Illinois.

†From the Department of Pharmacy Practice, University of Illinois at Chicago, Chicago, Illinois.

‡From Searle, Skokie, Illinois.

This research was supported in part by the US Pharmacopeia, Inc. (USP) Fellowship Program, the Drug Information Association, the US Food and Drug Administration, and the Campus Research Board of the University of Illinois at Chicago.

Aspects of this research were reported at the conference on Enhancing Patient Safety and Reducing Errors in Health Care, November 8–10, 1998, Rancho Mirage, CA.

Address correspondence to: Bruce L. Lambert, PhD, 833 S. Wood Street (M/C 871), Chicago, IL 60612–7231. E-mail: lambertb@uic.edu

Received March 12, 1999; initial review completed May 7, 1999; accepted June 4, 1999.

a pair of drugs whose names look or sound alike (eg, amiodarone and amrinone; E-Vista [Seatrace Pharmaceuticals, Inc., Gadsden, AL] and Evista [Lilly, Indianapolis, IN]; cisplatin and carboplatin; Celebrex [GD Searle & Co., Chicago, IL] and Cerebyx [Parke-Davis, Morris Plains, NJ]; Dynacin [Medicis Pharmaceutical Corp., Phoenix, AZ] and DynaCirc [Novartis Pharmaceuticals Corp., East Hanover, NJ]; Retrovir [Glaxo Wellcome, Research Triangle Park, NC] and Ritonavir [Abbott Laboratories, North Chicago, IL]).[1–5] Various organizations strive to prevent confusing new names from reaching the marketplace. Proposed names are subjected to pre-approval screening by the pharmaceutical manufacturers, the International Nonproprietary Name Committee of the World Health Organization, the United States Adopted Names (USAN) Council, the US Pharmacopeia, Inc., the US Patent and Trademark Office, and the US Food and Drug Administration (FDA).[6–11] Despite these efforts, new names that are similar to existing names continue to be approved, and name-confusion errors continue to occur.

Failures in the name review process occur partially because reviewing organizations have different goals. For example, drug companies want trademarks that will facilitate recognition and recall, distinguish products from the competition, and generate brand loyalty.[8] International Nonproprietary Name and USAN want nonproprietary names that are reasonably free from confusing similarities while remaining useful to health professionals. USAN accomplishes this by using a standard set of stems (eg, -ac for anti-inflammatory agents such as bromfenac, -mab for monoclonal antibodies such as abciximab).[12] The US Pharmacopeia, Inc. wants established names that are consistent with existing compendial nomenclature.[8] The US Patent and Trademark Office wants to prevent new trademarks from harming the commercial interests of firms holding existing marks.[8,13] The FDA wants to minimize threats to patient safety that may result from use of a confusing name. Unlike USAN, International Nonproprietary Name, and US Patent and Trademark Office reviews, FDA reviews consider indication, dosage form, and dosing schedule (much of this information is not even available when USAN designations are granted).[8] Thus, name review processes are motivated by multiple, competing goals and decisions are based on multiple, conflicting criteria.

Current pre-approval screening methods are also hindered by the lack of a systematic procedure for integrating computerized searching and expert human review. Drug manufacturers routinely conduct computerized screening of proposed names, but regulatory agencies do not require results of these computer searches to be submitted as part of the evaluation process.[7,8,10,11,14,15] Instead, expert panels rely on subjective judgments about the acceptability of new names.[8] Although agencies who provide names have established guidelines for the acceptability of proposed names, these guidelines are vague and are subject to varying interpretation.[7,8,16] The lack of objective criteria is compounded by the enormity of the review task. With roughly half a million pharmaceutical trademarks registered in the major industrialized countries alone, unaided experts could never consider the full range of names that might conflict with one another. As a result of these and other deficiencies in the name review and approval processes, confusing names continue to slip through the cracks. Furthermore, society lacks any effective procedure for modifying the drug name once it has been determined that nomenclature contributes significantly to medication errors.

An improved system for evaluating the acceptability of new drug names would integrate expert judgment and computerized name searches into a systematic and scientifically valid manner. Such a process would let computers do what they do best, to recall information from large databases quickly, precisely, and comprehensively. At the same time, human experts would be empowered to do what they do best, to make abstract judgements about complex patterns and real-time practice factors (eg, poor handwriting, abbreviations, storage of drug products on shelves or crash carts, stress, fatigue, and distractions).

The envisioned review system would rely on validated, objective measures of lexical (ie, word-to-word) similarity. Many such measures exist already. In fact, techniques for computerized searching of trademark databases are well-developed and are routinely used by trademark attorneys.[14,17–19] One problem with commercial search algorithms is that they are kept as trade secrets. They have not been described in peer-reviewed publications and their performance has not been adequately assessed. Fortunately, the computer science literature describes standard methods for computing orthographic (ie, spelling) similarity between words.[20–22] In addition, pho-

netic (ie, sound-based) methods continue to be developed as tools for searching databases of proper names.[19,23,24] If these methods could be validated and systematically combined with evaluation by human experts, the probability of allowing a confusing new drug name to reach the marketplace might be reduced.

Initial work on the development and evaluation of a computerized name searching system has previously been published.[25] In that previous, case-control study, three different measures were used to calculate orthographic (ie, spelling) similarity scores for drug pairs. It was shown that orthographic similarity was a significant risk factor for errors associated with name confusion. A prognostic test, based on orthographic similarity, was able to distinguish between cases and controls with 94% accuracy (when accuracy was measured by resubstitution).

The goal of the present study was to identify and validate objective measures of lexical similarity that could serve as the basis for a computerized drug-name searching system. The objective was to evaluate several measures of drug-name confusion, alone and in combination, with respect to their ability to distinguish between errors (ie, pairs of names reported to have been confused) and matched controls. This study expanded on the previous work by evaluating additional measures of orthographic similarity, adding phonetic (ie, sound) similarity measurements to the model, assessing dose-response relationships, building multivariate models, and using more sophisticated techniques (ie, crossvalidation) to assess predictive accuracy. We expected that crossvalidation would yield more conservative estimates of predictive accuracy than those made by resubstitution. We also expected a multivariate model, including orthographic and phonetic measures, that would perform better than univariate, orthographic models.

Patients and Methods

The study used a case-control design to examine the relationship between similarity and the probability of drug-name confusion errors.

Data

In the following analyses, the pair of names was the unit of analysis. Cases (n = 1,127 pairs) were drawn from published reports of look-alike/sound-alike medication errors.[2,3,26–28] The cases included those used in an earlier investigation and 158 new cases added from additional published sources.[25] Controls were selected by a two-step process. First, all of the individual names among the cases were listed (a total of 2,254 names). Duplicate names were deleted which left 1,423 unique (ie, nonrepeated) names. Names from this list were then randomly paired to create 1,127 controls. For example, among the cases, Achromycin (Lederle Laboratories, Pearl River, NY) was paired with Adriamycin (Pharmacia and Upjohn, Kalamazoo, MI) and bupivacaine was paired with Mapivacaine (Astra USA, Westborough, MA). One of the control pairs generated at random was Achromycin and bupivacaine. The process of selecting control pairs was purely random except for the following constraints: repeated control pairs were not allowed, and no control pairs included the same name (ie, amoxicillin and amoxicillin). This method of selecting controls minimized confounding by word length, word frequency, and other possible factors because both cases and controls were drawn from the same population of individual names; only the pairings were different. A similar approach has been used in the analysis of speech errors[29] (Note: It would be unusual to select controls from among the cases if the unit of analysis were the individual, but here the pair was the unit of analysis. Similarity is a property of pairs, not individual names). With this sample size, and two-tailed alpha set to 0.05, t tests had greater than 99% power to detect small effects. Correlational analyses had greater than 90% power to detect small effects.[30] Assuming a 1% exposure rate among controls, tests of association had greater than 90% power to detect an odds-ratio greater than or equal to 3.[31]

Measures

Twenty-two different measures of lexical (ie, word-to-word) similarity were evaluated. These measures were grouped into three categories: orthographic similarity, orthographic distance, and phonetic distance.

Orthographic Similarity. The term "orthographic" refers to the spelling of a word; "phonetic" refers to the sound pattern of a word. The first category, orthographic similarity, was comprised primarily of n-gram measures (the names of var-

ious measures are printed in boldface type to facilitate crossreferencing between tables and text). **N-gram** measures computed similarity by breaking words down into n-letter subsequences and then by counting the subsequences that occurred in both words.[20,25] The **bigram** method used two-letter subsequences and the **trigram** method used three-letter subsequences. For example, to compute the similarity between Acthar and Acular and for Acular each word was broken down into its two-letter subsequences. For Acthar, this yielded (ac, ct, th, ha, ar) and for Acular (ac, cu, ul, la, ar). All n-grams were converted to lowercase letters before comparisons were made. The Dice coefficient was used to compute a similarity score between sets of bigrams[32]:

$$Similarity = 2C/(B + A)$$

in which A was the number of bigrams in the first word, B the number of bigrams in the second word, and C the number of bigrams that occur in both words. Acthar (Rhône-Poulenc Rorer, Collegeville, PA) and Acular (Allergan, Irvine, CA) share two bigrams: (ac, ar). Hence, the bigram similarity between Acthar and Acular is

$$2*2/(5 + 5) = 0.4.$$

To increase sensitivity to the beginnings and endings of words, spaces were added before and after each word in some methods. Thus, the **bigram-1b1a** method used two-letter subsequences and added one space before and after each word. Eleven of the 22 measures used some variant of the n-gram method. Also in this category was a measure called **longest common subsequence**. The **longest common subsequence** between two character strings is "a subsequence common to both having maximal length, ie, it is at least as long as any other common subsequence of the strings."[20] The length of the **longest common subsequence** was used as the numerical similarity measure here. Thus, the **longest common subsequence** between Acthar and Acular is "acar," with a sequence of length 4. Note that the letters in the common subsequence need not be adjacent in the original sequences.[20]

Orthographic Distance. The second category, orthographic distance,[25] was comprised of **edit-distance** measures (previously termed Levenshtein distance). **"Edit distance"** refers to the number of edits (ie, insertions, deletions, or substitutions) required to transform one word into another.[20,25] To transform Ambien (GD Searle & Co., Chicago, IL) into Amen (Carnrick Laboratories, Inc., Cedar Knolls, NJ), one must delete the b and the i, so the **edit distance** between Ambien and Amen equals 2. In addition to raw **edit distance**, this category also included a **normalized edit distance**, in which the raw **edit distance** was divided by the maximum possible **edit distance** between two given words (ie, the length of the longer of the two words). Thus, the **normalized edit distance** between Ambien and Amen is

$$2/6 = 0.33.$$

Soundex Code	0	1	2	3	4	5	6
Letters	aeiouyhw	bpfv	cgjkqsxz	dt	l	mn	r

FIG. 1.

Phonetic Distance. The third category, phonetic distance, included hybrid measures that combined edit distance with various phonetic transcription methods. Phonetic transcription methods transform the orthographic representation of a word into a representation that is designed to capture regularities in the sound pattern of English. The phonetic transformation methods included **soundex, phonix, editex, tapered edit distance, omission key, and skeleton key**.[23,33–35] Space limits prevent the description of each of these measures in detail (interested readers should consult the references given earlier or query the author for detailed descriptions). A couple of examples illustrate the general approach. In the **soundex** system, letters are recoded as numbers according to the scheme outlined in Fig. 1.[36]

The **soundex** system leaves the first letter alone, recodes all subsequent letters, and deletes all 0 codes (ie, vowels, letter "h", and letter "w").

Editex Code	0	1	2	3	4	5	6	7	8	9
Letters	aeiouy	bp	ckq	dt	lr	mn	gj	fpv	sxz	csz

FIG. 2.

TABLE 1. Descriptive Statistics for Similarity/Distance Measures

		Cases (n = 1,127)			Controls (n = 1,127)			Overall (n = 2,254)		
Type	Measure	Mean*	SD	Range	Mean*	SD	Range	Mean	SD	Range
	Bigram	0.422	0.206	0–1	0.066	0.096	0–0.5	0.244	0.240	0–1
	Bigram-1b	0.467	0.182	0–1	0.068	0.093	0–0.5	0.268	0.246	0–1
	Bigram-1a	0.444	0.197	0–1	0.082	0.107	0–0.6	0.263	0.240	0–1
	Bigram-1b1a	0.482	0.173	0–1	0.083	0.101	0–0.5	0.282	0.245	0–1
	Trigram	0.265	0.228	0–1	0.009	0.042	0–0.4	0.137	0.208	0–1
	Trigram-1b	0.306	0.216	0–1	0.010	0.043	0–0.4	0.158	0.215	0–1
Orthographic	Trigram-1a	0.287	0.224	0–1	0.018	0.057	0–0.5	0.152	0.212	0–1
similarity	Trigram-2b	0.366	0.199	0–1	0.018	0.053	0–0.46	0.192	0.227	0–1
	Trigram-2a	0.325	0.218	0–1	0.038	0.080	0–0.6	0.181	0.218	0–1
	Trigram-1b1a	0.320	0.207	0–1	0.018	0.053	0–0.4	0.169	0.214	0–1
	Trigram-1b2a	0.350	0.198	0–1	0.036	0.073	0–0.5	0.193	0.217	0–1
	Trigram-2b1a	0.371	0.188	0–1	0.023	0.058	0–0.42	0.197	0.223	0–1
	Trigram-2b2a	0.393	0.179	0–1	0.039	0.073	0–0.43	0.216	0.224	0–1
	LCS	5.444	2.279	0–20	2.201	1.133	0–7	3.823	2.471	0–20
Orthographic	Edit Distance	4.075	1.960	0–19	8.788	2.983	2–27	6.432	3.453	0–27
distance	NED	0.446	0.158	0–1	0.837	0.112	0.4–1.0	0.641	0.239	0–1
Phonetic	Editex	10.807	5.406	0–55	23.673	8.242	4–76	17.240	9.485	0–76
distance	TED	104.197	90.009	0–1367	243.533	162.046	34–1527	173.865	148.420	0–1527
	Edit-Soundex	1.810	1.086	0–8	4.295	1.363	1–11	3.053	1.750	0–11
	Edit-Phonix	1.784	1.149	0–9	4.512	1.526	1–13	3.148	1.920	0–13
	Edit-Omission	3.098	1.399	0–8	5.992	1.592	2–12	4.545	2.060	0–12
	Edit-Skeleton	3.276	1.478	0–11	6.582	1.590	2–14	4.929	2.256	0–14

Note: Trigram-2b, trigram with two spaces added before the word; Trigram-1a, trigram with one space added after the word, etc.

*For every measure, mean similarity/distance scores differed significantly between cases and controls (t tests, $P < 0.0001$).

LCS, longest common subsequence; NED, normalized edit distance; TED, tapered edit distance; SD, standard deviation.

Transformed strings are then truncated to a length of four symbols. Thus, the **soundex** codes for clonidine and Clonopin (Roche, Nutley, NJ) would be c4535 and c4515, respectively. Once a pair of words had undergone phonetic transformation, raw edit distance was used to calculate the distance between them.[23,33] **Editex** is a variant on **soundex** which uses slightly different letter groups (Fig. 2).[36] With **editex**, edit distance was computed as usual, but the cost of a letter substitution depended on the letter groups. If two letters were the same, the cost was 0. If two letters were in the same **editex** letter group, the cost was 1. Otherwise, the cost of an insertion, deletion, or substitution was 2.

A **tapered edit distance**, which computed the cost of each edit as a function of the position in thestring (with higher costs given to edits near the start of the string), was the final measure in this category.[23]

Analysis Plan

Descriptive statistics were computed for all 22 measures. Differences in mean similarity scores were examined by the t test and the Mann-Whitney U test. Correlations between the mea-

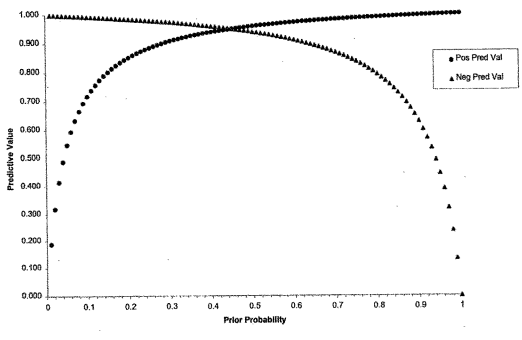

FIG. 3.

TABLE 2. Average Performance of Similarity and Distance Measures in Case-Control Analyses of Name-Confusion Errors (n = 1,127 cases, n = 1,127 controls) Based on Ten-fold Crossvalidation

Measure	Accuracy	95% CI	Sensitivity	95% CI	Specificity	95% CI	Cutoff	95% CI	OR	95% CI
Trigram2b	0.929	(0.916, 0.941)	0.920	(0.903, 0.937)	0.937	(0.918, 0.957)	0.116	(0.113, 0.119)	364.994	(12.735, 717.253)
Trigram2b1a	0.924	(0.911, 0.936)	0.936	(0.919, 0.952)	0.912	(0.892, 0.931)	0.110	(0.110, 0.110)	210.141	(118.140, 302.141)
NED	0.917	(0.897, 0.937)	0.901	(0.867, 0.935)	0.933	(0.910, 0.956)	0.659	(0.652, 0.666)	266.452	(82.427, 450.478)
Bigram1b	0.916	(0.904, 0.927)	0.886	(0.875, 0.896)	0.946	(0.926, 0.965)	0.246	(0.243, 0.249)	183.527	(115.520, 251.535)
Bigram1b1a	0.909	(0.894, 0.923)	0.905	(0.887, 0.924)	0.913	(0.888, 0.937)	0.245	(0.237, 0.253)	126.294	(91.815, 160.773)
Trigram2b2a	0.901	(0.883, 0.919)	0.885	(0.859, 0.911)	0.917	(0.894, 0.940)	0.168	(0.152, 0.184)	130.067	(63.210, 196.923)
Trigram1b1a	0.889	(0.871, 0.906)	0.860	(0.836, 0.883)	0.918	(0.898, 0.937)	0.110	(0.110, 0.110)	89.840	(55.448, 124.232)
Trigram1b2a	0.884	(0.864, 0.905)	0.903	(0.881, 0.924)	0.866	(0.842, 0.890)	0.110	(0.110, 0.110)	79.951	(46.294, 113.609)
Trigram1b	0.883	(0.866, 0.901)	0.805	(0.777, 0.834)	0.962	(0.948, 0.975)	0.116	(0.110, 0.122)	195.798	(71.728, 319.868)
Bigram1a	0.882	(0.863, 0.901)	0.857	(0.830, 0.884)	0.907	(0.885, 0.929)	0.240	(0.240, 0.240)	78.189	(48.728, 107.651)
Edit-Distance	0.882	(0.861, 0.902)	0.822	(0.790, 0.855)	0.941	(0.925, 0.957)	5.123	(5.109, 5.137)	98.397	(59.876, 136.917)
Editex	0.881	(0.861, 0.901)	0.839	(0.808, 0.871)	0.923	(0.904, 0.942)	15.002	(14.755, 15.249)	96.382	(37.314, 155.450)
Bigram	0.874	(0.856, 0.892)	0.813	(0.779, 0.847)	0.935	(0.921, 0.949)	4.062	(4.058, 4.066)	83.325	(44.497, 122.154)
Edit-Omission	0.866	(0.851, 0.882)	0.870	(0.841, 0.898)	0.863	(0.831, 0.894)	0.177	(0.160, 0.194)	50.827	(38.399, 63.254)
Edit-Phonix	0.862	(0.848, 0.876)	0.933	(0.923, 0.943)	0.791	(0.772, 0.811)	3.116	(3.108, 3.124)	59.184	(43.781, 74.587)
Edit-Soundex	0.859	(0.835, 0.883)	0.774	(0.733, 0.815)	0.945	(0.935, 0.954)	2.090	(2.090, 2.090)	75.305	(47.537, 103.074)
Edit-Skeleton	0.850	(0.835, 0.864)	0.850	(0.822, 0.878)	0.849	(0.831, 0.868)	4.080	(4.080, 4.080)	35.968	(28.112, 43.823)
Trigram2a	0.848	(0.829, 0.867)	0.840	(0.808, 0.873)	0.855	(0.835, 0.876)	0.116	(0.113, 0.119)	37.661	(26.012, 49.310)
Trigram1a	0.844	(0.828, 0.860)	0.773	(0.747, 0.800)	0.914	(0.894, 0.935)	0.115	(0.112, 0.118)	44.509	(31.753, 57.265)
LCS	0.837	(0.816, 0.859)	0.846	(0.827, 0.866)	0.829	(0.802, 0.855)	3.006	(2.994, 3.018)	31.153	(21.169, 41.138)
Trigram	0.836	(0.823, 0.850)	0.727	(0.707, 0.747)	0.946	(0.932, 0.959)	0.070	(0.050, 0.090)	59.919	(36.585, 83.252)
TED	0.812	(0.787, 0.836)	0.803	(0.777, 0.828)	0.821	(0.791, 0.851)	137.855	(137.022, 138.688)	22.275	(14.647, 29.902)

Note: On each fold of crossvalidation, a locally optimal cut-off point was chosen by evaluating 100 different cut-off points, evenly spaced across the range of the given measure and by picking the one that performed best in resubstitution accuracy. Ninety percent of the n = 2,254 pairs of cases and controls were used to select the cut-off point for each fold. Each cut-off point was then tested on the remaining 10% of the data; this process was repeated 10 times. The table reports mean results from ten trials. The ten test sets were nonoverlapping. Cut-off point is the mean of the 10 locally optimal cut-off points. Odds ratio is the mean of the odds ratios at ten locally optimal cut-off points. Trigram-2b, trigram with two spaces added before the word; Trigram-1a, trigram with one space added after the word, etc. Measures are sorted in decreasing order of accuracy.

OR, odds ratio; CI, confidence interval; NED, normalized edit distance.

sures were examined. We then estimated the sensitivity, specificity, overall accuracy, best cutoff value, and odds ratio for each individual measure using 10-fold crossvalidation.[37–39] On each fold of crossvalidation, 90% of the data was used as a training set. A locally optimal cutoff point was chosen by evaluating 100 different cutoff points evenly spaced across the range of the given measure and then selecting the point that most accurately discriminated between cases and controls in the training set. That cutoff point was then tested on the remaining 10% of the data. This process was repeated 10 times. The ten test sets were nonoverlapping.

One orthographic similarity measure, one orthographic distance measure, and one phonetic distance measure were then selected for further analysis. For each of these measures, we examined the relationship between similarity and the odds of confusion. Similarity scores were broken into six exposure levels, and the odds ratio was examined as a function of exposure level. The resulting relationships were tested for the presence of a significant trend.[31]

In constructing a multivariate model, we pursued a two-part strategy. Part I was based on the manual selection of a small set of individual variables. First, we built separate single-predictor logistic regression models using the most accurate orthographic similarity, orthographic distance, and phonetic similarity measures. We then formed a multiple logistic regression model of confusion probability using these same three measures as predictors.[40] The predictive accuracy of the model was evaluated by 10 fold crossvalidation. On each fold of crossvalidation, a logistic regression model was formed using 90% of the cases and controls. Each model was then tested on the remaining 10% of the data. The process was repeated 10 times using nonoverlapping test data. Positive and negative predictive values were plotted as a function of previous probability of error.[38]

Part II of our multivariate modeling strategy began with data reduction. First, we performed principal components analysis on the covariance matrix of the 22 measures. A small number of principal components was selected, and a multivariate model was formed using the component scores as predictors. Accuracy was again evaluated by 10 fold crossvalidation. On each fold of the crossvalidation, components were extracted, and a logistic regression model was formed on the resulting component scores.

Results

Table 1 displays descriptive statistics for cases and controls for each of the measures. Mean differences between cases and controls were large and statistically reliable according to the t test ($P < 0.000$). Concerns about the skewed distribution of similarity scores (especially for cases) also led us to examine differences with nonparametric tests. Differences remained significant in a Mann-Whitney U test and in a χ^2 goodness-of-fit test comparing 10-bin histograms ($P < 0.000$, details not shown). The 22 measures were highly interrelated. Correlations ranged in absolute value from 0.13 to 0.98, with most having an absolute value greater than 0.80. All correlations were significant at the .0001 level.

Table 2 shows means and 95% confidence intervals for sensitivity, specificity, overall accuracy, cutoff values, and odds ratios for each of the individual measures tested.

Dose-response data are in Table 3. For this analysis, we selected the most accurate measure of each type (by "most accurate," we mean those with the highest sample means, ignoring for the moment that several measures had overlapping 95% confidence intervals). **Trigram-2b** was the most accurate orthographic similarity measure. **Normalized edit distance** was the most accurate orthographic distance measure, and **editex** was the most accurate phonetic distance measure (Table 2). For each measure, we divided the scores into six ranges, computed the odds and odds ratios, and tested for the significance of a trend in odds ratios. The midpoint of each range was used as a cutoff point in these analyses.[31] As indicated in Table 3, there was an increasing trend in the odds ratios for each measure. Compared with names with similarity scores at the lowest levels (or greatest distances), pairs of names with the highest similarities (shortest distances) were thousands of times more likely to be errors.

Multivariate Model I

We formed three single-predictor logistic regression models of error probability using **trigram-2b**, **normalized edit distance**, and **editex**, respectively.[40] Model parameters are given in Table 4. Next, we formed a logistic regression model including all three predictors. The estimated parameters to that model are indicated in Table 5. Using resubstitution, the

TABLE 3. Relation of Similarity/Distance to Odds Ratio of Error for Selected Univariate Measures

Measure	Similarity or Distance	Errors	Controls	Odds	OR
	0	60	983	0.061	1.000
	0–0.1	9	46	0.196	3.213
Trigram-2b	0.1–0.2	196	79	2.481	40.672
	0.2–0.3	181	14	12.929	211.951
	0.3–0.4	227	2	113.500	1860.656
	>0.4	454	3	151.333	2480.869
		χ^2 for trend = 1522.84, $P < 0.000$			
	0.9–1.0	5	277	0.018	1.000
	0.8–0.9	15	429	0.035	1.944
NED	0.7–0.8	51	283	0.180	10.000
	0.6–0.7	86	94	0.915	50.833
	0.5–0.6	158	32	4.938	274.333
	≤0.5	812	12	67.667	3759.278
		χ^2 for trend = 1579.74, $P < 0.000$			
	>50	1	19	0.053	1.000
	40–50	2	45	0.044	0.830
Editex (Distance)	30–40	9	80	0.113	2.123
	20–30	38	560	0.068	1.280
	10–20	458	419	1.093	20.624
	0–10	619	4	154.750	2919.811
		χ^2 for trend = 970.09, $P < 0.000$			

Note: Results displayed in increasing order of similarity, decreasing order of distance. Normalized edit distance and editex are distance measures; odds increase as distance decreases. Trigram-2b, trigram with two spaces added before the word.

OR, odds ratio; NED, normalized edit distance.

model had 93.6% sensitivity, 96.1% specificity, and 94.9% overall accuracy. Table 6 shows the results of tenfold crossvalidation for the three-predictor model. The model had crossvalidated sensitivity of 93.7%, specificity of 95.9%, and accuracy of 94.8%. Performance of this three-variable model was statistically equivalent to that of a univariate model based on trigram with two spaces before the word, but better than all other univariate models. The positive and negative predictive values for this model are given in Fig. 3 for previous probabilities from 0 to 1. The relevant previous probabilities for name confusion errors are almost certainly below 0.1.

Multivariate Model II

Two potential problems confronted Part I of the modeling strategy. First, the 95% confidence intervals for the univariate measures overlapped and, therefore, it was not possible to uniquely identify the most accurate measures. Second, the measures were highly intercorrelated. We reasoned that data reduction by principal components analysis would identify the main dimensions of variation in our measures. Scores on these dimensions could then be used as predictors in another model. Principal components analysis yielded two components that combined to account for 89.74% of the variance in similarity scores (the details of these analyses are not shown but are available from the author upon request). Based on the component coefficients, one component tapped similarity and the other tapped distance. A logistic regression model was formed using scores from these two components as predictors. On the entire data set, the model had 93.70% sensitivity, 95.03% specificity, and 94.37% accuracy. By 10 fold

TABLE 4. Single-Predictor Logistic Regression Models for Predicting Probability of Name-Confusion Errors

Variable	Beta	SE(B)	Wald χ^2	−2 Log Likelihood	$P > \chi^2$	OR	95% CI for OR
Trigram-2b	22.7319	1.0159	500.7169	1000.864	0.0001	7.45×10^9	$(1.50 \times 10^8, 3.69 \times 10^{11})$
						9.71†	(7.956, 11.847)
NED	−17.3343	0.7407	547.7006	969.419	0.0001	2.96×10^{-8}	$(1.27 \times 10^{-7}, 6.97 \times 10^{-9})$
						0.177*	(0.153, 0.204)
Editex	−0.3734	0.0150	618.9278	1474.477	0.0001	0.688	(0.668, 0.709)
						0.059‡	(0.047, 0.073)

SE, standard error; OR, odds ratio; CI, confidence interval; NED, normalized edit distance.
*Indicates odds ratio for changes in NED of 0.1 units.
†Indicates odds ratio for changes in trigram-2b of 0.1 units.
‡Indicates odds ratio for changes in editex of 7.6 units ($1/10^{th}$ of the editex range).

crossvalidation, the model had mean sensitivity of 93.48% (SD = .03) and mean specificity of 94.96% (SD = .03), and mean accuracy of 94.13% (95% CI: 93.51%–94.74%). The differences between the multivariate model I, model II, and the best univariate models were within the range of sampling error.

Limitations

These results should be interpreted in light of the following limitations. First, cases were drawn from voluntary reports, and their validity is compromised by weaknesses inherent in voluntary reporting systems (eg, selection biases, under-reporting, and reporting of near misses vs. actual errors).[41] No concept of error frequency or severity was included. Moreover, this paper ignored much of the complexity of oral communication. When drug names are spoken out loud, differences in regional dialects, personal diction traits, job pressure, nonstandard pronunciations, and so forth can also influence error rates associated with nomenclature. Beyond the name, this study ignored other drug product characteristics that may contribute to confusion errors[42]; that is, the measures presented earlier predicted confusion between drug names. A drug product, however, is more than just a name. To the extent that other characteristics (ie, dose, dosage form, etc.) contribute to drug confusion errors, the current measures are incomplete.

TABLE 5. Three-Variable Logistic Regression Model for Predicting Probability of Name-Confusion Errors

Variable	Beta	SE(B)	Wald χ^2	$P > \chi^2$	OR	95% CI for OR
Trigram-2b	14.5019	1.1446	160.5381	0.0000	1.98×10^6	$(1.16 \times 10^5, 3.38 \times 10^7)$
					4.263†	(3.407, 5.336)†
NED	−6.1107	0.9714	39.5702	0.0001	0.002	(0.000, 0.015)
					0.543*	(0.449, 0.657)*
Editex	−.1698	0.0221	59.2658	0.0000	0.844	(0.808, 0.881)
					0.275‡	(0.198, 0.383)‡
Constant	5.2786	0.6051	76.1097	0.0000		
−2 log likelihood = 674.378, $\chi^2(3) = 2450.330$, $P < 0.0001$						

Note: Trigram-2b, trigram with two spaces added before each word.
SE, standard error; OR, odds ratio; CI, confidence interval; NED, normalized edit distance.
*Indicates odds ratio and confidence interval for changes in NED of 0.1 units.
†Indicates odds ratio and confidence interval for changes in trigram-2b of 0.1 units.
‡Indicates odds ratio for changes in editex of 7.6 units ($1/10^{th}$ of the editex range).

Discussion

Summary of Findings

Some pairs of drug names are more likely to be involved in errors than others.[25] This analysis demonstrated that automated measures of orthographic and phonetic similarities can be used to distinguish between known error pairs and controls drawn from the same population of names. Specifically, we have shown (1) that errors and controls are distributed differently with respect to measures of similarity and distance, (2) that individual measures of similarity or distance can form the basis of sensitive and specific prognostic tests of error potential, (3) that the odds of being involved in a name-confusion error increase as similarity increases (distance decreases), and (4) that a multivariate model is no more accurate than the best univariate model.

Improvements on Previous Work

This investigation sought to build upon an earlier study.[25] Previous research examined only three measures of orthographic similarity[20,21,23]; however, research in psycholinguistics indicated that phonological (ie, sound) similarity played a role in lexical confusions in addition to that played by orthographic similarity.[43–46] The present study evaluated 22 distinct similarity measures–some orthographic and some phonetic. The previous study assessed predictive accuracy by resubstitution, but resubstitution tends to overestimate the accuracy of a model.[37,39,47–49] The present study used 10 fold crossvalidation to reduce optimistic bias. Whereas the original study only examined univariate models, the present investigation incorporated both univariate and multivariate models. Finally, the earlier effort did not draw cases and controls from the same population of names and did not provide information about dose-response relationships. The present investigation addressed both of these problems.

Error Prevention Strategy and Policy

This work suggests strategies for error prevention. For example, one could provide a list of confusing name pairs to vendors of drug order entry systems. These pairs could be integrated into software warning systems so that when a potentially confusing name is entered by a doctor, nurse, or pharmacist, a warning would caution the professional to double check that the correct drug was being identified.[50] This approach would seem to be most powerful for dealing with confusing drug names that are already on the market. A more efficient strategy would be to use the predictors described earlier to screen proposed drug names against a standard database of existing drug names (a reference-standard database of names does not currently exist, but one could be developed through cooperative efforts of the International Nonproprietary Name, USAN, US Pharmacopeia, Inc., FDA, commercial trademark search firms, and the drug industry). The output of that screening process would feature a list of existing names, ranked in order of similarity to the proposed name.

One problem with the methods demonstrated in this paper is that they might not be sufficiently accurate to serve as the sole basis for judgments about confusion potential. For example, the multivariate model we tested had a sensitivity of 93.7% and specificity of 95.9%. This translates to a false negative rate of 6.3% and a false positive rate of 4.1%. Concretely, when screening a new name against a database of existing names, our best model will miss roughly 1 of every 16 truly confusing names; 1 of every 24 names our model identifies as confusing will prove not to be. These error rates may be acceptable when computer screening is one part of a multistep process that includes evaluation by human experts, but they seem too high to justify using the computerized methods as the sole basis for regulatory decisions. Moreover, the three-variable model had poor positive predictive value at the relevant prior probabilities, meaning that additional screening of names returned by our search methods would be required to maximize the practical usefulness of the model's predictions.

Efforts are underway to reduce these error rates, but research in information retrieval confirms that there will always be a tradeoff between recall (sensitivity) and precision (specificity).[51] Consequently, we propose that US Pharmacopeia, Inc., USAN, FDA, the US Patent and Trademark Office and other name-review bodies begin using objective measures, such as those developed earlier, to screen proposed drug names. The list of names retrieved by the initial screening would be supplied as input to an expert review process that would judge error potential by

examining practice factors and drug product characteristics other than the nomenclature. The European counterpart to the FDA, the European Agency for the Evaluation of Medicinal Products (EMEA), has already taken a step in this direction by issuing a draft guidance paper on trademarks that requires new names to differ from existing names by at least three letters.[53] This overly simplistic criterion would not receive our endorsement as it is currently expressed, but it does represent an useful attempt to apply objective standards to the name approval process.

Conclusion

Similarity increases the risk of drug-name confusion errors. Sensitive and specific tests of confusion potential can be formed using objective measures of similarity. Organizations that approve drug names should consider using such prognostic indicators to screen proposed names against databases of existing names. Experts, who would make the final determination about the acceptability of new drug names, could then review potentially confusing names in light of other factors that may affect the "real world" potential for confusion. The quality of the drug-name approval process and the look-alike/sound-alike error rate ought to improve as a result.

Acknowledgments

The authors acknowledge the helpful assistance of Dan Boring, Bill Brewer, Patricia Byrns, Mike Cohen, and the staff of the Institute for Safe Medication Practices, Gary Dell, Prahlad Gupta, Keith Johnson, David Lambert, Robert Lee, Susan Proulx, Don Rucker, Gordon Schiff, and Justin Zobel.

References

1. **Edgar TA, Lee DS, Cousins DD.** Experience with a national medication error reporting program. Am J Hosp Pharm 1994;51:1335–1338.

2. **Davis NM.** Drug names that look and sound alike. Hosp Pharm 1997;32:1558–1570.

3. US Pharmacopeia. USP Quality Review. Stop, look, and listen. Rockville, MD: US Pharmacopeia, 1995.

4. **Brodell RT, Helms SE, KrishnaRao I, Bredle DL.** Prescription errors. Arch Fam Med 1997;6:296–298.

5. US Pharmacopeia. Similar names. USP DI Update. Vol. I and II. 1997:4–7.

6. US Pharmacopeia. Drug product quality review. What's in a name, nombre, nom, naam? Rockville, MD: US Pharmacopeia, 1993:1–2.

7. **Boring D, Homonnay-Weikel AM, Cohen M, Di Domizio G.** Avoiding trademark trouble at FDA. Pharm Exec 1996;16:80–88.

8. **Boring D.** The development and adoption of nonproprietary, established, and proprietary names for pharmaceuticals. Drug Inf J 1997;31:621–634.

9. **Freimanis R.** The naming of drugs: The role of the United States Adopted Names (USAN) Council. DE Monitor 1994;Summer.

10. **Boring DL.** The CDER labeling and nomenclature committee: Structure, function, and process. Drug Inf J 1997;31:7–11.

11. **Boring DL, Stein IA, Di Domizio G.** United States: Trademark trainwrecks at FDA. Trademark World 1998;July:29–33.

12. US Pharmacopeia. Appendix I: Guiding principles for coining United States adopted names for drugs. USP Dictionary of USAN and International Drug Names. Rockville, MD: US Pharmacopeia, 1998:867–877.

13. **Kane SD.** Trademark law: A practitioner's guide. 3rd. New York: Practicing Law Institute, 1997.

14. **Gundersen GA.** Trademark searching: A practical and strategic guide to the clearance of new marks in the US. New York: International Trademark Association, 1994.

15. Thomson and Thomson Home Page. Trademark and copyright services. 1999 [cited 1999 March 1]; Available from: URL: http://www.thomson-thomson.com/.

16. US Pharmacopeia. USP dictionary of USAN and international drug names. Rockville, MD: US Pharmacopeia, 1998.

17. **Lans MS.** Trademark searches are an ounce of prevention. Marketing News 1994;28:13.

18. **Ojala M.** Trademarks for the business searcher. Online 1996;20:52–54.

19. Avantiq. Avantiq home page. 1997 [cited December 21, 1998]; Available from: URL: http://www.avantiq.lu/.

20. **Stephen GA.** String searching algorithms. River Edge, NJ: World Scientific, 1994.

21. **Aoe J.** Computer algorithms: String pattern matching strategies. Washington, DC: IEEE Computer Society Press, 1994.

22. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** Basic local alignment search tool. J Mol Biol 1990;215:403–410.

23. **Zobel J, Dart P.** Phonetic string matching: Lessons from information retrieval. In: Frei HP, Harman D, Schauble P, Wilkinson R, eds. SIGIR96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich, Switzerland. New York: Association for Computing Machinery, 1996:166–72.

24. Language Analysis Systems. Language Analysis Systems Home Page. 1998 [cited 1998 December 21]; Available from: URL: http://www.las-inc.com.

25. **Lambert B.** Predicting look- and sound-alike medication errors. Am J Health-Syst Pharm 1997;54:1161–1171.

26. **Grabenstein JD, Proulx SM, Cohen MR.** Recognizing and preventing errors involving immunologic drugs. Hosp Pharm 1996;31:791–804.

27. **Davis NM, Cohen MR, Teplitsky B.** Look-alike and sound-alike drug names: The problem and the solution. Hosp Pharm 1992;27:95–110.

28. US Food and Drug Administration. Medication errors. FDA Med Bull 1996;26:3.

29. **Dell GS, Reich PA.** Stages in sentence production: An analysis of speech error data. J Verbal Learning Verbal Behav 1981;20:611–629.

30. **Cohen J.** Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum, 1988.

31. **Schlesselman JJ.** Case control studies. New York: Oxford University Press, 1982.

32. **Frakes WB.** Stemming algorithms. In: Frakes WB, Baeza-Yates R, eds. Information retrieval: Data structures and algorithms. Englewood Cliffs, NJ: Prentice-Hall, 1992:131–160.

33. **Pfeifer U, Poersch T, Fuhr N.** Searching for proper names in databases. In: Kuhlen R, Rittberger M, eds. Proceeding of the Conference on Hypertext–Information Retrieval-Multimedia: Synergistic Effects of Electronic Information Systems. Konstanz, Germany: University of Dortmund, 1995:103–109.

34. **Gadd TN.** 'Fisching fore werds': Phonetic retrieval of written text in information systems. Program 1988;22:222–237.

35. **Gadd TN.** Phonix: The algorithm. Program 1990;24:363–366.

36. **Zobel J, Dart P.** Finding approximate matches in large lexicons. Software-Practice and Experience 1995;25:331–345.

37. **Kohavi R.** A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Mellish CS, editor. Proceedings of the International Joint Conference on Artificial Intelligence. Montreal: Morgan Kaufmann, 1995:1137–1143.

38. **Hulley SB, Cummings SR.** Designing clinical research. Baltimore, MD: Williams and Wilkins, 1988.

39. **Harrell FE, Lee KL, Mark DB.** Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–387.

40. **Hosmer DW, Lemeshow S.** Applied logistic regression. New York: John Wiley & Sons, 1989.

41. **Cullen DJ, Bates DW, Small SD, Cooper JB, Nemeskal AR, Leape LL.** The incident reporting system does not detect adverse drug events: A problem for quality improvement. Jt Comm J Qual Improv 1995;21:541–548.

42. **Cohen MR.** Drug product characteristics that foster drug-use-system errors. Am J Health-Syst Pharm 1995;52:395–399.

43. **Baddeley AD.** Semantic and acoustic similarity in short-term memory. Nature 1964;204:1116–1117.

44. **Baddeley AD.** Short-term memory for word sequences as a function of acoustic, semantic, and formal similarity. Q J Exp Psychol 1966;18:362–365.

45. **Baddeley AD.** Working memory. Oxford: Oxford University Press, 1986.

46. **Baddeley AD.** How does acoustic similarity influence short-term memory? Q J Exp Psychol 1968;20:249–264.

47. **Johnson RA, Wichern DW.** Applied multivariate statistical analysis. Englewood Cliffs, NJ: Prentice-Hall, 1988.

48. **Laupacis A, Sekar N, Stiell IG.** Clinical prediction rules: A review and suggested modifications of methodological standards. JAMA 1997;277:488–494.

49. **Wasson JH, Sox HC, Neff RK, Goldman L.** Clinical prediction rules: Application and methodological standards. N Engl J Med 1985;313:793–799.

50. **Cohen M.** Novel way to prevent medication errors. 1996 [cited 1998 March 1]; Available from: URL: http://www.ismp.org/ISMP/alert_toc.html July 31,1996.

51. **Salton G, McGill M.** Introduction to modern information retrieval. New York: McGraw-Hill, 1983.

52. **Yu C, Meng W.** Principles of query processing for advanced database applications. San Francisco, CA: Morgan Kaufmann, 1998.

53. Committee for Proprietary Medicinal Products (CPMP). Draft guidance paper on the acceptability of tradenames for medicinal products processed through the centralized procedure. 1998 [cited 1998 August 19]; Available from: URL: http://www.eudra.org/frame/frametest3.html.