ARTICLE

Predicting look-alike and sound-alike medication errors

BRUCE L. LAMBERT

Abstract: A model for predicting medication name confusion is described.

Many medication errors are caused by look-alike and sound-alike medication names, yet few procedures exist to ensure the safety of new drug nomenclature or to identify confusingly similar names from within existing databases. In this study, three automated, quantitative measures of orthographic similarity (i.e., similarity in spelling) were identified (bigram similarity, trigram similarity, and Levenshtein distance). The relationship between orthographic similarity and the likelihood of a medication error was examined. For each measure of similarity, the frequency distribution of similarity scores for pairs of drug names previously reported to cause confusion (error pairs) was compared with the distribution of similarity scores for control pairs randomly selected from the general index of *USP DI—Volume I: Drug Information for the Health Care Professional*. Then, three parallel, unmatched case–control studies were conducted to discover whether similarity was a significant risk factor for medication errors. Finally, on the basis of the three similarity measures, tests for predicting confusion were developed and evaluated.

For each similarity measure, the frequency distribution of error pairs was significantly different from that for control pairs, and orthographic similarity was a significant risk factor for medication errors. Pairs of names whose measures of similarity exceeded preset thresholds were between 25 and 523 times more likely to be involved in a medication error than pairs whose similarity did not exceed these thresholds. A prognostic test that correctly identified 91% of all pairs as either errors or controls was developed. This test had a sensitivity of 84% and a specificity of 99%.

Automated measures of similarities between medication names can form the basis of highly accurate, sensitive, and specific tests of the potential for errors with look-alike and sound-alike medication names.

Index terms: Computers; Errors, medication; Methodology; Models; Nomenclature

Am J Health-Syst Pharm. 1997; 54:1161-71

Medication errors present a serious threat to patient welfare and a significant liability to health professionals and their insurers.[1] Although there are many contributing factors, confusing pairs of medication names are consistently associated with errors. Look-alike and sound-alike medication names play a part in perhaps one quarter of all medication errors.[2,3] The agencies responsible for approving trademarks and established names (i.e., nonproprietary generic names) for new drug products, primarily the U.S. Food and Drug Administration (FDA) and the United States Adopted Names Council (USAN), lack valid and reliable methods for assessing the likelihood of errors from look- and sound-alike medication names.[4] Such methods could perhaps be used to reduce the number of confusing drug names that reach the marketplace and to identify confusing pairs within existing databases of medication names. Once these pairs were identified, safeguards could be built into drug information systems to reduce the probability of confusion in clinical practice.[5]

Until now, health care practitioners have had to rely

BRUCE L. LAMBERT, PH.D., is Assistant Professor, Department of Pharmacy Administration, and Clinical Assistant Professor, Department of Pharmacy Practice, University of Illinois at Chicago, 833 S. Wood Street (M/C 871), Chicago, IL 60612-7231 (lambertb@uic.edu).

The assistance of Dan Boring, Bill Brewer, Diane Cousins, Stephanie Crawford, Gary Dell, George DiDomizio, Jean Gallagher, Paul Grussing, Prahlad Gupta, Keith Johnson, Kim Keller, David Lambert, Don Rucker, Gordon Schiff, and Donna Szymanski is acknowledged. The United States Pharmacopeia is acknowledged for providing the author with an electronic version of the general index to the 1995 *USP DI—Volume I: Drug Information for the Health Care Professional*. Bruno Haible and Michael Stoll are acknowledged for placing their Common Lisp compiler (CLISP) in the public domain.

on voluntary reports of errors in identifying potentially confusing pairs of medication names.[2,3,6,7] The only available techniques for assessing the confusion potential of new trademark and established names have involved panels of experts completing a variety of rating scales.[6] The reliability and validity of these instruments have not been firmly established. In addition, the sheer number of existing medication names—more than 15,000 in the United States alone—makes it unlikely that a manual evaluation of the potential for confusion would ever be comprehensive enough to be viewed as reliable.[8,9] Roughly 15,000 comparisons would need to be made to assess the confusion potential of a single new name. To identify confusing pairs from among existing names, the number of unique pairs of names that would need to be considered would be $(N^2 - N)/2$. With $N = 15{,}000$, a staggering 112,492,500 comparisons would be needed. Clearly, such a task is impossible without automated methods for evaluating the potential for confusion.

Fortunately, it is now possible to design objective, computer-based measures of orthographic (spelling) and phonological (sound) similarity. These similarity measures can serve as the basis for predictions about the potential for confusion. Although lacking some of the features of manual evaluation by experts (e.g., consideration of dosage, indication, or physical appearance of the drug), the computerized measures of lexical similarity (i.e., similarity of words) are objective, reliable, and based on well-established psycholinguistic theory. Automated methods for searching trademark databases for similarities are already available from commercial vendors and are used routinely by intellectual property attorneys and other interested parties.[10,11] The automated methods allow exhaustive comparisons between proposed new medication names and existing names in databases. However, these methods need to be validated before they can be used in a regulatory or error-prevention context.

This report describes a series of validation experiments designed to assess and optimize the error-predicting potential of computerized measures of orthographic similarity. The result of these experiments was a prognostic test that can identify confusingly similar pairs among new and existing medication names.

Background

To date, the literature on errors resulting from look- and sound-alike medication names has been largely atheoretical. However, a vast literature in psycholinguistics and experimental psychology describes the mental and cognitive processes that produce lexical confusion in speaking, listening, reading, and writing, as well as in short-term and long-term memory.[12-24] These psychological experiments have been conducted primarily on college undergraduates, with common English words used as stimuli. A similar pattern of results might be expected if the participants were health professionals and the stimulus words were medication names.

The heart of all lexical processing is the mental lexicon, or "mental dictionary," where information about words is stored. Words are indexed in the mental lexicon by their orthographic, phonological, syntactic (grammatical), and semantic (meaning-related) representations.[25] These representations are directly involved in the cognitive processes that allow healthy adults to retrieve the correct words from memory when speaking, listening, writing, and reading. In general, words with similar orthographic, phonological, syntactic, or semantic representations are more likely than others to be confused. This general pattern has been observed across a wide range of experimental paradigms. For example, when people make errors in recall from immediate memory, they tend to recall words that sound similar to their target word.[13,26] Errors in recognition are more likely when distractor items are semantically similar to the target item.[27] Words with many orthographically or phonologically similar "neighbors" take longer to recognize and are more likely to be incorrectly recognized than words with few such neighbors.[21,23] Speech errors involving the substitution of one word for another (e.g., saying *pollution* instead of *population*) are more likely to involve semantically or phonologically similar words.[24,28] Thus, evidence from psycholinguistics clearly demonstrates that lexical similarity increases the likelihood of errors in recall, in recognition of written and spoken words, and in spontaneous speech.

Purpose

This study was designed to identify relationships between lexical similarity and lexical confusion in drug names. We sought answers to three specific questions:

1. What is the relationship between orthographic similarity and the likelihood of lexical confusion?
2. Is lexical similarity a significant risk factor for lexical confusion?
3. What measure of orthographic similarity will most accurately predict lexical confusion?

The central goal was to discover whether automated measures of lexical similarity could serve as valid predictors of medication error potential and thus be used to facilitate the development of safer, more error-resistant drug nomenclature.

The ability of three measures of orthographic similarity to distinguish between confusing and nonconfusing pairs of medication names was evaluated. Three hypotheses were tested. In the hypotheses, the term *error pair* refers to a pair of drug names that has been reported in the literature to cause confusion, and *control pairs* refers to pairs of names selected at random from a large database.

H1: The frequency distribution of orthographic similarity scores for known error pairs differs from that observed for control pairs; known error pairs are more similar, on average, than control pairs.

H2: Using a threshold (i.e., a cutoff value) to define similarity as a dichotomous exposure variable shows that orthographic similarity is a significant risk factor for look- and sound-alike medication errors.

H3: When receiver operator characteristic (ROC) curves are plotted, orthographic similarity can be used to construct a prognostic test that has at least 80 ± 5% sensitivity and 90 ± 5% specificity to predict lexical confusion.

Methods

Research design. Two related research designs were used to test the stated hypotheses. Each was both observational and retrospective. To compare the frequency distribution of error pairs and control pairs and to examine the association between orthographic similarity and the probability of lexical confusion, a case–control design was employed. Cases (i.e., error pairs) were drawn from published reports of medication errors, and controls were drawn at random from all possible pairs of medication names.[29] To determine how useful lexical similarity measures are in predicting whether or not a pair of names would be confused in clinical practice, a modified case–control design, appropriate for the evaluation of new prognostic tests, was used.[29-32]

Source of medication names. *Error pairs (cases).* Pairs of medication names either known to have been confused in clinical practice or judged by experts to be confusingly similar were compiled from several published lists.[3,6,33,34] These lists were combined and, after duplicate pairs were deleted, 969 unique error pairs were identified. Table 1 shows 20 of the error pairs used in the study.

Control pairs. Control pairs of medication names (969) were selected at random from an electronic version of the general index to the United States Pharmacopeia's *USP DI—Volume I: Drug Information for the Health Care Professional*.[8] Table 2 shows 20 of the control pairs used in this study.

Measurement of orthographic string similarity. Most computer programs view words as sequences, or strings, of letters. In the experiments reported here, orthographic string similarity was measured by three different methods: bigram, trigram, and Levenshtein distance.[35]

Bigram and trigram methods are examples of *n*-gram measures of string similarity. For a given pair of medication names, unique *n*-grams (i.e., *n*-letter subsequences) in each name were generated. Then, the number of *n*-grams common to the two names was tallied. Finally, *n*-gram string similarity (*S*) was defined by the Dice coefficient:

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

where *A* is the number of unique *n*-grams in the first word, *B* the number of unique *n*-grams in the second word, and *C* the number of unique *n*-grams common to the two words.[36,37] Both bigram (i.e., two-letter subsequence) and trigram (i.e., three-letter subsequence) measures were used in these investigations. For example, the name Tylenol has the unique trigrams *tyl*, *yle*,

Table 1.
Similarity and Distance for 20 Error Pairs[a]

Name 1	Name 2	Levenshtein Distance	Bigram Similarity	Trigram Similarity
Dimetane	Dimetapp	2	0.71	0.67
Atarax	Marax	2	0.67	0.57
Citracal	Citrucel	2	0.43	0.33
Diphenatol	Diphenidol	2	0.67	0.50
Cytotec	Cytoxan	3	0.50	0.40
Oracin	Orasone	3	0.36	0.22
Docusate	Doxinate	3	0.43	0.17
Enflurane	Isoflurane	3	0.71	0.67
Lopid	Slo-bid	3	0.40	0.00
Percocet	Percodan	3	0.57	0.50
Voltaren	Vontrol	4	0.31	0.00
Pralidoxime	Pyridoxine	4	0.42	0.35
Accubron	Accutane	4	0.43	0.33
Phos-Flur	Phoslo	4	0.46	0.36
Catapres	Combipres	4	0.40	0.31
Auralgan	Ophthalgan	5	0.50	0.43
Imferon	Intropin	5	0.17	0.00
Dipyridamole	Disopyramide	6	0.45	0.10
Chlor-Trimeton	Chloromycetin	6	0.40	0.26
Actacel	Actimmune	6	0.31	0.17

[a]This subset was randomly selected from the full list of 969 error pairs. Pairs are sorted in increasing order of Levenshtein distance. Although capital letters are used here, comparison of the names was not case sensitive. A complete list of the error pairs used can be obtained from the author.

Table 2.
Similarity and Distance for 20 Control Pairs[a]

Name 1	Name 2	Levenshtein Distance	Bigram Similarity	Trigram Similarity
Alaxin	Beta-HC	6	0.00	0.00
Dalgan	Atasol-8	7	0.00	0.00
6-MP	Canesten	8	0.00	0.00
Nitropress	Anectine	9	0.00	0.00
Diphen Cough	Citrolith	10	0.00	0.00
K-Phos Neutral	Lipisorb	11	0.00	0.00
Myciguent	Velosulin Human	13	0.00	0.00
Metaprel	Amrinone Lactate	13	0.09	0.00
Promote with Fiber	Diarrest	16	0.00	0.00
Anti-thymocyte Serum	Infumorph	16	0.15	0.00
Potassium Gluconate	Phentermine Hydrochloride	18	0.05	0.00
Sodium Dichloroacetate	Klerist-D	19	0.00	0.00
Aminophylline	Technetium Tc 99m HSA	20	0.07	0.00
Immun-Aid	Ascomp with Codeine No. 3	22	0.00	0.00
Preemie SMA 20	Minocycline Hydrochloride	22	0.11	0.00
Novo-Tamoxifen	Carbinoxamine Compound-Drops	24	0.15	0.00
Noxzema Anti-Acne Pads	Regular Strength Allerest Children	28	0.04	0.00
Apo-Metoclop	Contac Jr. Children's Cold Medicine	30	0.05	0.00
Poliovirus Vaccine Inactivated Enhanced Potency	Tarpaste	42	0.05	0.00
Propoxyphene Hydrochloride, Aspirin, and Caffeine	Slow Fe	44	0.08	0.00

[a]This subset was randomly selected from the full list of 969 control pairs. Pairs are sorted in increasing order of Levenshtein distance. Although capital letters are used here, comparison of the names was not case sensitive. A complete list of the control pairs used can be obtained from the author.

len, *eno*, and *nol*. The name atenolol has the unique trigrams *ate*, *ten*, *eno*, *nol*, *olo*, and *lol*. The trigram string similarity between Tylenol and atenolol, which share two trigrams in common (*eno* and *nol*), is $(2 \times 2)/(5 + 6) = 0.364$.

Levenshtein distance is a measure of orthographic string similarity that forms the basis for several widely used spell-checking and text-processing utilities.[35] It is the number of edit operations (e.g., substitutions, insertions, or deletions) needed to transform one word into another. The specific algorithm used to implement Levenshtein distance in these investigations was designed by Wagner and Fischer.[35,38] Consider the names Zantac and Xanax. In order to transform the word Zantac into the word Xanax, one must change the *Z* to an *X*, delete the *t*, and change the *c* to an *x*. Three edit operations are required; thus, the Levenshtein distance between the two names is 3.

All measures of string similarity were computed with programs that I wrote in the programming language Lisp. All comparisons were case insensitive; in effect, all names were converted to a single case (either upper or lower) before similarity measures were taken.

Analysis plan. If lexical similarity were to distinguish between error pairs and control pairs, the frequency distribution of similarities for error pairs should, at least, be significantly different from the frequency distribution of similarities for control pairs (i.e., hypothesis 1 should hold). To test this hypothesis, orthographic similarity was calculated for 969 error pairs and for 969 control pairs, and a chi-square test of independence was performed on the resulting 2×11 contingency table (error or control for 11 similarity ranges). With 1938 pairs overall, the chi-square contingency test at $\alpha = 0.01$ had greater than 99% power to detect effect sizes larger than $w = 0.20$.[39]

To establish whether lexical similarity represented a significant risk factor for the occurrence of a look- or sound-alike medication error (i.e., to test hypothesis 2), an unmatched case–control study was conducted with 969 error pairs and 969 controls.[40] Relative risk was estimated, with the odds ratio computed from a 2×2 contingency table (e.g., exposure or no exposure for case or control). Significance of the odds ratio was tested by using the chi-square statistic with 1 degree of freedom.[40] For *n*-gram methods, exposure was defined as similarity ≥0.10 by the Dice coefficient. For Levenshtein distance, exposure was defined as distance ≤10 edit operations. Assuming an exposure rate of 1% among controls, $\alpha = 0.01$, and 969 pairs for both

error pairs and controls, chi-square tests had 90% power to detect a relative risk greater than 4.[40]

The third phase of the study attempted to construct a prognostic test that would use automated measures of lexical similarity to predict which pairs of names were error pairs. ROC curves were plotted for each measure of lexical similarity previously described. The curves were plotted by systematically varying the threshold of the similarity score that corresponded to a positive prognostic test. Error pairs whose similarity scores exceeded the threshold were termed true positives. Control pairs that exceeded the threshold were termed false positives. Error pairs that did not exceed the threshold were termed false negatives. Control pairs that did not exceed the threshold were termed true negatives. Predictive accuracy, i.e., (number of true positives + number of true negatives)/(number of true positives + number of true negatives + number of false positives + number of false negatives), was measured and plotted at various thresholds. At each threshold, sensitivity, i.e., number of true positives/(number of true positives + number of false negatives), was plotted against 1 – specificity, where specificity was defined as number of true negatives/(number of true negatives + number of false positives).[32] The resulting ROC curves were used to select the optimal cutoff value for each test. With 969 error pairs and 969 control pairs, it was possible to estimate 99% confidence intervals for sensitivity and specificity of ± 5%.[29]

Positive predictive value, i.e., the probability that a pair was an error pair, given a positive test, was computed by the following formula[29]:

$$\frac{\text{Sensitivity} \times \text{Prior probability}}{(\text{Sensitivity} \times \text{Prior probability}) + [(1 - \text{Specificity}) \times (1 - \text{Prior probability})]}$$

Negative predictive value, i.e., the probability that a pair was a control pair, given a negative test, was computed by the following formula[29]:

$$\frac{\text{Specificity} \times (1 - \text{Prior probability})}{[\text{Specificity} \times (1 - \text{Prior probability})] + [(1 - \text{Sensitivity}) \times (\text{Prior probability})]}$$

The predictive value of a positive and a negative test was reported for various prior probabilities at the optimal threshold for each measure of similarity.[29,32] That is, once the optimal threshold was identified for each measure of orthographic similarity, a test based on that threshold was used to generate positive and negative predictive values over a range of prior probabilities.

Results

Frequency distributions. The frequency distributions of similarity scores for error pairs and control pairs, for each measure of similarity (bigram, trigram, and Levenshtein distance), are given in Figures 1–3. By definition, the Dice coefficient for bigram and trigram scores took on values from 0 to 1. Bigram and trigram scores were divided into 11 intervals. For bigram and trigram measures (Figures 1 and 2), the similarity scores for error pairs were skewed to the high end of the scale and the scores for control pairs were skewed to the low end. For the Levenshtein distance measure (Figure 3), scores for error pairs were skewed to the low end of the scale, i.e., few edit operations, while scores for control pairs were skewed to the high end. In each case, the chi-square test of independence was highly significant: For bigram string similarity, $\chi^2 = 1281.08$, d.f. = 10, $p <$

Figure 1. Histogram of bigram string similarities for 969 error pairs and 969 control pairs. Vertical axis is on a logarithmic scale. Light bars represent error pairs. Dark bars represent control pairs. Values at ends of vertical bars are frequencies. Values on the horizontal axis represent the bins for the histogram. For example, (0, 0.1) means "greater than 0 and less than 0.1," and [0.1, 0.2) means "greater than or equal to 0.1 and less than 0.2."

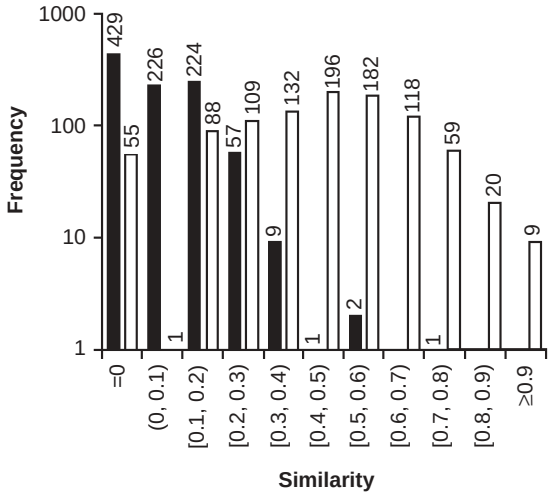


Figure 2. Histogram of trigram string similarities for 969 error pairs and 969 control pairs. Vertical axis is on a logarithmic scale. Light bars represent error pairs. Dark bars represent control pairs. Values at ends of vertical bars are frequencies. Values on the horizontal axis represent the bins for the histogram. For example, (0, 0.1) means "greater than 0 and less than 0.1," and [0.1, 0.2) means "greater than or equal to 0.1 and less than 0.2."

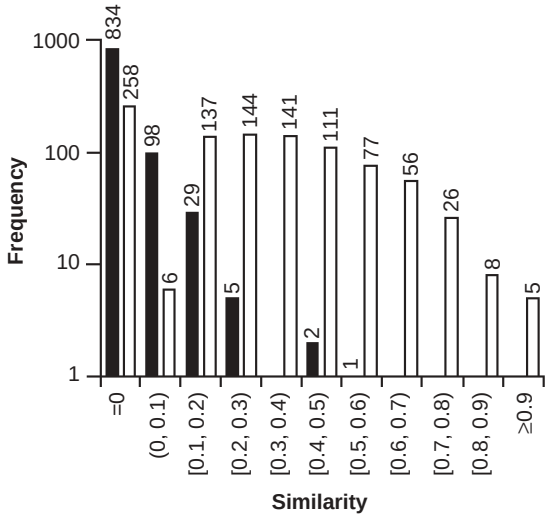


Figure 3. Histogram of Levenshtein distances for 969 error pairs and 969 control pairs. Vertical axis is on a logarithmic scale. Light bars represent error pairs. Dark bars represent control pairs. Values at ends of vertical bars are frequencies. Values on the horizontal axis represent the bins for the histogram. For example, [2, 4) means "greater than or equal to 2 and less than 4." Note that Levenshtein distance is not a similarity measure, so error pairs are skewed to the low end, and controls to the high end, of the distance scale.

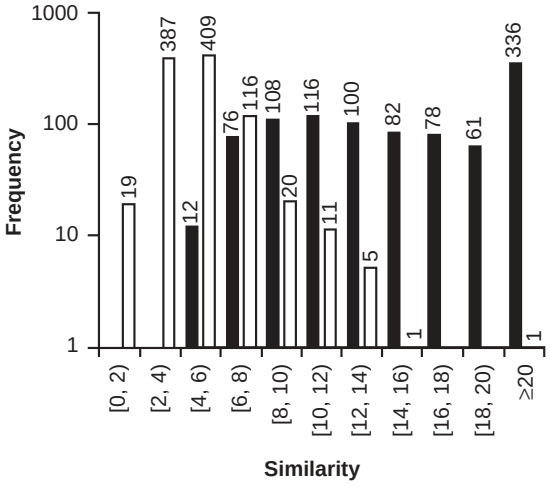


Table 3.
Frequency of Exposure for an Unmatched Sample of 969 Error Pairs and 969 Control Pairs

Measure of Exposure	Error	Control	Total
Bigram similarity ≥0.1			
Yes	913	314	1227
No	56	655	711
Trigram similarity ≥0.1			
Yes	705	37	742
No	264	932	1196
Levenshtein distance ≤10			
Yes	959	250	1227
No	10	719	711

Figure 4. Predictive accuracy of test based on bigram similarity at several thresholds. Accuracy values appear above each plotted point.

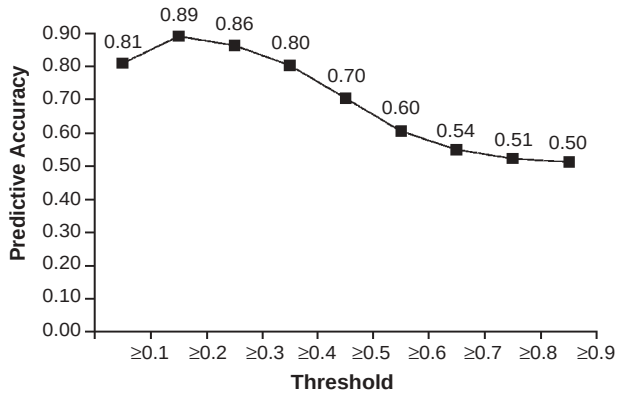
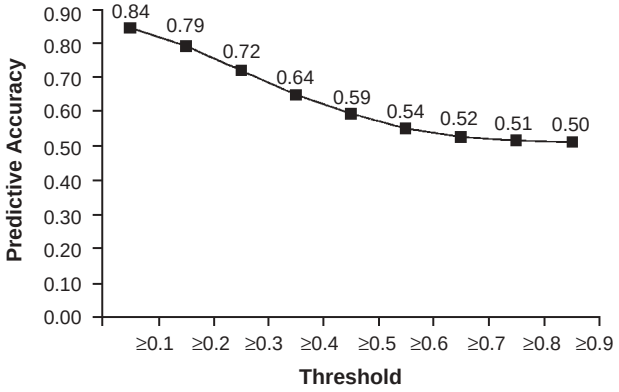


Figure 5. Predictive accuracy of test based on trigram similarity at several thresholds. Accuracy values appear above each plotted point.



0.00000; for trigram string similarity, $\chi^2 = 1000.34$, d.f. = 10, $p < 0.00000$; for Levenshtein distance, $\chi^2 = 1573.02$, d.f. = 10, $p < 0.00000$.

Relative risk. For bigram and trigram string similarity measures, exposure was defined as similarity ≥0.1. For Levenshtein distance, exposure was defined as distance ≤10. Although these thresholds were chosen somewhat arbitrarily, they were intended to be conservative values (i.e., relatively low similarity and relatively great distance). For each measure of exposure (Table 3), relative risk was approximated by the odds ratio.[40] For bigram string similarity, the odds ratio (OR) was substantially greater than 1 (OR = 34.01, 95% confidence interval [CI] = [25.16, 45.98], $\chi^2 = 797.06$, d.f. = 1, $p < 0.00000$), as it was for trigram similarity (OR = 67.27, 95% CI = [47.04, 96.19], $\chi^2 = 974.48$, d.f. = 1, $p < 0.00000$) and Levenshtein distance (OR = 275.81, 95% CI = [145.52, 522.76], $\chi^2 = 1105.33$, d.f. = 1, $p < 0.00000$). String similarity, regardless of how measured, was a significant risk factor for being involved in a look- or sound-alike medication error. Pairs of names whose similarity exceeded the specified threshold were between 25 and 523 times more likely to be involved in a medication error than those whose similarity did not exceed the threshold.

Predictive accuracy. The predictive accuracy of each measure at each threshold is given in Figures 4–6. The test with the highest overall accuracy (94%) was a Levenshtein distance-based test with a threshold of distance ≤6 (Figure 6). The bigram test with the highest accuracy classified 89% of the pairs correctly and was based on a threshold of similarity ≥0.2 (Figure 4). The trigram test with the highest accuracy classified 84% of the pairs correctly and was based on a threshold of similarity ≥0.1 (Figure 5).

Sensitivity and specificity: ROC curves. Figures 7–9 show the ROC curves for bigram, trigram, and Levenshtein distance, respectively. ROC curves display the tradeoff between sensitivity and specificity and are used to select the best threshold for a given diagnostic

Figure 6. Predictive accuracy of test based on Levenshtein distance at several thresholds. Accuracy values appear above each plotted point.

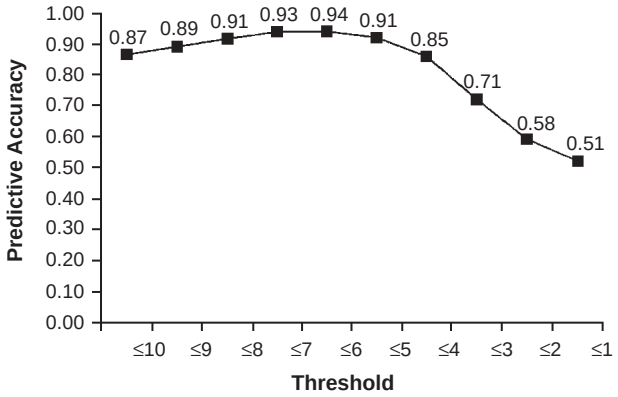


Figure 7. Receiver operator characteristic (ROC) curve for bigram string similarity in the prediction of look-alike and sound-alike medication errors. The bigram similarity values at various thresholds are in parentheses.

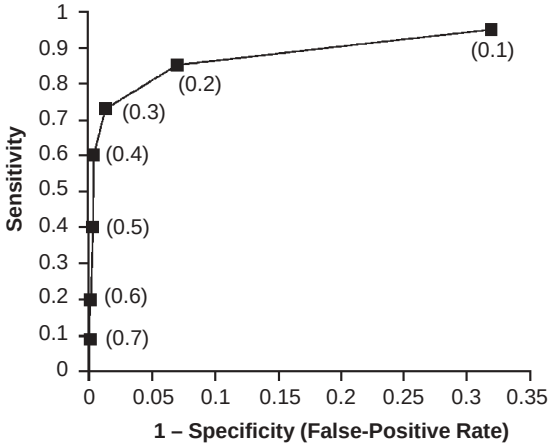
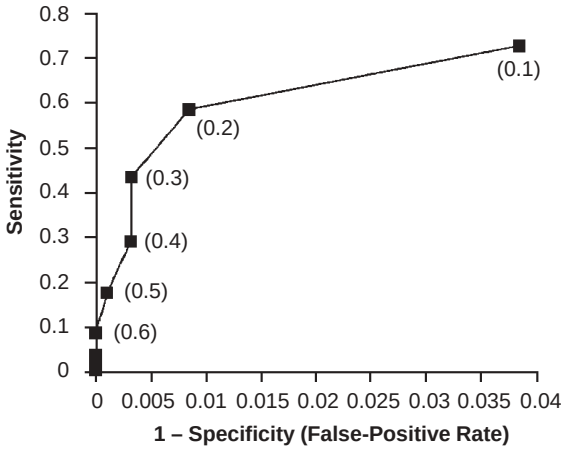


Figure 8. Receiver operator characteristic (ROC) curve for trigram string similarity in the prediction of look-alike and sound-alike medication errors. The trigram similarity values at various thresholds are in parentheses.



or prognostic test. The best point is at the shoulder of the ROC curve, i.e., the point of diminishing returns where further increases in sensitivity are offset by decreases in specificity.[29,32] For the bigram measure, the best threshold was similarity ≥0.3, where the test achieved 73% sensitivity and 98.6% specificity (Figure 7). For the trigram measure, the best threshold was similarity ≥0.2, where the test achieved 58.6% sensitivity and 99% specificity (Figure 8). For the Levenshtein distance measure, the best threshold was distance ≤5, where the test achieved 84% sensitivity and 98.8% specificity (Figure 9).

Positive and negative predictive value. Estimates of positive and negative predictive value help clinicians to interpret the results of diagnostic and prognostic tests. Both positive and negative predictive value are dependent on the prior probability of the event being predicted. In this study, prior probabilities corresponded to the error rate for look- and sound-alike medication errors. In practice, the rate is likely to be less than 5%.[41] For each measure of similarity (or of distance), the positive and negative predictive values were plotted at various prior probabilities (e.g., 0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 0.9, 0.95, 0.99, 0.999). The positive predictive values for bigram, trigram, and Levenshtein distance are plotted in Figures 10–12. The negative predictive values are plotted in Figures 13–15. Note that more specific tests yield higher positive predictive value, and more sensitive tests yield higher negative predictive value.

Discussion

Usefulness of the prognostic tests. If a prognostic test is to be viable, three conditions must be met. The first, and least difficult, condition is that cases and controls must be distributed differently with respect to the prognostic measure. Next, the prognostic measure must be a significant risk factor for the condition being predicted. Finally, the test must have sufficiently high sensitivity, specificity, and positive and negative predictive value.[29,32] All three of these conditions were met by the measures evaluated. Hypotheses 1–3 were supported by the data. In each case, error pairs and control pairs were distributed differently with respect to the measures of orthographic similarity. In each case, orthographic similarity was a significant risk factor for occurrence of an error. For at least one measure (Levenshtein distance) it was possible to construct a test with high values of sensitivity and specificity and sufficiently high positive and negative predictive values at the relevant prior probabilities. Given its sensitivity and specificity at a threshold of distance ≤5, Levenshtein distance was the best measure tested. By using this threshold in screening tests, one would expect to correctly identify 84% of all true error pairs and 98.8% of all nonerror pairs. In general, the rarer a condition, the more specific a test must be to be useful.[29] If a test for a

Figure 9. Receiver operator characteristic (ROC) curve for edit distance in the prediction of look-alike and sound-alike medication errors. The Levenshtein distance values at various thresholds are in parentheses.

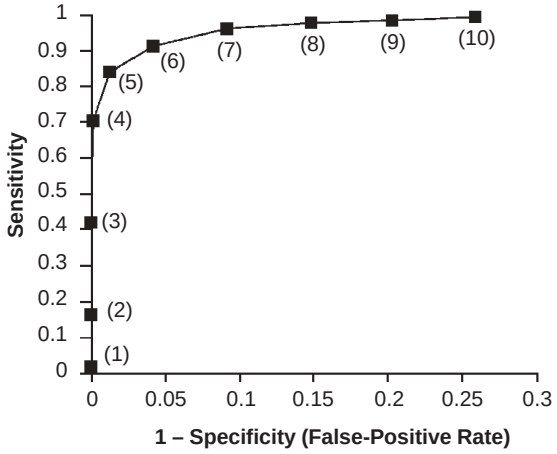


Figure 10. Positive predictive value of prognostic test based on bigram string similarity with 73% sensitivity and 99% specificity. The threshold for this test was similarity ≥0.3.

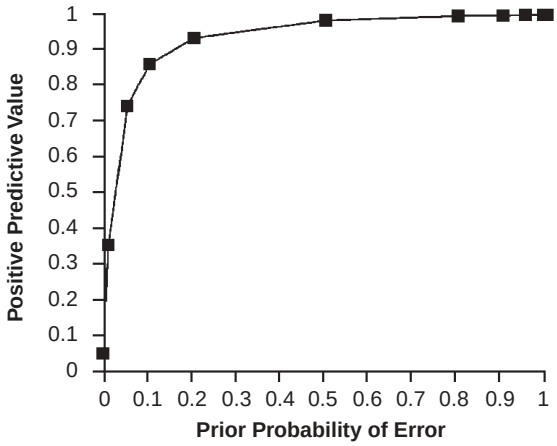


Figure 11. Positive predictive value of prognostic test based on trigram string similarity with 59% sensitivity and 99% specificity. The threshold for this test was similarity ≥0.2.

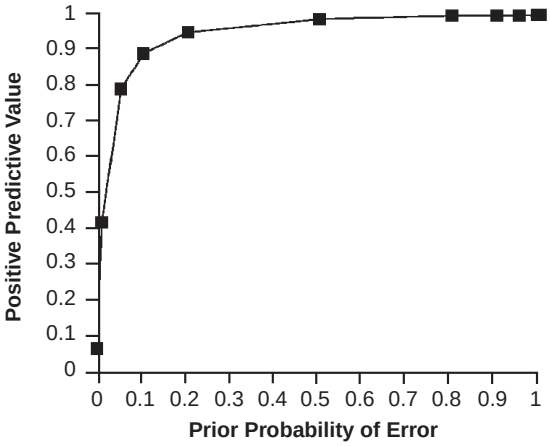
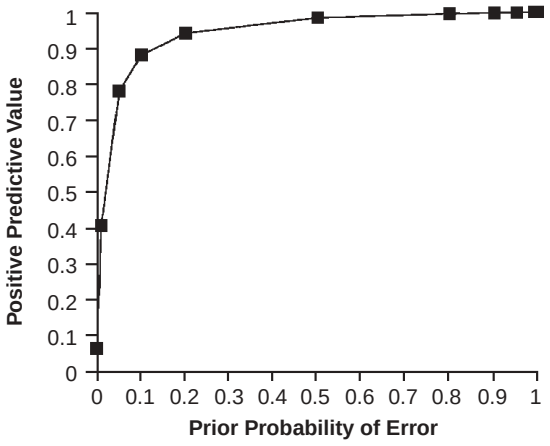


Figure 12. Positive predictive value of prognostic test based on Levenshtein distance with 84% sensitivity and 98.8% specificity. The threshold for this test was edit distance ≤5.



rare condition is not sufficiently specific, an unacceptably high number of false positives will be reported. Because medication errors were assumed to be rare events, i.e., to have an incidence of less than 1%, a more specific test was identified as the best for this study, even though another test had slightly better overall accuracy. However, in the final analysis, the precise placement of a threshold depends on a careful consideration of the societal costs of false positives versus false negatives.

Regulatory implications. Given these results, regulatory agencies approving new drug nomenclature would be well advised to explore the possibility of integrating automated tests of similarity into their routine name-approval procedures. One could imagine a scenario in which candidate names submitted to USAN and FDA were routinely screened against all existing names. Under this scenario, if the similarity between a candidate name and an existing name exceeded some established threshold (e.g., if the Levenshtein distance was ≤5), the candidate name would be refused, or perhaps contingently accepted with appropriate precautions. If the similarity between the candidate name and any of the existing names did not exceed the threshold, then the name would be approved. Similarly, every pair of names in the existing pharmacopeia could be screened in an effort to identify confusingly similar pairs. Once such pairs were identified, information about them could be added to the precautions section of drug references and to the contraindications field of electronic drug information systems.[5]

Need for debate on policy. If a test such as the one developed here were incorporated into the name-approval process, discussion would be required on how to

Figure 13. Negative predictive value of prognostic test based on bigram string similarity with 73% sensitivity and 99% specificity. The threshold for this test was similarity ≥0.3.

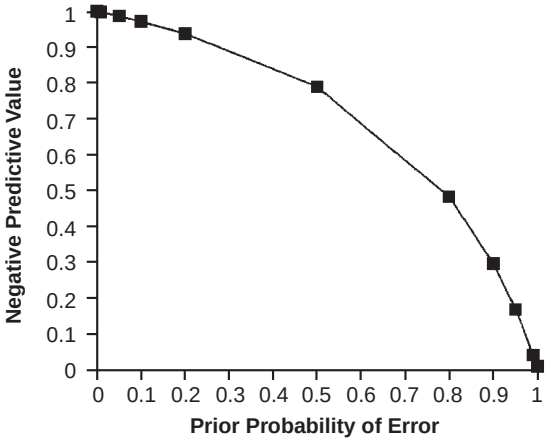
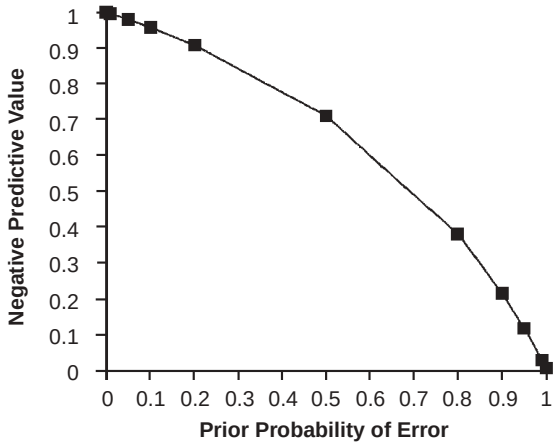
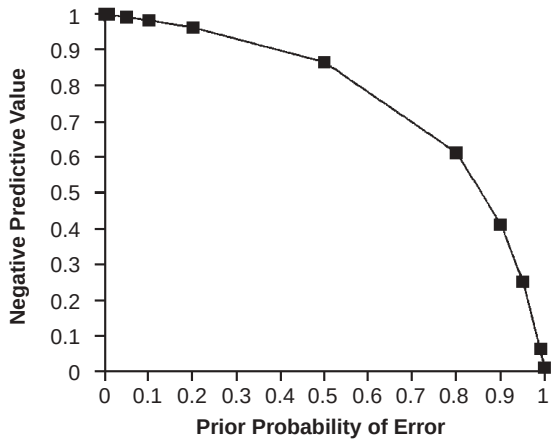


Figure 14. Negative predictive value of prognostic test based on trigram string similarity with 59% sensitivity and 99% specificity. The threshold for this test was similarity ≥0.2.



set the threshold. This discussion would need to focus on the costs associated with false positives and false negatives. A false positive could result in the prohibition of a name that was not, in fact, likely to cause an error. The main cost would be the opportunity cost to the company wanting to use a particular drug name. Not being able to use the name would mean forgoing the profits associated with one name while being forced to use an ostensibly less desirable, but potentially safer, name. A false negative, on the other hand, could result in approval of a medication name that might, in fact, be likely to cause a look-alike or sound-alike error. The costs here would be patient suffering as well as the monetary costs of avoidable hospitalizations and malpractice or liability litigation. Thus, the setting of a threshold becomes an important policy question that deserves public discussion and debate.

Figure 15. Negative predictive value of prognostic test based on edit distance with 84% sensitivity and 98.8% specificity. The threshold for this test was edit distance ≤5.



Conflicting concerns in medication naming. The effort to design error-resistant drug nomenclature is complicated by the need for new drug names that simultaneously satisfy commercial, professional, and safety concerns. New names must be reasonably safe and free from confusion but must also be meaningful and memorable to physicians, nurses, pharmacists, and patients. Although names need to be distinct, drugs that share an indication, mechanism of action, or chemical constituent are often intentionally given the same prefix or suffix.[9] Similarly, slight variants of some drugs are often given similar names, with only single letters distinguishing between products (e.g., Claritin and Claritin D), so that the value invested in one trademark can be easily transferred to another related mark owned by the same company. Safety, marketing, and professional concerns are all valid, but conflicts may arise during efforts to develop error-resistant nomenclature. Representatives of industry, the health professions, and government must work together to establish administrative procedures so that conflicting concerns can be resolved safely and efficiently.

Importance of theory. The tests developed here succeeded largely because they were based on sound, up-to-date psycholinguistic theory. Future attempts to reduce the incidence of medication errors will also benefit from a strong theory base, whether it be psychology, human factors, engineering, or computer science. With respect to look- and sound-alike medication errors, success will depend on an accurate and comprehensive understanding of the mental processes that underlie language production and comprehension.

Beyond medication errors. The general principles that support the tests developed here are applicable in other situations in which confusing nomenclature causes errors (e.g., in identifying confusing pairs of names of medical procedures and diagnoses) and could

even be used outside the medical field. In fact, several commercial trademark searching services, based on technologies related to those described here, already exist.[10]

Limitations. This investigation had several limitations. The database of known error pairs was small compared with the total possible number of confusing pairs. The published lists of error pairs were gathered by ad hoc methods that may have reflected selection biases: In effect, the error pairs were drawn from referral centers (e.g., the Medication Error Reporting Program, the Institute for Safe Medication Practices, and FDA MedWatch). The error pairs were thus likely to be more strikingly similar than unreported pairs of confusing names. The sensitivity of a prognostic test tends to be exaggerated under these circumstances.[29] At the same time, however, control pairs were selected at random and were not necessarily immune from error. It is possible that some of the control pairs used could have been involved in unreported medication errors. This would tend to cause specificity to be underestimated.

Even though a prognostic test was being evaluated, the research design was retrospective. So, rather than saying that test results accurately *predicted* errors, it would be more realistic to say that test results accurately distinguished between known errors and controls. A more convincing test would measure the similarity of all possible pairs in advance, track error rates prospectively, and then examine the relationship between similarity and error rate. Of course, because of well-known problems in error reporting and error surveillance (e.g., low incidence rates and underreporting), such an investigation would be difficult.

The cases and controls were not matched for frequency of prescribing, which is an important variable to consider when estimating the probability of confusion in actual practice. Each error listed in published reports was assumed to have occurred an equal number of times, even though errors involving the most frequently prescribed medications were likely to have occurred most frequently. The similarity measures studied ignored similarity in product labeling and packaging, dosage, indication, and physical appearance of the dosage form. International variations in spelling were ignored. Finally, the experiments ignored degrees of error severity: All errors were viewed as equally severe.

The phonological dimension of similarity was captured only indirectly by the orthographic measure used. Most English words obey regular rules that map spelling onto pronunciation.[16] That is, words that are spelled similarly are normally pronounced similarly. But similarities among irregular words, for which pronunciation is not a simple function of spelling, would not be captured by the present model, nor would mispronunciations and regional variations in pronunciation. These investigations did not draw clear distinctions among perceptual (e.g., visual or auditory) modalities or communication media (e.g., handwriting, typewriting, fax, computer monitor, telephone, face-to-face dialogue). Nor were distinctions drawn among recall, recognition, and short- or long-term memory, all of which are known to be distinct psychologically. The position of letter bigrams or trigrams within a given name was ignored, even though there is evidence that similarity in initial syllables is much more likely to cause errors than similarity in later syllables. Abbreviations and the number of syllables in each medication name were also ignored, even though both of these features are known to contribute to the potential for confusion. Some abbreviations may actually mitigate errors; unfortunately, research on this question is limited. Research on a model of phonological similarity is currently being undertaken to address these shortcomings.

Perhaps the most significant limitation is that these experiments treated medication errors as if they occurred in an abstract psychological realm. In reality, medication errors occur within complex and dynamic physical and organizational environments. Although we know that orthographic and phonological similarity contributes to errors, we still need a satisfactory model of medication errors that explains how contextual factors—psychological, environmental, and organizational—combine to cause or prevent errors. In light of these limitations, the results presented must be interpreted cautiously.

Conclusion

Automated measures of similarity between medication names can form the basis of highly accurate, sensitive, and specific tests of the potential for errors with look- and sound-alike names. These tests are comprehensive, theory-based, inexpensive, objective, and reliable. They lack certain features of expert evaluation of error potential, especially with respect to potential similarity in indication, packaging, and dosing, as well as methods for directly assessing phonological similarity.

References

1. Manasse HR Jr. Toward defining and applying a higher standard of quality for medication use in the United States. *Am J Health-Syst Pharm*. 1995; 52:374-9.
2. Cohen MR. Drug product characteristics that foster drug-use-system errors. *Am J Health-Syst Pharm*. 1995; 52:395-9.
3. United States Pharmacopeial Convention. Stop, look, and listen! *USP Qual Rev*. 1995; no. 49.
4. Boring D, Homonnay-Weikel AM, Cohen M et al. Avoiding trademark trouble at FDA. *Pharm Exec*. 1996; 16(6):80-8.
5. Cohen M. Novel way to prevent medication errors [resource on World Wide Web]. URL: http://www.ismp.org/ISMP/Novel.html. Available from Internet. Accessed 1996 Oct 11.
6. Davis NM, Cohen MR, Teplitsky B. Look-alike and sound-alike drug names: the problem and the solution. *Hosp Pharm*. 1992; 27:95-110.
7. DiDomizio G, Cohen M. International conference targets medication errors. *Trademark World*. 1994(Sep):32-7.
8. USP DI—Volume I: Drug information for the health care professional. Rockville, MD: United States Pharmacopeial Convention; 1995.

9. United States Pharmacopeial Convention, Inc. USP dictionary of USAN and international drug names. Rockville, MD: United States Pharmacopeial Convention; 1996.
10. Imsmarq Home Page [resource on World Wide Web]. URL: http://www.denpat.lu/imsmarq/imsmarq0.htm. Available from Internet. Accessed 1996 Oct 11.
11. Zobel J, Dart P. Phonetic string matching: lessons from information retrieval. In: 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval. Frei HP, Harman D, Schauble P et al., eds. Zurich, Switzerland: Association for Computing Machinery; 1996:166-72.
12. Dell GS. A spreading activation theory of retrieval in sentence production. *Psychol Rev*. 1986; 93:283-321.
13. Gathercole SE, Baddeley AD. Working memory and language. Hillsdale, NJ: Erlbaum; 1993.
14. Marslen-Wislon W, ed. Lexical representation and process. Cambridge, MA: MIT Press; 1989.
15. Seidenberg MS, McClelland JL. A distributed, developmental model of word recognition and naming. *Psychol Rev*. 1989; 96:523-68.
16. Plaut DC, McClelland JL, Seidenberg MS et al. Understanding normal and impaired word reading: computational principles in a quasi-regular domain. *Psychol Rev*. 1996; 103:56-115.
17. Levelt WJM. Speaking: from intention to articulation. Cambridge, MA: MIT Press; 1989.
18. Anderson JR. Learning and memory: an integrated approach. New York: Wiley; 1995.
19. Dijkstra T, de Smedt K, eds. Computational psycholinguistics. Bristol, PA: Taylor & Francis; 1996.
20. Altmann GTM, ed. Cognitive models of speech processing: psycholinguistic and computational perspectives. Cambridge, MA: MIT Press; 1990.
21. Luce PA, Pisoni DB, Goldinger SD. Similarity neighborhoods of spoken words. In: Cognitive models of speech processing: psycholinguistic and computational perspectives. Altmann GTM, ed. Cambridge, MA: MIT Press; 1990:122-47.
22. Frauenfelder UH. Computational models of spoken word recognition. In: Computational psycholinguistics. Dijkstra T, de Smedt K, eds. Bristol, PA: Taylor & Francis; 1996:115-38.
23. Grainger J, Dijkstra T. Visual word recognition: models and experiments. In: Computational psycholinguistics. Dijkstra T, de Smedt K, eds. Bristol, PA: Taylor & Francis; 1996:139-65.
24. Dell GS, Juliano C. Computational models of phonological encoding. In: Computational psycholinguistics. Dijkstra T, de Smedt K, eds. Bristol, PA: Taylor & Francis; 1996:328-59.
25. Emmorey KD, Fromkin VA. The mental lexicon. In: Linguistics, the Cambridge survey III: language: psychological and biological aspects. Newmeyer F, ed. Cambridge, MA: Cambridge Univ. Press; 1988:124-49.
26. Conrad R. Acoustic confusion in immediate memory. *Br J Psychol*. 1964; 55:75-84.
27. Underwood BJ, Freund JS. Errors in recognition learning and retention. *J Exp Psychol*. 1968; 78:55-63.
28. Martin N, Weisberg RW, Saffran EM. Variables influencing the occurrence of naming errors: implications for a model of lexical retrieval. *J Mem Lang*. 1989; 28:462-85.
29. Hulley SB, Cummings SR, eds. Designing clinical research. Baltimore: Williams & Wilkins; 1988.
30. Sackett DL, Haynes RB, Tugwell P. Cinical epidemiology: a basic science for clinical medicine. Boston: Little Brown; 1985.
31. Lilienfeld DE, Stolley PD. Foundations of epidemiology. New York: Oxford Univ. Press; 1994.
32. Fletcher R, Fletcher SW, Wagner EH. Clinical epidemiology: the essentials. Baltimore: Williams & Wilkins; 1996.
33. U.S. Food and Drug Administration. Medication errors. *FDA Med Bull*. 1996; 26(2):3.
34. Grabenstein JD, Proulx SM, Cohen MR. Recognizing and preventing errors involving immunologic drugs. *Hosp Pharm*. 1996; 31:791-804.
35. Stephen GA. String searching algorithms. River Edge, NJ: World Scientific; 1994.
36. Frakes WB. Stemming algorithms. In: Information retrieval: data structures and algorithms. Frakes WB, Baeza-Yates R, eds. Englewood Cliffs, NJ: Prentice-Hall; 1992:131-60.
37. Frakes WB, Baeza-Yates R, eds. Information retrieval: data structures and algorithms. Englewood Cliffs, NJ: Prentice-Hall; 1992.
38. Wagner RA, Fischer MJ. The string-to-string correction problem. *J ACM*. 1974; 21:168-73.
39. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum; 1988.
40. Schlesselman JJ. Case control studies. New York: Oxford Univ. Press; 1982.
41. Leape L. Preventing adverse drug events. *Am J Health-Syst Pharm*. 1995; 52:379-82.