A System for Multi-Attribute Drug Product Comparison

Bruce L. Lambert, Ph.D.
Associate Professor
Department of Pharmacy Administration
Department of Pharmacy Practice

Clement Yu, Ph.D.
Professor
Department of Computer Science

Mohanraj Thirumalai, M.S.
Graduate Student
Department of Computer Science

University of Illinois at Chicago

March 17, 2003

Short Title: Multi-Attribute Drug Comparison
Address: Department of Pharmacy Administration,
833 S. Wood Street (M/C 871), Chicago, IL 60612-7231
Phone: 312-996-2411
Fax: 312-996-0868
Email: lambertb@uic.edu
Word Count:

Abstract

We describe a system for multi-attribute drug product searching. We then demonstrate the system's performance on sample queries, and evaluate the name-based similarity searching component. Ten drug names were used to query a database of existing drug names using 5 different retrieval methods. Retrieved names were merged into master lists and presented to 15 pharmacists. Pharmacists rated the similarity between the query name and each retrieved names on a scale of 1 to 5. We report the precision of our 5 different retrieval methods at 11 levels of recall. The best single measure was editex, with a precision of 17.4% averaged across 11 levels of recall. A regression model using four objective measures of similarity as predictors accounted for 40.6% of the variance in observed mean similarity ratings. Automated, multi-attribute drug product searching may improve the effectiveness and efficiency of pre-approval screening process and thereby prevent medication errors.

Keywords: medication error, drug names, confusion, similarity, recall, precision, information retrieval, strength, dosage form, route of administration, dosing interval

Acknowledgments

The evaluation of this system was funded in part by a grant to Dr. Lambert from the National Patient Safety Foundation. We would like to thank several groups of computer science students who helped to implement earlier versions of the system. We also acknowledge the assistance of Dan Boring, Ken-Yu Chang, Michael Cohen and the Institute for Safe Medication Practices, Linda Kawano, David Lambert, Eric Lambert, Bob Lee, Swu-Jane Lin, Alan Sherr, Jill Tarzian-Sorensen, Michael Verga, and Justin Zobel. Send reprint requests to Dr. Lambert.

Introduction

Confusions between similar drug products continue to be among the most common kinds of medication mishaps, accounting for 15-25% of voluntary error reports.[1] One way to minimize the risk of such confusions is to screen proposed drug products against a database of existing products prior to approval, rejecting products that are too similar to existing products. Such pre-approval screening services are offered by several commercial trademark searching firms[2], but these systems have weaknesses. Notably, they lack the ability to search based on non-name attributes of drug products (e.g., dosage form, strength, route of administration, dosing schedule). In addition, their measures of similarity have not been disclosed, and their performance has not been subject to peer-reviewed, empirical validation. The inability to use non-name attributes as search criteria is especially problematic in light of reports that similarity in non-name attributes increases the risk of confusion (e.g., *Serzone*® 100mg and 200 mg capsules vs. *Seroquel*® 100mg and 200mg tablets).[3, 4] To show how some of these weaknesses may be overcome, we describe a prototype system for multi-attribute drug product searching. The system is based on recently patented techniques.[5] The system gives users the option to input multiple attributes of a drug product, including brand name, dosage form, dosage strength, and route of administration. Users can assign relative weights to each dimension of product similarity. After giving a technical description of our implementation, we demonstrate its performance on sample queries. Finally, we report the results of an empirical evaluation of the name-based similarity searching component.

System Overview

We designed and implemented a drug product searching system that enables a user to assess the confusability of a newly proposed product.[5] The user inputs a drug name and

(optionally) other drug product attributes such as strength, dosage form, and route of administration. The user then accepts the defaults or makes several choices about how to conduct the search and how to display the results. The system then retrieves and displays a list of drug products, ranked in descending order of similarity (or ascending order of distance) to the query product. With respect to the name-searching capabilities, the system is functionally similar to trademark search engines offered by a wide variety of commercial search firms.[2] However, several features make our system unique, namely: (a) similarity measures are explicitly described and have been subject to extensive experimental validation; (b) numerical similarity scores are displayed to the user along with ranked retrieval results; (c) mean similarity scores for large populations of drug names have been published elsewhere,[6] thus providing a context for interpreting similarity scores for new names; (d) the system has the capability to do exact and approximate matching on non-name attributes of drug products such as strength, dosage form, and route of administration; (e) the system permits the user to differentially weight product attributes based on their potential contribution to confusability; (f) the system permits users to see the non-name attributes of retrieved drug names by simply clicking on the retrieved name, unlike other systems, where non-name attributes must be looked up separately, and (g) the databases of existing drug products being searched are freely available to the public.[7, 8]

The system is implemented primarily in Java, with some modules written in C and C++. The interface is similar to many standard search engines, with fields for free-text input of query names as well as a series of buttons and drop-down menus used to specify the type of search, the number of results to be retrieved, and the manner in which retrieved results should be sorted and displayed (see Figure 1).

(Insert Figure 1 about here)

As noted above, the system has the ability to search based on name similarity, non-name attribute similarity, or some combination of both. We begin by describing the name-based similarity (and distance) measures. These measures have been described and experimentally validated elsewhere, so we only describe them briefly here.[6, 9-13] Within the name searching component, the user can specify whether similarity should be based on spelling or pronunciation.

Orthographic (Spelling) Similarity

In our current implementation, spelling similarity is assessed by way of bigrams, trigrams, or edit distance.[10, 14]

N-grams. N-gram measures capture spelling similarity by counting the number of *n*-letter subsequences that two words have in common. Currently, users must choose between bigram and trigram methods. The bigram method uses two-letter subsequences; the trigram method uses three-letter subsequences. Both measures range from zero to one, with zero representing the least similarity and one the most. For example, to compute the bigram similarity between *Acthar*® and *Acular*®, each word is broken down into its two-letter subsequences. For *Acthar*® this yields {*ac, ct, th, ha, ar*} and for *Acular*® {*ac*, *cu, ul, la, ar*}. All n-grams are converted to lower case before comparisons are made. The Dice coefficient is used to compute a similarity score between sets of bigrams (although other methods could be used to arrive at a numerical score once the number of common n-grams has been determined)[14-18]:

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

where *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*® and *Acular*® share two bigrams: {*ac, ar*}. Hence, the bigram similarity between *Acthar*® and *Acular*® is 2*2/(5+5) =

0.4.[a] To increase the sensitivity to similarity at the beginnings and endings of names, the user has the option of adding spaces to the beginning or ending of the names. For example, if the user selected trigram similarity with two spaces added to the beginning of each name, the similarity between *Accupril®* {*--A, -Ac, Acc, ccu, cup, upr, pri, ril}* and *Accutane®* {*--A, -Ac, Acc, ccu, cut, uta, tan, ane*}, which share 4 trigrams {*--A, -Ac, Acc, ccu* }, is (2*4)/(8 + 8) = 0.5.

N-Gram Position. As we have described it thus far, the similarity between two names based on n-grams has not taken into consideration *where* the common bigrams or trigrams occur within the two names. It has been observed that the beginnings and endings of words are more important than the middle part in causing confusion.[19] Although this emphasis on the beginning or ending of a word can be implemented by inserting one or more blank characters, a more general approach is to assign a weight associated with each bigram based on its position. More precisely, the highest weight is given to the first and the last bigrams, and the weights decrease toward the middle of the given name. Users may select an option which turns on this positional sensitivity.

Edit Distance. The final orthographic measure is edit distance. Edit distance refers to the number of edits (i.e., letter insertions, deletions, or substitutions) that must be made in order to transform one name into another.[9, 10, 14] For example, to transform *Ambien®* into *Amen®*, 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 case raw edit distance was divided by the maximum possible edit distance between two given

[a] An efficient way to compute the similarities between a given query name, *N*, and a set of existing drug names is to index all bigrams before the computation of similarities. That is, for each existing drug name, all bigrams are identified. Then, for each bigram, an inverted file for the drug names having that bigram is created. For example, if the bigram is "ty", an inverted file { 92, 143, 178} means that drug names with IDs 92, 143, and 178 contain that bigram. When the given name, *N*, is submitted, all its bigrams are extracted. For each bigram, the corresponding inverted file is retrieved. Then, the similarity of each drug name having that bi-gram is increased by an amount due to that bigram. This is repeated for each bigram of *N*.

words (i.e., the length of the longer of the two words). Thus the normalized edit distance between *Ambien*® and *Amen*® is 2/6=0.33. To transform edit distance into a similarity measure, one can use 1 minus the normalized edit distance. In this example, the similarity would be 1 - 0.33 = 0.67.

In the simplest implementation of edit distance, two characters that are not identical have an substitution cost of 1, but using a generalization of edit distance, some pairs of distinct characters can have substitution costs less than 1, depending on how similar they are to one another.[10, 16] The more similar the characters, the smaller their substitution cost. The similarity between two characters in turn depends on whether the characters are typed, printed or cursive. Consider the characters *m* and *n*. It is relatively easy to confuse these two characters. As a consequence, the substitution cost for the two characters should be less than 1. In general, the substitution costs between all pairs of characters are given by a matrix *M*, where $M(i, j)$ is the substitution cost between the *i*th character and the *j*th character.[14] Using this matrix, the edit-distance between two words can be computed using dynamic programming with the same time complexity given above. Whenever two characters from the two words differ, the substitution cost between the two characters is obtained from the matrix *M*.

Phonological (Pronunciation) Similarity

Confusing drug names do not just look alike, they often sound alike as well. In order to capture this sound-alike similarity, we have implemented a measure that assesses the distance between phonological transcriptions of drug names. The transformation of the written drug names into a sequence of phonemes is achieved using rules created by Fisher.[20, 21] Essentially, each character is converted into a phoneme based on the characters preceding it and those succeeding it. A phoneme is a basic unit of sound. Fisher's system uses a set of phoneme

symbols called the ARPAbet.[22] In Fisher's scheme, there are a total of 20,480 translation rules. Given a character and its preceding and succeeding characters, the rule which can utilize the highest number of adjacent characters to the given character is used for the conversion. More details are given as follows. The rules making use of the highest number of adjacent characters utilize 3 characters to the left of the given character and 4 characters to the right. It is denoted by L3R4. The rules with the next highest number of adjacent characters are L3R3, i.e. they utilize the closest 3 characters to the left and the closest 3 characters to the right of the given character. The other rules in descending order of priorities are L2R3, L2R2, L1R2, L1R1, L0R1 and L0R0.

As an example, consider the conversion of the name *Motrin*® into phonemes. The rules L3R4 are not applicable to the character "m", since the 3 characters preceding "m" are all blanks. The same applies to rules L3R3, L2R3, L2R2, L1R2, L1R1. Even the rule L0R1 is not applicable, when "m" is adjacent to "o" because no such rule exists for "m" and "o". The only applicable rule is L0R0, where "m" is converted to phoneme "*m*". When the character "o" is considered, the rule which is applicable and with the largest number of adjacent characters is L1R1, utilizing the adjacent characters "m" and "t". The phoneme is "*ow*". When this process is repeatedly executed for the entire name, the sequence of phonemes is "*m ow t r ih n*".

After names are converted to sequences of phonemes, the similarity between two names can be computed by any of the orthographic measures (n-gram or edit distance). In this case, a bigram is two consecutive phonemes. For example, for the sequence of phonemes *m ow t r ih n*, the bigrams would be (*m ow*), (*ow t*), (*t r*), (*r ih*) and (*ih n*). Internally, multi-letter ARPAbet phonemes (e.g., *ow, ih*) are replaced by unique symbols to facilitate orthographic comparisons using existing n-gram and edit distance measures.

Methods for Non-Name Attribute Searching

In addition to name-searching, the system has the capability to do exact and approximate searches on non-name attributes of drug products. The set of all non-name attributes can potentially be quite large, and might include: dosage strength, indication, dosage form, dosing interval, route of administration, manufacturer, pharmacologic category, storage requirements, color, shape, legal standing (prescription or over-the-counter), and trademark goods and services description.[5] The set of non-name attributes searchable in our prototype system is currently limited by the attributes that are available in the freely-available data sources we use (i.e., the FDA Orange Book and the Multum Lexicon).[7, 8] From the FDA Orange Book, these include: the active ingredient(s) for the product, dosage form, route of administration, trade name, applicant (manufacturer) name, strength or potency of the active ingredient, New Drug Application (NDA) number, and type or category of approved drugs (e.g., Rx, OTC, discontinued).[23] From the Multum Lexicon, it includes (among others): active ingredients, strength, route, dosage form, trade name, Controlled Substances Act (CSA) Schedule, Health Care Financing Administration (HCFA) payment codes, National Drug Code (NDC), pregnancy category, therapeutic classification, etc.[7] Many of these attributes can be displayed by double-clicking on a retrieved drug name (See, e.g., Figure 2 and 3).

(Insert Figure 2-3 about here)

Exact Matching. At the simplest level, the system has the capability to do exact matching on all non-name attributes. For example, if the proposed new product has a strength of 100 mg., then the system can identify all other products with the exact same strength. If the proposed product uses the oral route of administration, then the system can retrieve all other products that use the oral route.

Approximate Matching. More innovative approaches are required to achieve approximate matching between non-name attributes such as strength, dosage form, route of administration, and dosing interval. Our main approach is to define *equivalence classes* to capture degrees of similarity that are less than exact matching but greater than total distinctiveness. Consider, for example, the FDA Orange Book classification of dosage forms.[23] It includes, among others, those listed in the first column of the table below.

Dosage Form	Equivalence Class
Aerosol	1
Aerosol, metered	1
Bar, chewable	2
Capsule	3
Capsule, coated pellets	3
Capsule, delayed release pellets	3
Capsule, extended release	3
Concentrate	4
Cream	5
Cream, augmented	5
Cream, suppository	5
…	…
Troche/Lozenge	n

We assign each dosage form to an equivalence class, as illustrated in the second column of the table. Exact matching between dosage forms proceeds as described above. Approximate matching is based on membership in the same equivalence class, and non-matching is based in failure of exact and approximate matching. For example, if two products shared the exact same attribute value for dosage form they would get a similarity of 1 for the dosage form dimension. If the dosage forms were not identical but belonged to the same equivalence class (e.g., capsule vs. capsule, coated pellets), then they would get a score of 0.5 for the dosage form dimension. If they were neither identical nor members of the same equivalence class, their dosage form similarity score would be zero.

The equivalence classes displayed above are based on an intuitive grouping of dosage forms into similar sets. Finer gradations of approximate matching could be achieved by creating equivalence classes at different levels of granularity such that members of a Level 1 equivalence class were more similar than members of a Level 2 equivalence class, and so on. It should be noted that the equivalence class approach will work for all of the non-name attributes. Consider the route of administration attribute, whose values are also taken from the FDA Orange Book.

Route	Equivalence Class
Buccal	1
Buccal/sublingual	1
Dental	2
Endocervical	3
For Rx compounding	4
Implantation	5
Inhalation	6
Injection	7
…	…
Intramuscular	7
…	…
Vaginal	n

Note that these equivalence classes are tentative, and final class membership would be based on extensive expert review.

Weighted Combinations. Having the ability to search on multiple attributes raises the question of how much weight to attach to each type of attribute similarity. We handle this by allowing the user to associate a numerical weight (i.e., a real number between 0.0 and 1.0) with each attribute. The weights are constrained such that they summed to 1.0. If the sum of the weights suggested by the user exceeds 1 or is less than 1, then the system will automatically adjust the weights so that they sum up to 1 while preserving their relative magnitudes. These weights represent relative degrees of importance of the attributes. Default weights (set, for now, in an *ad hoc* fashion) are used if the user chooses not to assign explicit weights. For now, we

presume that the brand name is the most important attribute, followed, in no particular order, by the route of administration, the dosage form, the dosage strength, and the dosing interval or schedule. These elements are believed to be important because they typically appear on a written prescription, e.g., “Prozac 20mg tablets, PO (by mouth) BID (twice daily).”

Demonstration

In this section, we illustrate the basic functions of our system by way of examples. The first query name we have chosen is *Serzone*®. We chose this name because it has recently been the subject of error reports that involve both name and non-name attribute confusion.[3] The name most often confused with *Serzone*® appears to be *Seroquel*®. Thus, it will allow us to illustrate both name and attribute similarity capabilities of our system. Note however, that many other name could have been chosen.[1]

First, we do a simple orthographic name search, using trigram with two spaces inserted before the name as the similarity measure. The first 13 retrieved names are given in Figure 1.

(Insert Figure 1 about here)

Double-clicking on *Serax*® causes the window in Figure 3 to open. Notice that although the name *Serax*® appears only once in the search results, there are (at least) 10 distinct *Serax*® products in the database, each with a slightly different set of non-name attributes.

(Insert Figure 3 about here)

Note in Figure 3 that *Serax*® has the second largest similarity value but is listed (in the second column of Figure 3) as the 10th ranked product. That is because there are 9 distinct *Serzone*® products that precede it. All nine *Serzone*® products could be viewed by double-clicking on the name *Serzone*®.

Seroquel®, the product that has been confused with *Serzone*® on multiple occasions[3], has the fifth highest name-only similarity to *Serzone*®. Were one to consider non-name attribute similarity, the ranking would change from fifth to third. To search on non-name attributes, one must first input these attributes, and this is done by filling in the blanks in the window shown in Figure 4. In this case, we have input the brand name, the strength, the dosage form, the inner pack size and the route of administration. We weighted these attributes 0.5, 0.2., 0.1, 0.1, 0.1 respectively, as can be seen in the right hand column of Figure 4. As a result of weighted, multi-attribute searching, *Seroquel*® moves up in the rankings and several other names move as well (see Figure 5).

(Insert Figure 4-6 about here)

In addition to a traditional ranked list, we also allow the user to visualize the density of a query name's neighborhood graphically, as in Figure 6. In this figure, the radius of the circle represents the maximum similarity or distance (in this case a similarity of 1.0, occurring at the center of the circle). The center represents the query name. Names closer to the center are more similar to the query. Names further from the center are more distant from the query name. Clicking on any of the squares in this diagram would cause a product attribute window to appear, giving all of the relevant attribute similarity information about the selected drug. We are still experimenting with these diagrams as a user interface, but we believe they are important because they graphically illustrate the notion of *neighborhood density*, an idea that has great significance in the visual and auditory perception of words.[24-27] Note that, generally speaking, there is a large *zone of safety* around the target name, with most of the closest names being slightly different versions of *Serzone*® itself. *Serax*®, *Serentil*®, and *Seroquel*® are around the inner edge of the larger cluster of retrieved names. Presumably, users would begin to develop intuitions with

respect to these graphics, and could begin to make more refined judgments regarding whether or not an existing name violated the zone of safety of the new name.

We could illustrate several other features of our system, but space constraints prevent us from doing so. We have sought instead to illustrate the main features and the most novel features, including weighted, multi-attribute searching and graphical display of search results. Choosing different orthographic or phonological similarity measures or different weights for multi-attribute searching would, of course, alter the ranking of retrieved names. Thus, one might logically ask which search strategy or which set of parameters is optimal. Optimality in this context could be defined with reference to users' preferences or with reference to more formal measures of performance such as recall and precision.[28] Unfortunately, optimal search parameters are not presently known. In fact, there is no fixed set of search parameters that is optimal with respect to all possible query names and user preferences. Precisely how to select a set of search parameters is an ongoing focus of research in information retrieval.[16-18] In the next section, we describe a study of our orthographic and phonetic search measures designed to discover which measure or set of measures best captures experts' judgments of similarity.

Methods

We used standard information retrieval techniques to evaluate the name-searching aspect of our system.[16, 28] First, we selected five distinct similarity/distance measures to use as retrieval methods (edit distance, normalized edit distance, editex, edit-soundex, and trigram-2b).[10] We then created a set of 10 test queries. Each query was the name of a drug recently approved in the U.S. The names were selected at random from those approved by the FDA during 1998. They were *Avelox*®, *Curosurf*®, *Enbrel*®, *Ferrlecit*®, *Herceptin*®, *Ontak*®, *Priftin*®, *Provigil*®, *Raplon*®, and *Singulair*®. Each similarity measure was used to search a drug name database with each of the test queries. Since there

were 5 retrieval methods and 10 sample queries, we performed 50 searches total. The database of names was created by merging the 1998 USP Dictionary, the January 2000 update of the USP-DI General Index, and the USPTO Trademarks Registered database, December 1999 update edition.[29-31] (These databases are different from those currently used in our search system because this research was initiated in 1999, and we were using different databases at that time.)

For each query name, the top 50 names retrieved by each measure were combined into a master list with duplicates deleted. This resulted in a 1548 names, or an average of about 155 existing names being retrieved for each query name. In exchange for an honorarium, a panel of 15 practicing pharmacists, drawn from the Institute for Safe Medication Practice's (ISMP) practitioner network, rated, on a scale of 1 (not at all similar) to 5 (extremely similar) the similarity of the retrieved names compared to the query names. Thus, practitioner judgments served as the gold standard by which the retrieval results were evaluated. Recall and precision scores (see below) for each similarity measure were computed based on practitoners' pooled judgments. These scores were then used to evaluate the various retrieval methods.

Measures of Lexical Similarity and Distance

Five different measures of lexical similarity and distance were used to retrieve names from the database of existing names. The retrieval methods were: edit distance, normalized edit distance, editex, edit-soundex, and trigram-2b, are described in detail above or elsewhere.[6, 9-14, 16, 17, 24]

Assessment of Retrieval Effectiveness

We compared the retrieval effectiveness of five distinct methods using the method of pooled relevance judgments.[16] Since the National Institute of Standards and Technology began using this method in its annual text retrieval conferences, it has become the *de facto* standard procedure for assessing the performance of information retrieval systems.[32] The method has three steps. First, a set

of retrieval methods is selected for evaluation (e.g., edit distance, normalized edit distance, editex, edit-soundex, and trigram-2b). Next, a set of sample queries is identified (e.g., *Avelox*®, *Curosurf*®, *Enbrel*®, *Ferrlecit*®, *Herceptin*®, *Ontak*®, *Priftin*®, *Provigil*®, *Raplon*®, and *Singulair*®). Each method is then used to search the target database with each query. For each query, the retrieval results from all methods are merged into a single, de-duplicated list. Domain experts (in this case, practicing pharmacists) then assess, for each retrieved name, whether it is relevant or not relevant. In information retrieval, a relevant item is one that an expert user would deem useful and appropriate if it were retrieved in response to a given query. In this context, where names rather than whole documents are being retrieved, judges rated the relevance of each retrieved name on a semantic differential-type scale that ranged from 1 (not at all similar) to 5 (extremely similar). Raters were instructed to base their judgments of similarity on either spelling or pronunciation or both, but were instructed not to consider similarity in non-name attributes. Similarity scores were normalized to the range 0-1, with scores 1 through 5 mapping to 0, 0.25, 0.50, 0.75, 1.0 respectively.

Based on these relevance judgments, recall and precision scores were computed for each retrieval method.[33] *Recall* is defined as the number of relevant names retrieved in response to a given query divided by the total number of relevant names in the database. Recall is an index of the true-positive rate or sensitivity of a given retrieval method. *Precision* is defined as the number of relevant names retrieved in response to a given query divided by the total number of names retrieved (both relevant and irrelevant).

Relevance judgments are typically given as 0 or 1, all or nothing, but, as noted, we used relevance judgments given on a normalized 0 to 1 scale. When a document is judged to be partially relevant, one can apply the same definitions of recall and precision. Specifically, if a set of *n* retrieved documents has degrees of relevance r_1, r_2, .., r_n, then the total number of relevant documents retrieved is

the sum i.e. $r_1 + r_2 + r_3 + \ ... + r_n$. For example, if three retrieved documents have degrees of relevance = 0, 0.5 and 0.75, then the number of relevant documents retrieved is $0 + 0.5 + 0.75 = 1.25$. Thus, even using degrees of relevance, both recall and precision can be computed as indicated above. In our evaluations, for each query name, the total number of relevant names in the database was determined by summing the mean expert ratings across all the names retrieved for a that query. For each distinct retrieval method, the precision at 11 levels of recall was averaged across the 10 test queries. For each method, then, average precision was computed at 11 levels of recall (0, 0.1, 0.2, …, 0.9, 1.0). The resulting recall-precision graphs were plotted and used to visualize the comparative performance of each retrieval method.

Combining Retrieval Methods

Previous research has shown that combinations of retrieval methods often perform better than any of the methods when used individually.[16] Based on this insight, we performed a hierarchical linear regression analysis using multiple measures as predictors and using expert ratings as the outcome to be predicted. Presently, expert ratings are assumed to be the best possible basis for evaluating the similarity between new and existing names. But human experts are highly paid and busy. Therefore, it is desirable to have an automated system that could mimic expert judgment. The purpose of this analysis was to determine whether a combination of measures could predict experts' mean similarity ratings more accurately than any single measure. Because the edit distance measure was highly correlated with the other edit distance type measures, plain edit distance was omitted from our model. The other predictors were entered in the order of their correlation with the outcome. The outcome itself was skewed (with a long right-hand tail), and so we used a square-root transformation of the outcome to linearize the relationship between predictors and the outcome.[34] Thus, our final regression model used the following independent

variables: (a) editex, (b) normalized edit distance, (c) trigram-2b, and (d) edit-soundex. The dependent variable was the square root of the mean expert rating of each of the 1548 rated names.

Expert Judgments

As noted, the expert judgments themselves can be viewed as a type of retrieval method. (e.g., if one had an objective measure that exactly reproduced the mean judgments of our 15 expert raters.) For a given query, one can rank the names in descending order of the mean expert similarity ratings. Then one can compute recall and precision for the 50 most highly ranked names in the sorted list in the manner described above. Recall and precision based on expert judgments can be viewed as the *practical upper limit* of performance on this task because the judgments of experts are presumed, for the present purposes, to represent the gold standard of similarity. The other retrieval methods should be evaluated against this upper limit, not against the ideal performance of 100% recall and 100% precision. We computed and graphed recall-precision curves based on expert ratings, and we also reported precision averaged across 10 queries and 11 levels of recall.

Results

Mean Similarity Ratings

The mean similarity rating across all 1548 names and 15 judges was 0.18 on a scale of 0 to 1 (SD = 0.12, median = 0.15, mode = 0.10). According to our rating scale and normalization procedure, the mean, median, and modal scores all fell between the similarity levels labeled “not at all similar” and “not very similar.” Figure 7 is a histogram of mean similarity ratings tabulated across all 1548 retrieved names and averaged across 15 expert judges. It shows that 87.21% of all names had mean ratings less than 0.4 (0.5 meant “slightly similar”). Only 2.38% had mean ratings greater than 0.5. Thus, for any given query name, only a very small percentage of existing

names were found that were more than "slightly similar" to the query name, in the opinion of experienced pharmacists. This is both consistent with what has previously been reported in an analysis of the U.S. drug name lexicon, and it is a logical consequence of the pre-approval screening process that seeks to keep highly similar names from entering the marketplace.[6, 35-37]

Recall-Precision Analysis

Based on summed mean similarity judgments from expert raters, the total number of relevant names in the database was as follows: *Avelox*® (30.53), *Curosurf*® (21.07), *Enbrel*® (32.21), *Ferrlecit*® (30.62), *Herceptin*® (21.77), *Ontak*® (30.72), *Priftin*® (24.51), *Provigil*® (27.14), *Raplon*® (30.78), and *Singulair*® (23.1). Because experts rated most names as only partially relevant/similar, the total number of relevant names is somewhat misleading. For *Avelox*®, for example, the database contained not 30 *fully* relevant names but greater than 30 *partially* relevant names. All the fractional relevance scores for *Avelox*® summed to 30.53, so that is the total we used in our recall and precision calculations. Figures 8-13 illustrate the recall-precision curves for the five tested retrieval methods and for the expert judgments.

Retrieving the 50 most highly rated names, the expert judgments yielded mean precision of 26.72% averaged across 10 queries and 11 levels of recall. Thus, 26.72% mean precision should be viewed as the upper limit on performance in this task. The performance of edit distance (mean precision across 11 levels of recall = 17.4%), editex (mean precision = 15.4%), and normalized edit distance (mean precision = 14.7%) was similar. Trigram-2b performed somewhat worse (mean precision = 11.37), followed by edit-soundex (mean precision = 5.5%). Note that only 50 names were retrieved by each individual method. In contrast, the total set of relevant names was determined by summing the mean expert ratings across the larger set of roughly 155 names that resulted from merging and de-duplicating the names retrieved by all five methods combined. As a

result, for each method, some (partially) relevant names was not among the top 50 names retrieved by that method. Thus, recall did not reach 100 percent for any single method. Precision was, by definition, zero when a given recall level is not reached.

(Insert Figures 7-13 about here.)

Correlation and Regression Analysis of Multiple Measures

Table 1 gives the correlations between all five individual methods and the mean similarity ratings produced by human experts. All but the edit-soundex measure had strong and significant correlations with the outcome, and these correlations were strengthened further by taking the square root of the experts' mean ratings as the outcome. As expected, the similarity and distance measures captured significant dimensions of expert similarity judgments.[12] Also, the edit distance measures (plain edit distance, editex, and normalized edit distance) were highly correlated with one another. These same measures were negatively correlated with trigram-2b because trigram-2b is a similarity measure.

Table 2 gives the result of the hierarchical multiple linear regression analysis. The four variables were entered in stepwise fashion using their simple correlations with the outcome to set the order of entry. The table gives test statistics indicating the significance of the contribution of each newly added term. The table shows that each variable adds significantly to the fit of the model and that the combined model is better than any model based on fewer variables. The final model was y = 0.69 - 0.01*Editex - 0.30*NED + 0.22*Trigram2b - 0.02*EditSoundex. The model accounted for 40.6% of the variance in the square root of the mean expert ratings. Figure 14 displays the fit between observed and predicted similarity ratings.

(Insert Tables 1-4 and Figure 14 about here.)

To illustrate the performance of the various methods, Tables 3 and 4 list the top 20 names retrieved by each method and the 20 names most highly rated by experts for two separate query names (*Avelox*® and *Curosurf*®) . Close inspection of these tables reveals that the combined method comes closer to the expert method than any of the individual methods. In several cases, there were ties among the similarity or distance scores of the top 50 names. When two names had the same similarity score, they were sorted in alphabetical order. Consequently, some more similar names were lower down the list because names beginning, for example, with "Av" were sorted lower alphabetically than names beginning with "Ab". In retrospect, we could have avoided this arbitrary tie-breaking by retrieving the names with the top 50 *scores*, rather than retrieving the top 50 *names* and then truncating ties beyond 50 names.[b]

Discussion

Evaluation of Name Retrieval Methods

In terms of recall, or the percentage of relevant names included among the 50 most highly ranked names, the five methods we tested retrieved between 40% and 60% of the (partially) relevant names in the database. For example, using the editex method to search for names similar to *Avelox*®, the 50 most similar names included names whose partial similarity scores summed to roughly 15; whereas, expert ratings indicated that all of the relevant names in the database would have partial relevance scores that summed to 30.53. Practically speaking, this means a user would need to look beyond the top 50 names in order to identify all relevant existing names in the database. The precision (i.e., 1 - false positive rate) of individual retrieval methods, averaged across 10 query names and 11 levels of recall, ranged from between 5.5% and 17.4%. At first glance, the false positive rate seems quite high (i.e., precision is quite low). Remember, however,

[b] Zobel and Dart dealt with ties by permuting the order of tied names and computing recall and precision based on the average of 10 permuted orderings.

that the practical upper limit on precision (e.g., that which would be achieved using expert ratings to rank existing names for retrieval) was found to be only 26.72%.

In absolute terms, this is not a very impressive level of performance. But this is a hard problem, as evidenced by the continued failure of existing methods to identify potential confusion problems before new drugs enter the marketplace.[1, 38] What's more, we are aware of no previous study that assesses the retrieval performance of a drug name searching system. To our knowledge, none of the dozens of commercial trademark searching firms has ever published a performance evaluation of their own system. Among published research studies, a study by Zobel and Dart is the most similar to our study, although it is by no means identical.[17] That study examined phonetic retrieval methods in a task that involved retrieving proper names from a telephone directory. They reported mean precision across 11 levels of recall ranging between 7% and 28%. Thus, the performance of our measures is in the same range, if perhaps a bit worse on average. One should be cautious even in this comparison however, since the domain are different (drug names vs. proper names) and the techniques for assessing relevance were also different.

In terms of automated measures, the combined model is likely to perform better than any of the individual methods evaluated above but worse than the expert judgments. We have attempted to illustrate the combined model's performance in Tables 1-4 and in Figure 14. Unfortunately, it was not possible to evaluate the recall and precision of the combined model, because it retrieved several names that were not retrieved by the individual methods. We had no relevance (similarity) ratings on these names and so could not compute recall or precision. Still, the fact that the combined model is more highly correlated with experts' ratings than any individual method, combined with the assumption that expert ratings constitute the current gold standard for assessing similarity, leads us to assume that it would yield a higher level of recall-

precision performance than any individual method tested above. One aim of our ongoing research is to build a predictive model, based only on automated measures, that can more closely mimic expert judgments.

Multi-Attribute Searching

Medication error reports continue to highlight the role of non-name attributes as causes of drug product confusions. When the drug name on a chart or prescription is ambiguous, health professionals often refer to the dosage strength, route of administration, dosage form, or dosing schedule to help them correctly identify the intended product. It stands to reason, therefore, that the more attributes two products have in common, the more difficult it is to distinguish between them and the more likely it is that they will get confused. Few argue with this logic, but surprisingly, none of the current systems for pre-approval screening of drug products allows one to take these non-name attributes into account when conducting searches. In fact, most existing search systems are *name searching* systems. We have described and demonstrated what we believe is the first *drug product* searching system capable of exact or approximate, weighted or unweighted, multi-attribute similarity searching.[5] Although, in this report, we have not produced direct evidence that the ability to search on non-name attributes increases search effectiveness in terms of recall or precision, the advantages to the user of such capabilities seem obvious. Producing evidence of the benefits of multi-attribute searching is high on our future research agenda.

Limitations

The results presented above should be interpreted in light of certain limitations. We only examined 10 query names, all approved in the same year. Averaging over a larger number of queries will give a better indicator of real-world performance. For each query, we only retrieved

50 names. This decision was made to minimize the burden on human raters, but it meant that we could not estimate recall and precision when more than 50 names were retrieved. In the high-stakes setting of drug name screening, users often examine many more than just the top 50 names. We only evaluated 5 retrieval methods. In our previous work, we have implemented more than 20 such methods. Even though some of these measures are highly correlated, our present results show that any additional independent information about similarity can be used to improve the predictive accuracy of a model. Thus, the four variable model we present above could likely be improved by adding additional measures of look alike and sound alike similarity.

With regard to our search system, it is still a prototype. A more polished web interface would be preferable to the current Java applet. We currently have no way to dynamically update the underlying databases. The default weights on multiple attributes are set in an ad hoc fashion. The equivalence classes that support approximate matching of non-name attributes need refinement and validation, and some important attributes (e.g., dosage schedule) still need to be integrated into the system. Procedures for sorting and saving search results also need improvement. In spite of these limitations, we contend that our system represents a significant improvement over existing search systems, for reasons outlined above. Concretely, we imagine that the system would be used: (a) by trademark attorneys within drug companies when screening potential new trademarks; (b) by FDA or other regulators during pre-approval screening of new drug names; and (c) by health-system pharmacy-and-therapeutic committees when assessing the confusability of a name being considered for addition to the formulary or when looking for confusing pairs within an existing formulary. In each context, the user would submit the name and other product attributes, conduct a search, and select the highly ranked

names for more detailed scrutiny. In the end, the decision about the acceptability or confusability of a name would be made by human experts.

Conclusion

Medication errors involving confusion between similar drug products are a source of ongoing concern to health professionals.[3, 10, 11, 13, 24, 35, 39-41] There is a growing awareness that what used to be viewed as a problem of *drug name* similarity is actually a problem of *drug product* similarity. Similarity in non-name attributes of drug products such as dosage strength, dosage form, route of administration, and dosing interval (schedule), is increasingly seen as contributing to the potential for confusion. In fact, the FDA recently issued a request for proposals and funded a contract to build a prototype system for multi-attribute drug product searching.[42] We described and demonstrated a system for drug product searching that enables users to search for similar products based on name and non-name attributes, or any weighted combination thereof. This capability may increase the efficiency and effectiveness of pre-approval screening of new drug products. We also presented the results of an evaluation of the name searching component of our system. The results of that evaluation demonstrated again the difficulty of the underlying problem. They also showed that, compared to a single measure, an automated model that incorporates multiple measures of similarity, based on both spelling and sound, will more accurately predict expert similarity judgments. Systems that exploit these insights have the potential to prevent medication errors that involve confusion between similar drug products.

References

1. U. S. Pharmacopeia. USP Quality Review. *U. S. Pharmacopeia*. March. Available at: http://www.usp.org/reporting/review/qr76.pdf. Accessed July 19, 2001.

2. Avantiq. Avantiq home page. Available at: http://www.avantiq.com/visitors/main.htm. Accessed March 3, 2003.

3. Kim H-J, Phillips J. Medication errors associated with Serzone and Seroquel. *Drug Topics.* Vol January 7; 2002:38, 40.

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

5. Lambert BL, inventor. Apparatus, method, and product for multi-attribute drug comparison. US patent 6,529,892. March 4, 2003.

6. Lambert BL, Chang KY, Lin SJ. Descriptive analysis of the drug name lexicon. *Drug Inf J.* 2001;35:163-172.

7. Multum Information Services Inc. The Multum Lexicon. *Multum Information Services Inc.* Available at: http://www.multum.com/Lexicon.htm. Accessed January 28, 2003, 2003.

8. U. S. Food and Drug Administration. Orange Book data files. *U. S. Food and Drug Administration* [Web page]. January 29, 2003. Available at: http://www.fda.gov/cder/orange/obreadme.htm. Accessed February 28, 2003.

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

10. Lambert BL, Lin S-J, Gandhi SK, Chang K-Y. Similarity as a risk factor in drug name confusion errors: The look-alike (orthographic) and sound-alike (phonological) model. *Med Care.* 1999;37(12):1214-1225.

11. Lambert BL, Chang KY, Lin SJ. Effect of orthographic and phonological similarity on false recognition of drug names. *Soc Sci Med.* 2001;52:1843-1857.

12. Lambert BL, Donderi D, Senders J. Similarity of drug names: Objective and subjective measures. *Psychology and Marketing.* 2002;19(7-8):641-661.

13. Lambert BL, Chang K-Y, Lin S-J. Immediate free recall of drug names: effects of similarity and availability. *Am J Health-Syst Pharm.* 2003;60:156-168.

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

15. Ukkonen E. Approximate string-matching with q-grams and maximal matches. *Theoretical computer science.* 1992;92:191-211.

16. Zobel J, Dart P. Phonetic string matching: Lessons from information retrieval. Paper presented at: 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; August 18-22, 1996; Zurich, Switzerland.

17. Zobel J, Dart P. Finding approximate matches in large lexicons. *Software-Practice and Experience.* 1995;25(3):331-345.

18. Pfeifer U, Poersch T, Fuhr N. Searching for proper names in databases. Paper presented at: Hypertext - Information Retrieval - Multimedia: Synergistic Effects of Electronic Information Systems; 5-7 April 1995, 1995; Konstanz, Germany.

19. Gernsbacher MA, ed. *Handbook of psycholinguistics*. San Diego, CA: Academic Press; 1994.

20. Fisher WM. A statistical text-to-phone function using ngrams and rules. Paper presented at: IEEE International Conference on Acoustics, Speech and Signal Processing; March 15-19, 1999; Phoenix, AZ.

21. Fisher WM, Fiscus JG, Martin A, Pallett DS, Pryzbocki MA. Further studies in phonological scoring. *Proceedings of the Spoken Language Systems Technology Workshop.* 1995:181-186.

22. Jurafsky D, Martin JH. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall; 2000.

23. U. S. Food and Drug Administration. Electronic Orange Book data files. *U. S. Food and Drug Administration* [Web page]. Available at: http://www.fda.gov/cder/orange/obreadme.htm. Accessed January 28, 2003, 2003.

24. Lambert BL, Chang K-Y, Gupta P. Effects of frequency and similarity neighborhoods on pharmacists' visual perception of drug names. *Soc Sci Med.* in press.

25. Luce PA, Pisoni DB. Recognizing spoken words: The neighborhood activation model. *Ear & Hearing.* 1998;19(1):1-36.

26. Luce PA, Pisoni DB, Goldinger SD. Similarity neighborhoods of spoken words. In: Altmann GTM, ed. *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Cambridge, MA: MIT Press; 1990:122-147.

27. Grainger J, Jacobs AM. Orthographic processing in visual word recognition: A multiple read-out model. *Psychol Rev.* 1996;103(3):518-565.

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

29. U. S. Pharmacopeia. *USP DI, Vol. I: Drug information for the health care professional*. Rockville, MD: U. S. Pharmacopeia; 1998.

30. U. S. Pharmacopeia. *USP dictionary of USAN and international drug names*. Rockville, MD: U. S. Pharmacopeia; 1998.

31. U.S. Patent and Trademark Office. Trademarks Registered. *U.S. Patent and Trademark Office--Office of Electronic Information Products* [CD-ROM]. February 1999. Available at: http://www.uspto.gov/web/offices/ac/ido/oeip/catalog/tmcassis.htm#TMregistered. Accessed June 15, 1999, 1999.

32. Voorhees EM, Harman D. Overview of the sixth text retrieval conference (TREC-6). Paper presented at: The sixth text retrieval conference (TREC-6); November 19-21, 1997, 1998; Gaithersburg, MD.

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

34. Cohen J, Cohen P. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1983.

35. Phillips J. The name game: New realities at FDA. *Pharmaceutical Executive.* Vol July; 2000:66-69.

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

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

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

39. Cohen M, ed. *Medication errors*. Washington, DC: American Pharmaceutical Association; 1999.

40. Kohn LT, Corrigan J, Donaldson MS, eds. *To err is human: building a safer health system*. Washington, DC: Institute of Medicine; 2000.

41. Institute for Safe Medication Practices. *ISMP Medication Safety Alert!* Huntingdon Valley, PA: Institute for Safe Medication Practices; January 9, 2003 2003. 7(7).

42. U. S. Food and Drug Administration. Automated method of minimizing medication errors due to similar proprietary and established names (RFQ #223-02-5618). Center for Drug Evaluation and Research: U. S. Food and Drug Administration; 2002.

43. Pedhazur EJ. *Multiple regression in behavioral research: Explanation and prediction.* New York: Harcourt Brace; 1997.

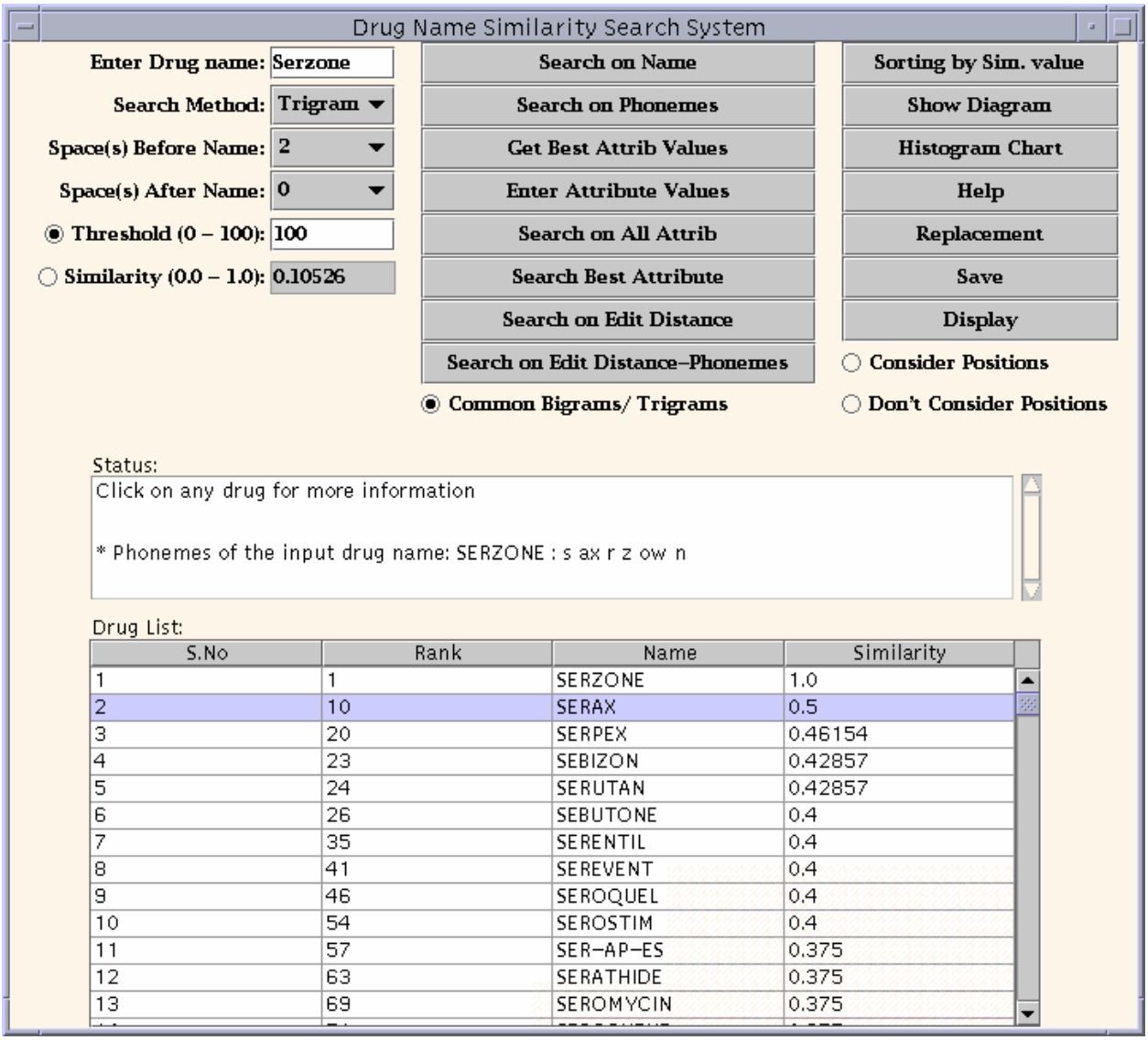

Figure 1. Main search screen with results from trigram search on query name *Serzone*®.

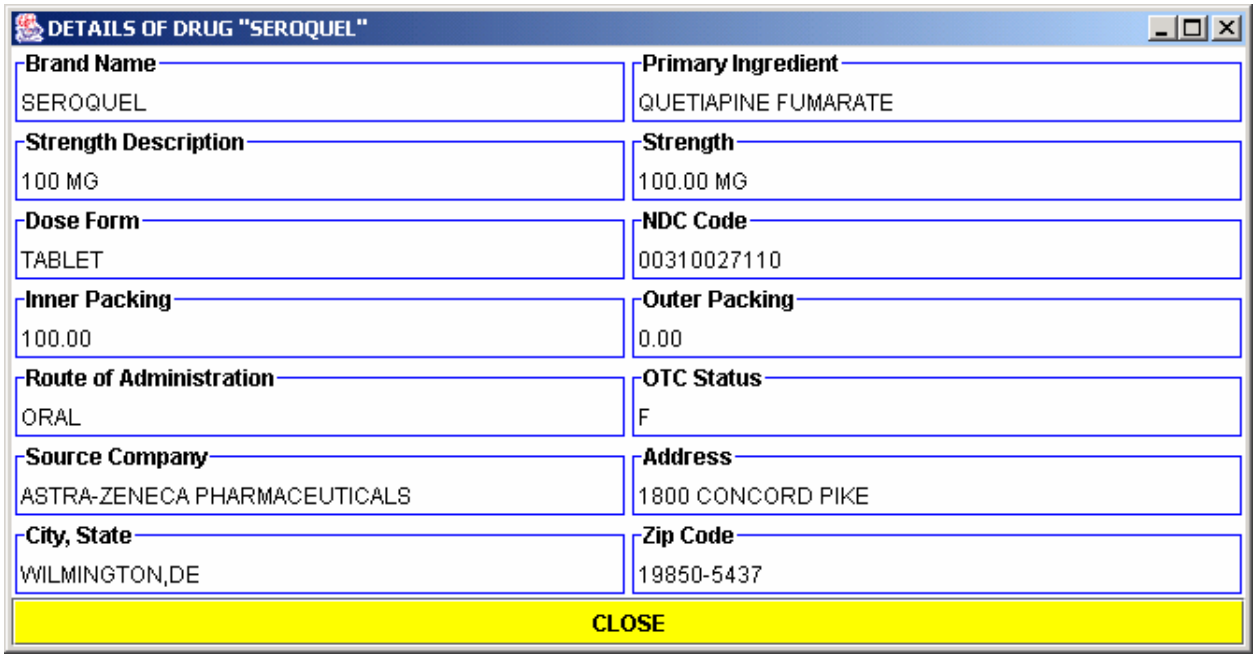

Figure 2. Multiple attribute information about *Seroquel*®.

Drug List

S.No	Rank	Name	Similarity	Phonemes	Strength	Doseform	NDC Code	Inner Packing
1	10	SERAX	0.5	s eh r ae k s	10.00 MG	CAPSULE	00008005...	25.00
2	11	SERAX	0.5	s eh r ae k s	10.00 MG	CAPSULE	00008005...	100.00
3	12	SERAX	0.5	s eh r ae k s	10.00 MG	CAPSULE	00008005...	500.00
4	13	SERAX	0.5	s eh r ae k s	15.00 MG	CAPSULE	00008000...	25.00
5	14	SERAX	0.5	s eh r ae k s	15.00 MG	CAPSULE	00008000...	100.00
6	15	SERAX	0.5	s eh r ae k s	15.00 MG	CAPSULE	00008000...	500.00
7	16	SERAX	0.5	s eh r ae k s	15.00 MG	TABLET	00008031...	100.00
8	17	SERAX	0.5	s eh r ae k s	30.00 MG	CAPSULE	00008005...	25.00
9	18	SERAX	0.5	s eh r ae k s	30.00 MG	CAPSULE	00008005...	100.00
10	19	SERAX	0.5	s eh r ae k s	30.00 MG	CAPSULE	00008005...	500.00

OK

Figure 3. Multiple instances of the drug *Serax*®.

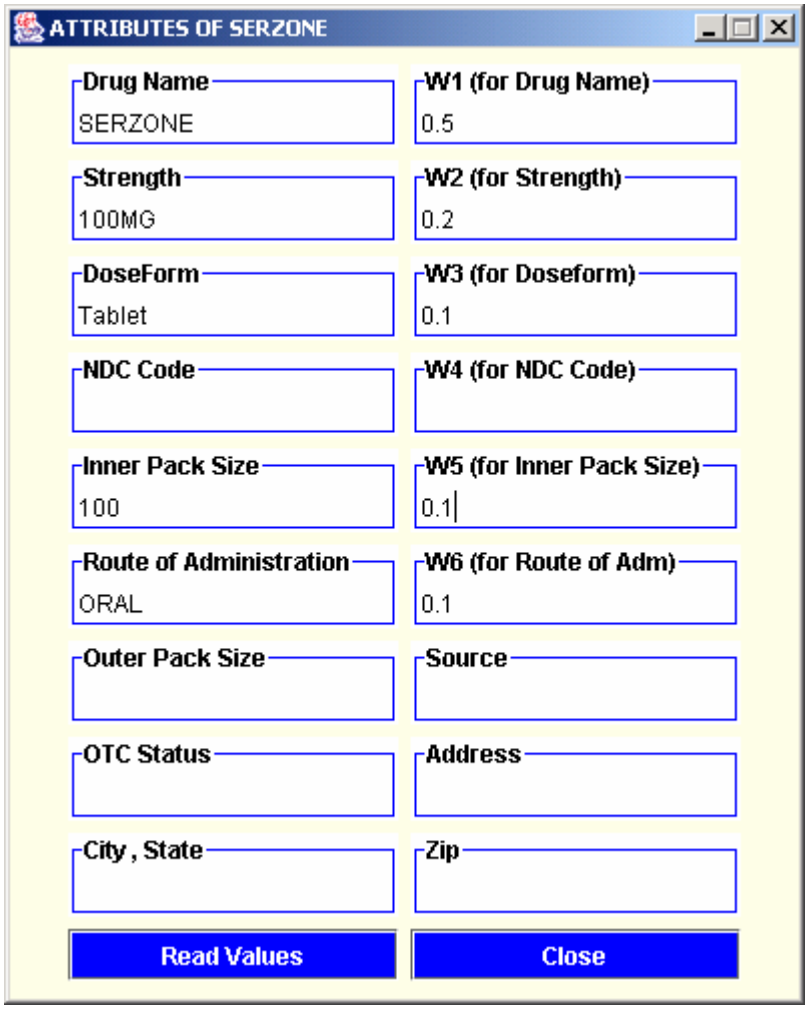

Figure 4. Data entry screen for multi-attribute values and weights.

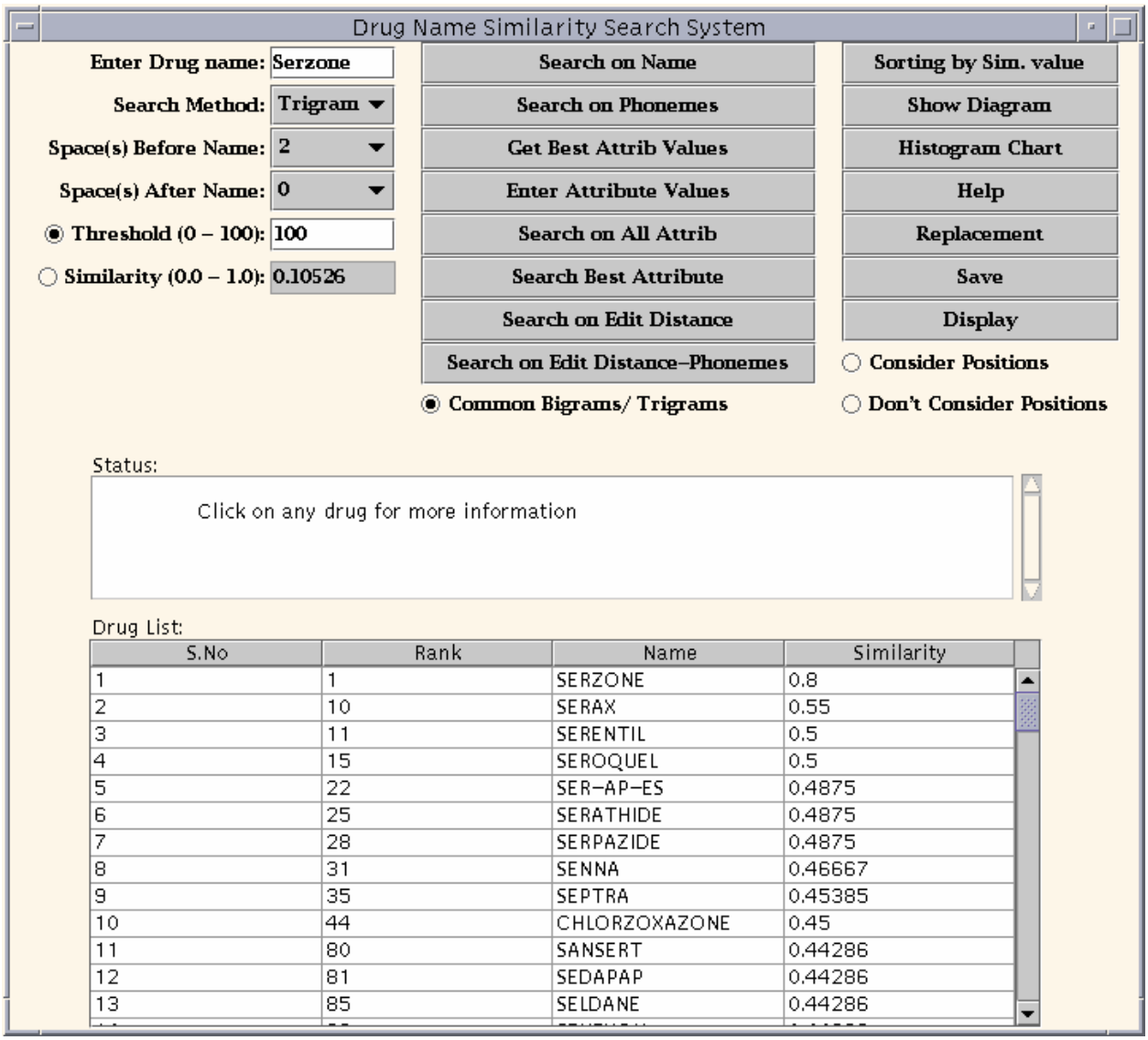

Figure 5. Search results for multi-attribute search on Serzone 100mg tablets, oral route of administration

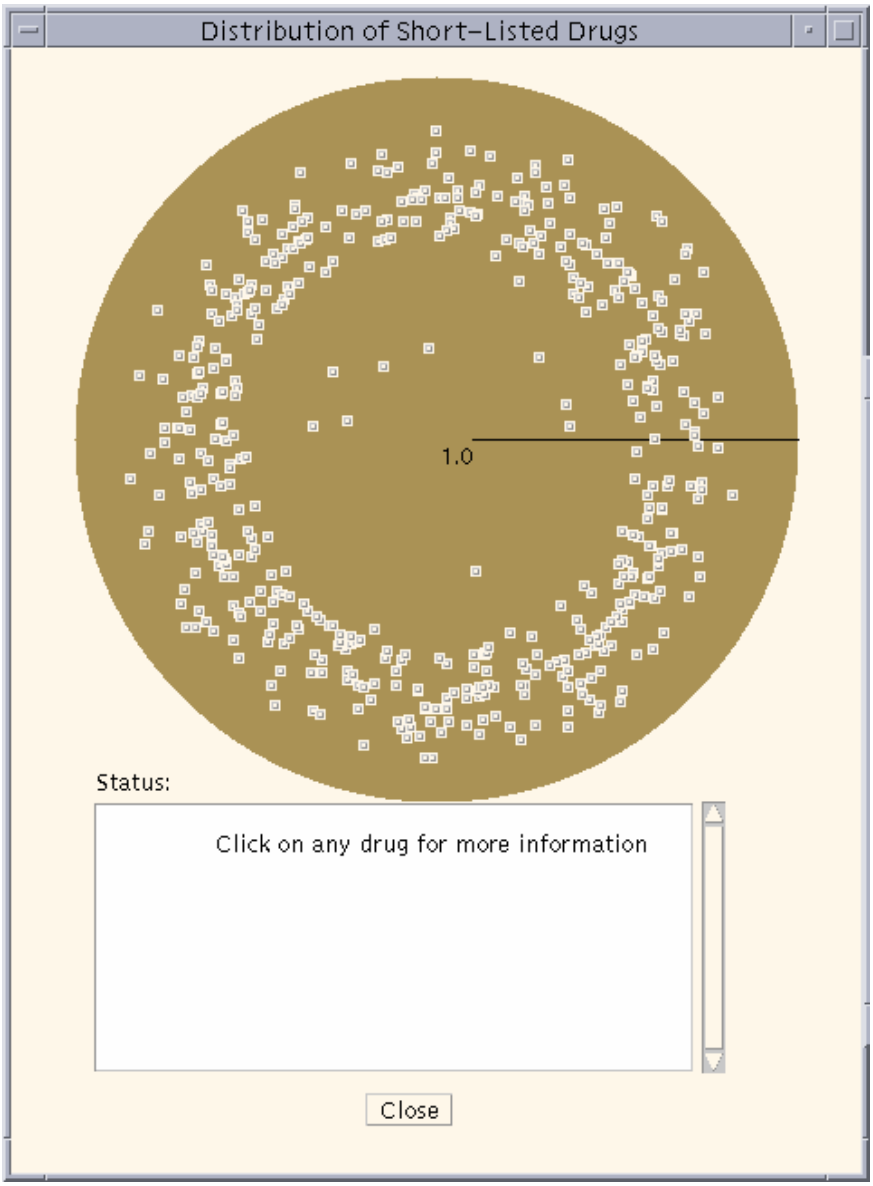

Figure 6. Neighborhood density diagram for query name *Serzone*®

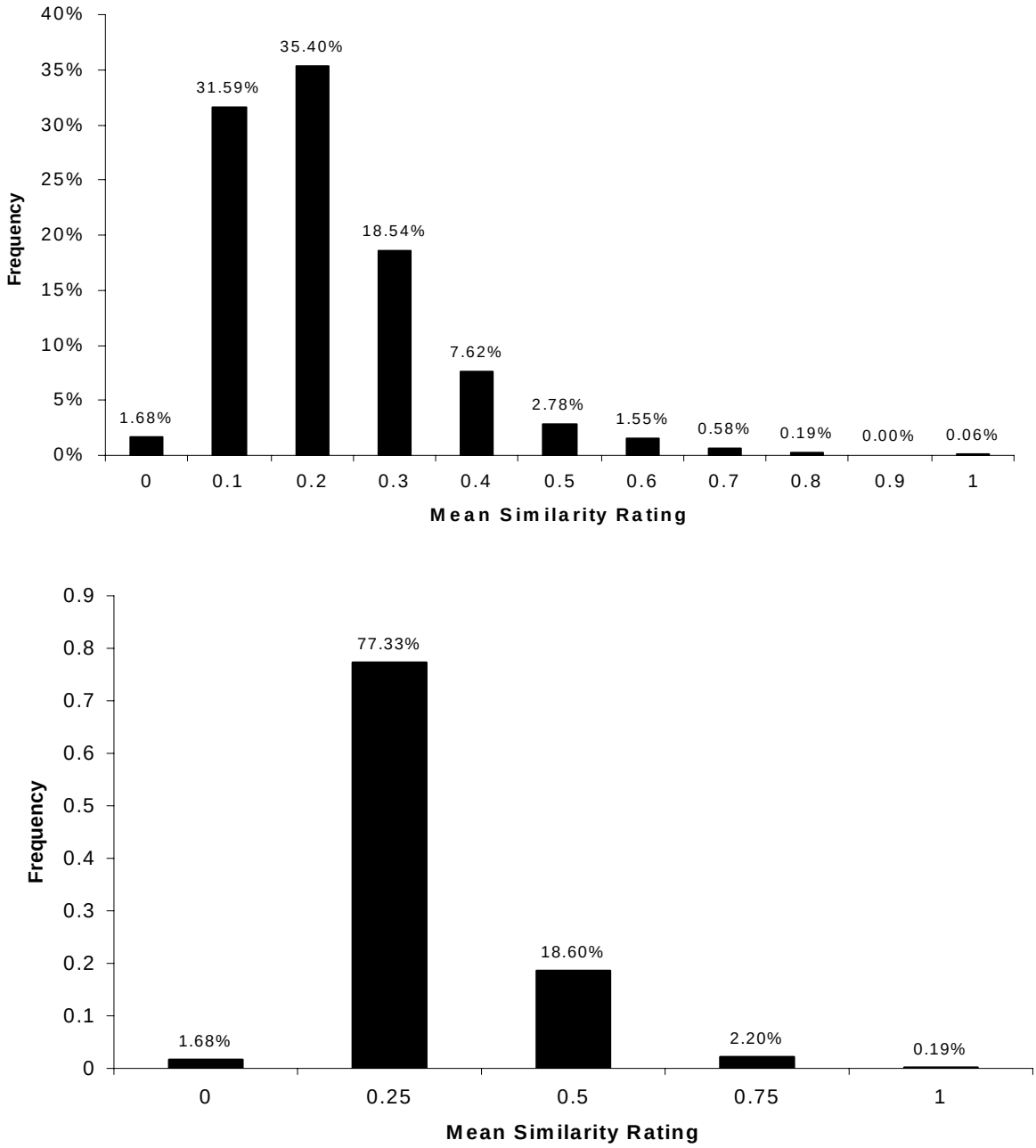

Figure 7. Histograms of mean similarity ratings for 1548 retrieved names, averaged across 15 expert judges. Original similarity scores were given on a scale of 1 to 5, with levels labeled “not at all similar,” “not very similar,” “slightly similar,” “quite similar,” and “extremely similar.” These scores were normalized to the interval 0 to 1, with levels 1 to five mapping to 0, 0.25, 0.5, 0.75 and 1.0 respectively.

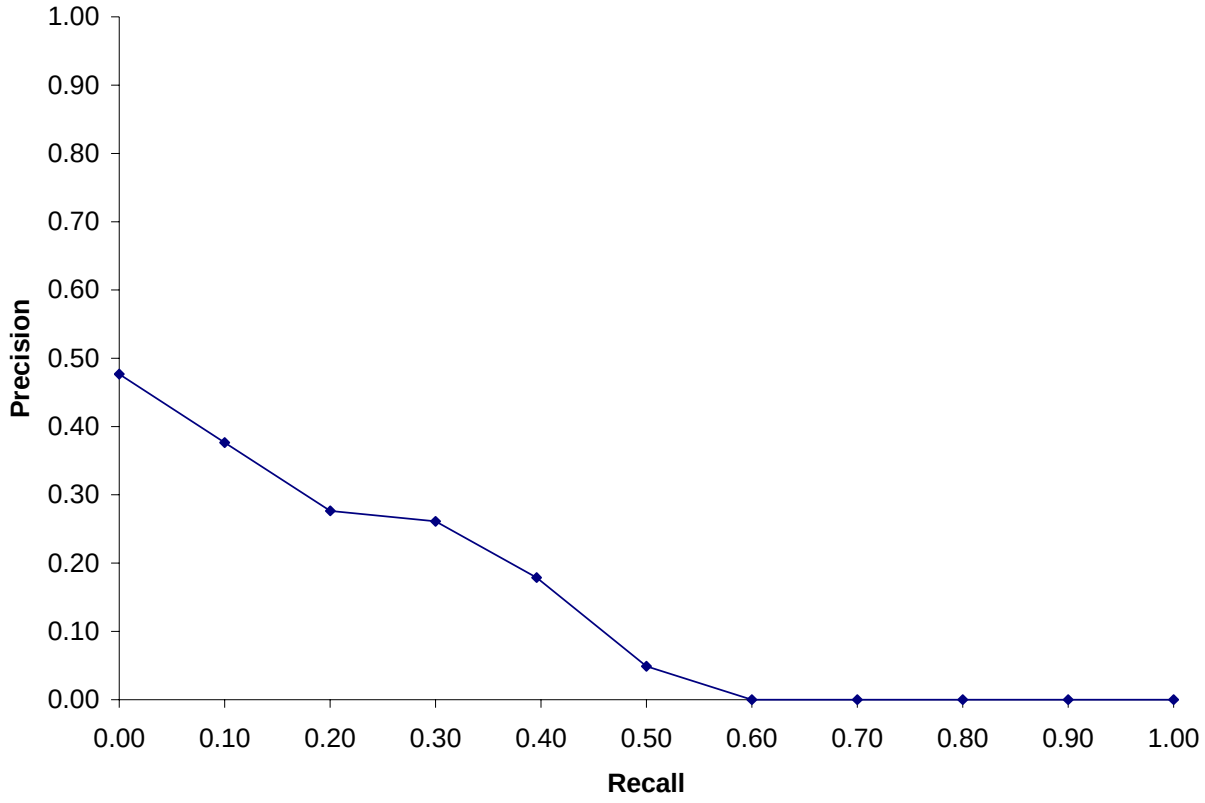

Figure 8. Precision of edit distance retrieval method at 11 levels of recall (mean precision = 14.7%). Only 50 names were retrieved by each method. For each method, some number of relevant names was not among the top 50 names retrieved. Thus, recall does not reach 100 percent for any method. Precision is, by definition, zero when a given recall level is not reached. See text for details.

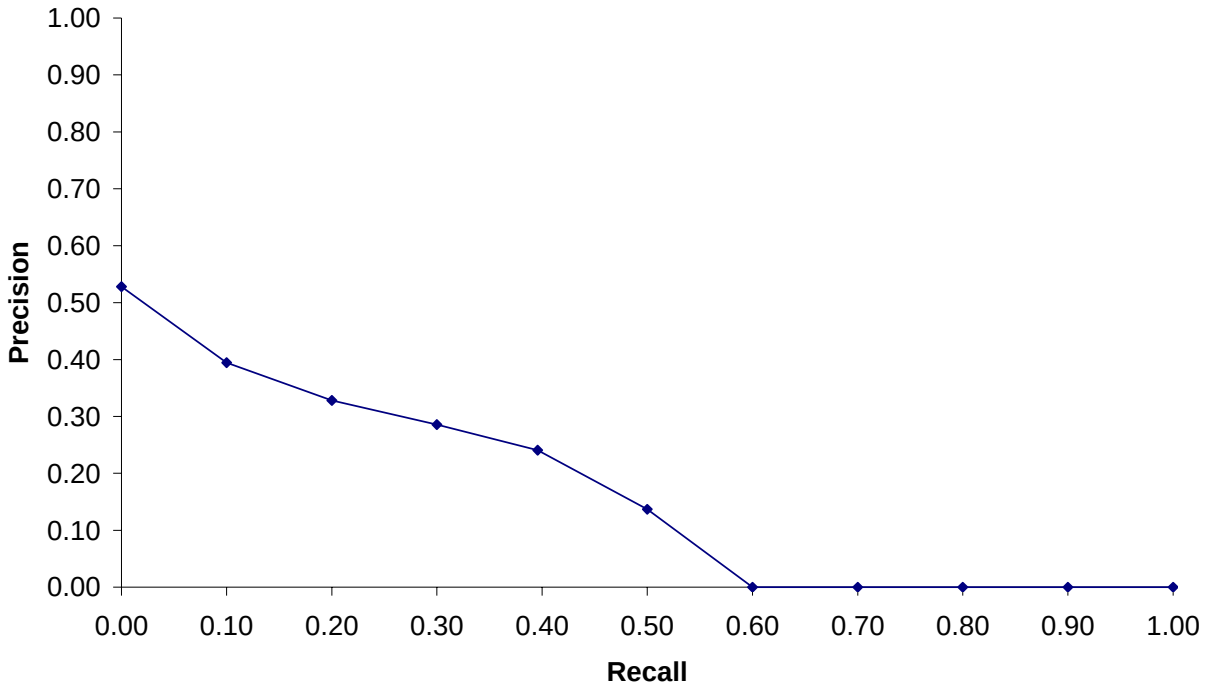

Figure 9. Precision of editex retrieval method at 11 levels of recall (mean precision = 17.4%).

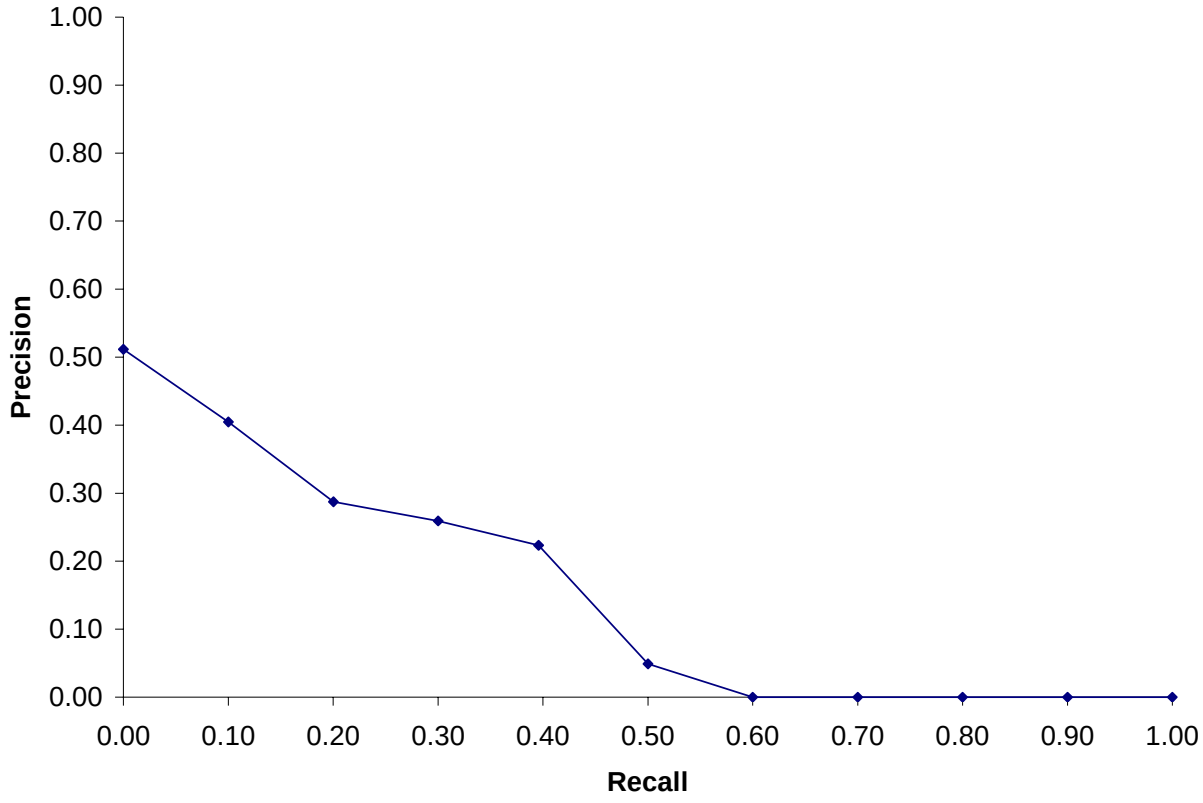

Figure 10. Precision of normalized edit distance retrieval method at 11 levels of recall (mean precision = 15.8%).

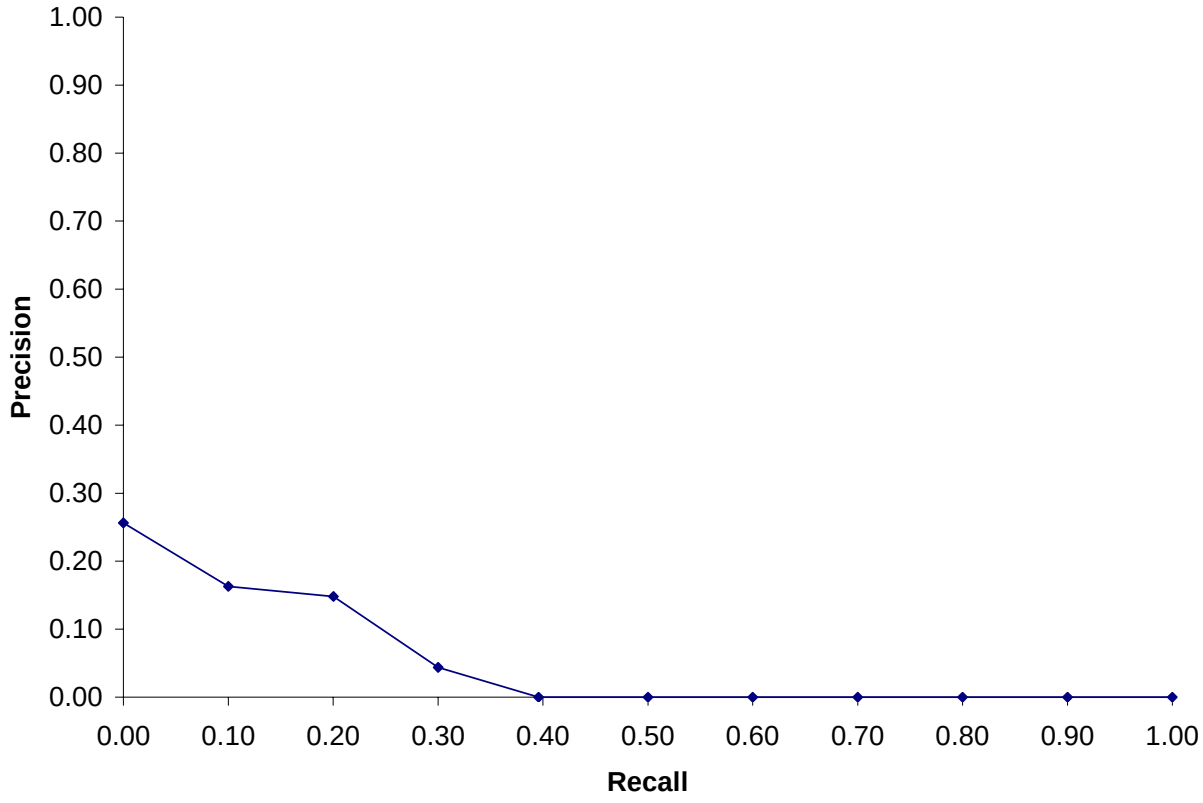

Figure 11. Precision of edit soundex retrieval method at 11 levels of recall (mean precision 5.5%).

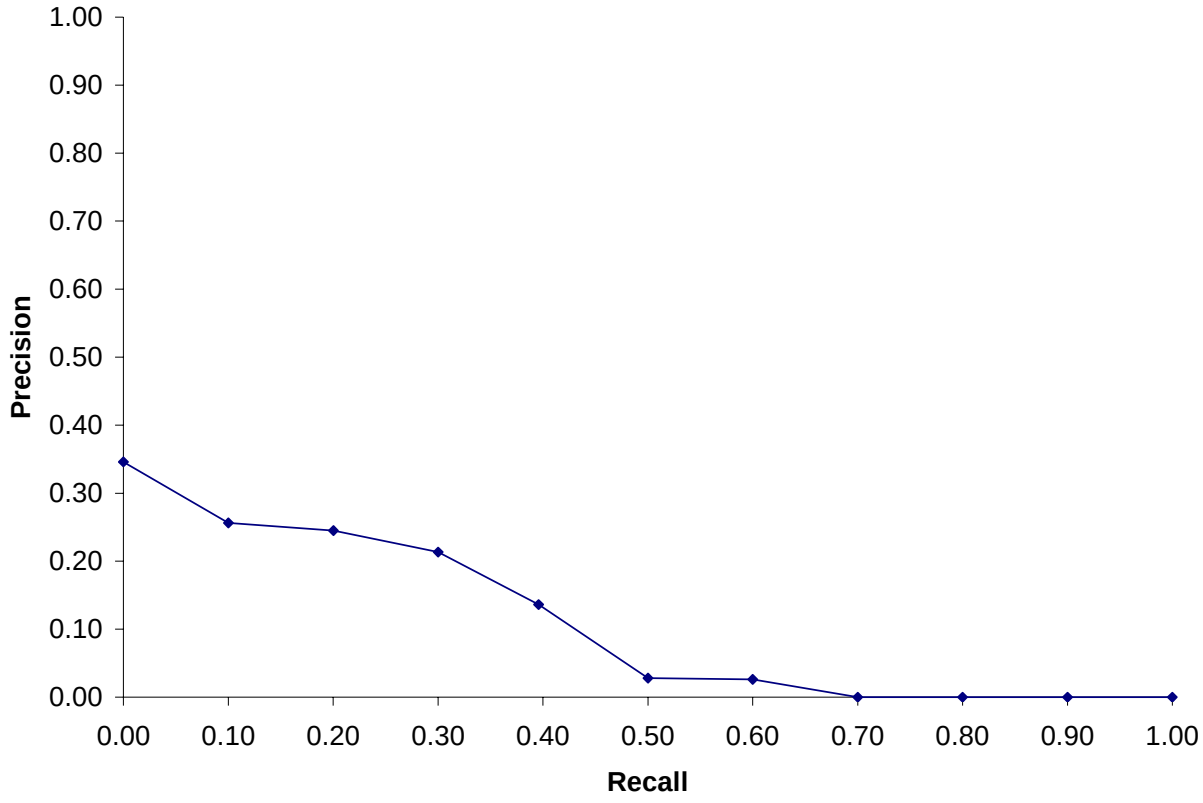

Figure 12. Precision of trigram-2b retrieval method at 11 levels of recall (mean precision = 11.4%).

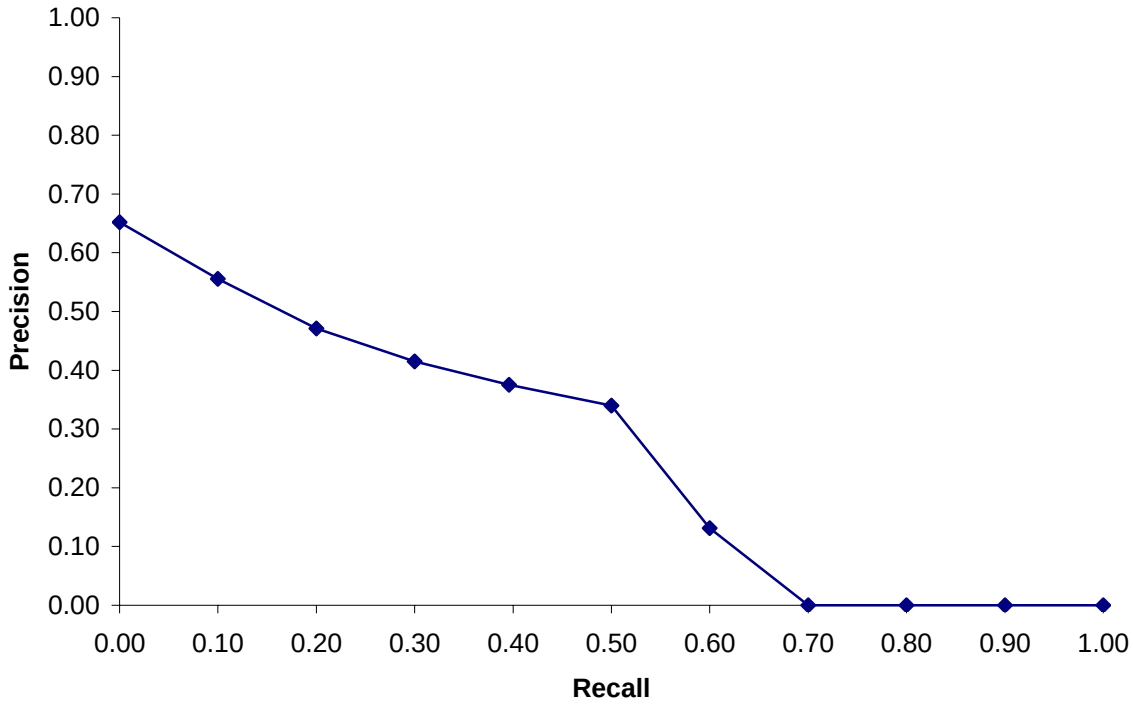

Figure 13. Precision of expert rating retrieval method at 11 levels of recall (mean precision = 26.7%).

Table 1. Correlation between multiple similarity and distance measures (N =1458).

	ED	Editex	Edit-soundex	NED	Trigram-2b	Mean Rating	Sq.Rt. Mean Rating
ED	-						
Editex	0.9530	-					
Edit-soundex	-0.1930	-0.0846	-				
NED	0.7476	0.6854	-0.2493	-			
Trigram-2b	-0.2006	-0.1422	0.1228	-0.2317	-		
Mean Rating	-0.4670	-0.4895	-0.0346	-0.4927	0.3054	-	
SqRt Mean Rating	-0.5223	-0.5396	0.0122	-0.5258	0.3483	0.9618	-

Note. ED = edit distance, NED = normalized edit distance, SqRt = square root. Correlations with absolute values greater than 0.042 are significantly greater than zero at alpha =0.05.

Table 2. Hierarchical regression analysis.

Step	Variable Added	Multiple R	R-Squared	Adj. R-Squared	F	P
1	Editex	0.5396	0.2912	0.2907	635.08	< 0.0000
2	NED	0.5806	0.3371	0.3362	105.95	< 0.0000
3	Trigram-2b	0.6277	0.3940	0.3928	144.21	< 0.0000
4	Edit Soundex	0.6372	0.4060	0.4045	30.20	< 0.0000

Note. F statistics for NED, Trigram-2b, and Edit Soundex are partial F statistics testing the change in R^2 associated with the addition of each new variable to the model.[43] F tests had 1 and 1546 degrees of freedom.

Table 3. Top 20 names retrieved by each retrieval method for query name *Avelox*®.

Retrieval Method						
Edit Distance	Editex	NED	Edit Soundex	Trigram-2b	Combined Model	Expert Ratings
Asulox	Allelix	Salvelox	A Plus	Aveco	Aveco	Azelex
Aveco	Asulox	Asulox	Ablc	Avert	Salvelox	Avadex
Azelex	Azelex	Aveco	Ables	Aved-M	Asulox	Avirax
Salvelox	Aveco	Azelex	Apo-Folic	Aveeno	Azelex	Avonex
Abelia	Aveeno	Kalvelax	Appebloc	Avenge	Aveeno	Salvelox
Abenol	Avirax	Marvelon	Applause	Aventyl	Avert	Aviax
Abtox	Salvelox	Adeflor	100 Plus	Avc	Avenge	Asulox
Aceon	Apollo	Alcloxa	2 Plus	Avo	Avirax	Lovenox
Adeflor	Apollon	Allelix	33 Plus	Salvelox	Aviax	Ava-Pox
Adexol	Avadex	Aloelax	4-Plex	Avenarius	Ava-pox	Abtox
Aero	Aviax	Ava-Pox	A L C	Avid	Allelix	Allertox
Aerx	Avonex	Camelot	A.P.L.	Avon	Avonex	Aloelax
Agrox	Abelia	Fieldox	A/B Otic	Avail	Avadex	Aloex
Alcloxa	Aero	Javelin	Abafilcon A	Avast	Aved-m	Kalvelax
Aldox	Aerx	Juvelon	Abas	Aviax	Juvelon	Opalux
Allelix	Agrox	Lovenox	Abbo-Pac	Avita	Javelin	Alcloxa
Aloe	Allertox	Pamelor	Abco	Aviva	Avo	Maalox
Aloelax	Aloelax	Pavalor	Abelcet	Asulox	Alcloxa	Agrox
Aloex	Aloex	Pavulon	Abelia	Avadex	Avon	Aldox
Alor	Amilon	Zavedos	Able	Availa	Aloelax	Ardrox

Table 4. Top 20 names retrieved by each retrieval method for query name *Curosurf*®.

Retrieval Method						
Edit Distance	Editex	NED	Edit Soundex	Trigram-2b	Combined Model	Expert Ratings
Atrosulf	Curasorb	Atrosulf	Creacarb	Curfew	Curasorb	Curasorb
Caropure	Curasore	Caropure	Curasorb	Curb	Curasore	Curasore
Curasorb	Atrosulf	Curasorb	Brush-Rhap	Curad	Exosurf	Curasilk
Curasore	Caropure	Curasore	Cal Group	Curay	Virosure	Exosurf
Exosurf	Exosurf	Exosurf	Calcarb 600	Curex	Urocur	Curasol
Proturf	Urocur	Proturf	Calsorb	C Cure	Atrosulf	Curisone
Urocur	Virosure	Urocur	Carbubarb	Curare	Curagard	Curasalt
Virosure	Curasalt	Virosure	Cardio-Herb	Curbit	Curasol	Infasurf
Auro-Dri	Curasilk	Luroscrub	Carlo Erba	Curity	Curasalt	Curafil
Aurora	Curasol	Nutrisure	Cascara Sagrada	Curves	Curasilk	Curecal
Biosure	Proturf	Auro-Dri	Char-Care	Cuprose	Caropure	Curagel
Cardura	Cerose	Aurora	Chiro.Care	Curafas	Curaderm	Curafas
Carmofur	Colostrx	Biosure	Chitosorb	Curafil	Curisone	Alusulf
Cerose	Copasure	Carboguard	Chromocarb	Curagel	Cuticura	Curagard
Colostrx	Curaderm	Cardiasure	Circu-Care	Curalan	Proturf	Curatek
Copasure	Curagard	Cardura	Citrus Grove	Curapid	Curare	Curaderm
Croesus	Curisone	Carmofur	Claysorb	Curasol	Cuprose	Curi-Strip
Cubosome	Cuticura	Cerose	Coli-Curb	Curatek	Cerose	Curvelle
Cuprose	Cytosar	Chronosule	Conce-Carb	Curecal	Luroscrub	Curretab
Curaderm	Farrowsure	Colostrx	Conserve	Curitas	Copasure	Curapid

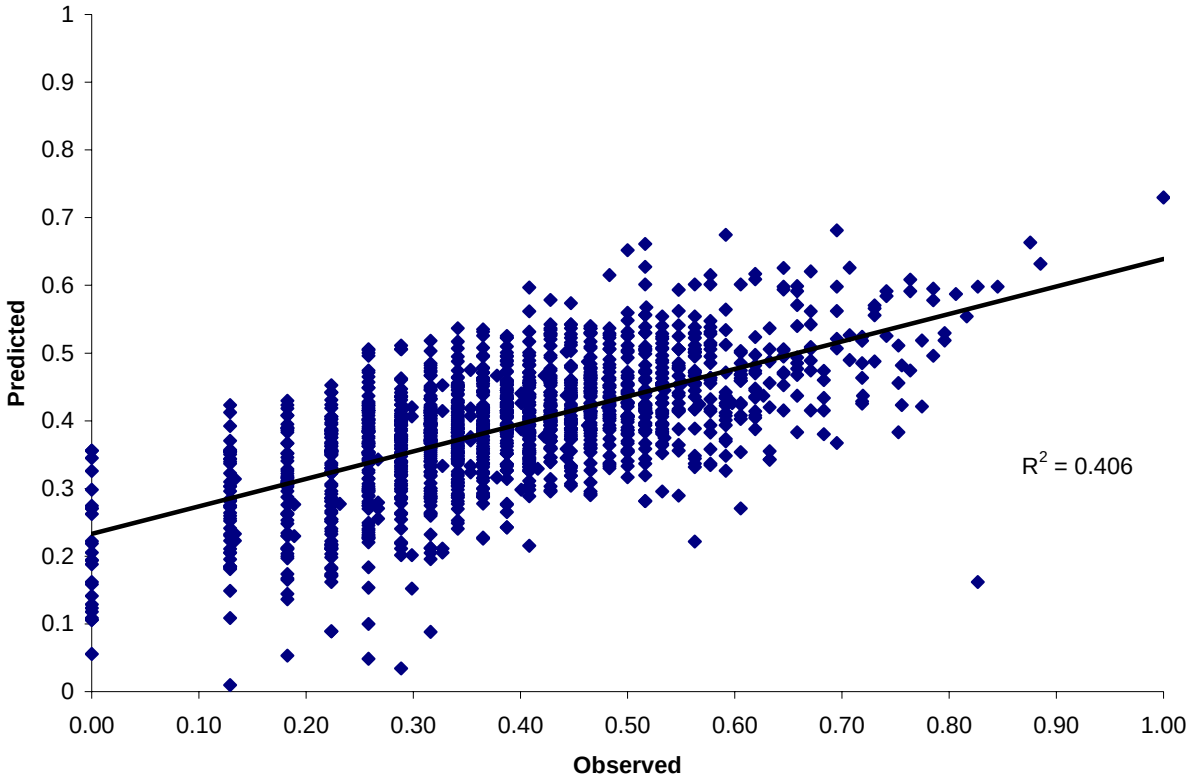

Figure 14. Fit between observed and predicted (square root) mean similarity ratings. Observed values based on mean ratings of 15 pharmacists. Predicted valued based on multiple linear regression model.