Predicting Look- and Sound-Alike Medication Errors

Bruce L. Lambert, Ph.D. is Assistant Professor, Department of Pharmacy Administration and Clinical Assistant Professor, Department of Pharmacy Practice, the University of Illinois at Chicago.

The assistance of Dan Boring, Bill Brewer, Stephanie Crawford, Diane Cousins, Gary Dell, George DiDomizio, Jean Gallagher, Paul Grussing, Prahlad Gupta, Keith Johnson, Kim Keller, David Lambert, Don Rucker, Gordon Schiff, and Donna Szymanski is gratefully acknowledged. The U. S. Pharmacopeia (USP) 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.

Address: 833 S. Wood St. (M/C 871), Chicago, IL 60612-7231

Phone: (312) 996-2411

Fax: (312) 996-0868

Email: lambertb@uic.edu

Abstract

Many medication errors are caused by look-alike and sound-alike medication names, yet few procedures exist to assure the safety of new drug nomenclature or to identify confusingly similar names from within existing databases. Until now, a theoretical model to explain look- and sound-alike errors has also been lacking, as has any quantitative measure of similarity between names. Drawing on recent research in psycholinguistics, this report outlines a theoretical model of medication name confusions. From this model, three automated, quantitative measures of orthographic (i.e., spelling) similarity were identified (bigram similarity, trigram similarity, and Levenshtein distance). The relationship between orthographic similarity and the likelihood of being involved in a medication error was examined.

Known look-alike and sound-alike error pairs (N=969) were identified from published reports. Control pairs (N=969) were selected at random from the general index of USP-DI, Volume I: Drug Information for the Health Care Professional. For each measure of similarity, the frequency distribution of similarity scores for error pairs was compared to the distribution of similarity scores for control pairs. Three parallel, unmatched case-control studies were conducted to discover whether similarity was a significant risk factor for medication errors. Finally, prognostic tests based on the three similarity measures were developed and evaluated.

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

Automated measures of medication name similarity can be used to form the basis of highly accurate, sensitive and specific prognostic tests for look- and sound-alike medication errors. These methods are comprehensive, theory-based, efficient, objective and reliable. Their use in the medication name approval process has the potential to reduce the incidence of look- and sound-alike medication errors.

Index terms: Drug information; Drug nomenclature; Look-alike; Medication errors; Prognostic test; Psycholinguistics; Safety; Similarity; Sound-alike.

In a society where nearly 2 billion prescriptions are filled annually in the community setting, and 30 million medication doses are administered daily in long term care settings, medication errors present a serious threat to patient welfare and a significant liability to health professionals and their insurers.[1] Many factors contribute to medication errors, but one factor consistently associated with errors is the existence of confusing pairs of medication names. 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 (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), are in need of valid and reliable methods that can assess the likelihood of look-and sound-alike medication errors before they occur.[4] If such methods could be developed, it might be possible to reduce the number of confusing names that reach the marketplace. The same methods could be used to identify confusing pairs within existing databases of medication names. Once identified, safeguards could be built into drug information systems to reduce the probability of confusion in clinical practice.[5]

Until now, practitioners have had to rely on voluntary reports of errors in order to identify potentially confusing pairs of medication names.[2,3,6,7] What's more, 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 confusion potential would ever be exhaustive enough to establish confidence in the resulting assessments.[8,9] To illustrate the difficulty of manual evaluation, consider that

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, $(N^2-N)/2$ unique pairs of names would need to be considered. With $\underline{N}$=15,000, a staggering 112,492,500 comparisons are needed! Clearly such a task is impossible without automated methods for evaluating confusion potential.

Fortunately, it is now possible to design objective, computer-based measures of orthographic (i.e., spelling, look-alike) and phonological (i.e., sound-alike) similarity. These similarity measures can then serve as the basis for predictions about confusion potential. Although lacking some of the features of manual evaluation by experts (e.g., no consideration of dose, indication, or physical appearance of the drug), the computerized measures of lexical (i.e., word-word) similarity are objective, reliable, and are based on well-established psycholinguistic theory. Automated methods for similarity-based searching of trademark databases are already available from commercial vendors, and they are used routinely by intellectual property attorneys and other interested parties.[10,11] The automated methods make it possible to do exhaustive comparisons between proposed new medication names and databases of existing names. Still, these methods require validation before considering their use 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 end-product of these experiments was a prognostic test that could be used to identify confusingly similar pairs among new and existing medication names.

Theoretical Background

To date, the literature on look-alike and sound-alike medication errors has been largely atheoretical. However, there is a vast literature in psycholinguistics and experimental psychology that describes the mental representations and cognitive processes that produce

lexical confusions in speaking, listening, reading, and writing, as well as in short- and long-term memory.[12-24] These psychological experiments have been conducted primarily on college undergraduates using common English words as stimulus materials. A similar pattern of results might be expected if participants were health professionals and the stimulus words were medication names. Hence, the measures of lexical similarity (i.e., similarity between words) developed and tested here are grounded in psycholinguistic theory.

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 (i.e., spelling), phonological (i.e., sound), syntactic (i.e., grammatical) and semantic (i.e., meaning) representations.[25] These representations are directly involved in the cognitive processes that allow healthy adults to access 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 the target word.[13,26] Errors in recognition memory 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 and/or phonologically similar words.[24,28] In summary, evidence from psycholinguistics clearly demonstrates that lexical similarity increases the likelihood of errors in recall, visual and auditory word recognition, and spontaneous speech.

Research Questions

This research was designed to identify and illuminate relationships between lexical similarity and lexical confusions in the domain of medication names. Answers were sought for several specific questions: What is the relationship between orthographic similarity and the likelihood of lexical confusion? Is lexical similarity a significant risk factor for lexical confusion? What measure of orthographic similarity will most accurately predict lexical confusions? The central practical concern was to discover whether automated measures of lexical similarity could serve as valid predictors of medication error potential, such that they might be used to facilitate the development of safer, more error-resistant drug nomenclature.

Hypotheses

This study evaluated the ability of three measures of orthographic similarity to distinguish between confusing and non-confusing pairs of medication names. Three specific hypotheses were tested. In the following hypotheses, the term "error pair" refers to a published pair of confusing medication names. Pairs of names selected at random from a large database of names are referred to as "control pairs." Details about the selection of error and control pairs are given in the Method section.

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

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

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

Research Design

Two related research designs were used to test the stated hypotheses. Both were 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 were drawn from published reports of medication errors and controls were drawn at random from all possible pairs of medication names.[29] To evaluate the usefulness of lexical similarity measures as the basis for a prognostic test that would predict 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]

Method

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, N=969 unique error pairs were identified. As an illustration, Table 1 shows 20 of the error pairs used in the study.

Table 1 about here.

Control pairs. Control pairs of medication names (N=969) were selected at random from an electronic version of the General Index to the U. S. Pharmacopeia's USP-DI, Volume I: Drug

Information for the Health Care Professional.[8] Table 2 shows 20 of the control pairs used in the current study. (A complete listing of the error and control pairs used is available from the author.)

Table 2 about here.

Measurement of Orthographic String Similarity

From the point of view of most computer programs, and for the purpose of assessing similarity, words are viewed 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] All measures of string similarity were computed using computer programs written by the author 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.

Both bigram and trigram methods are examples of n-gram measures of string similarity. For a given pair of medication names, n-gram measures were defined as follows.[35,36] First, the unique n-grams (i.e., n-letter sub-sequences) in each name were generated. Next, the number of n-grams common to the two names was tallied. Finally, n-gram string similarity (S) was defined by the Dice coefficient:

$$S = \frac{2C}{A + B}$$

where A was 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.[37] Both bigram (i.e. two letter sub-sequences) and trigram (i.e. three letter sub-sequences) measures were used in these investigations. Consider an example. For the drug Tylenol®, the unique

trigrams are tyl, yle, len, eno, and nol. For the drug atenolol, the unique trigrams are ate, ten, eno, nol, olo, and lol. The trigram string similarity between the names Tylenol® and atenolol, which share two trigrams in common (eno and nol), is (2*2)/(5+6)=.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] Levenshtein distance was defined as 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.

Analysis Plan

If lexical similarity was to distinguish between error pairs and control pairs, then the frequency distribution of similarities for error pairs should, at minimum, have been significantly different than the frequency distribution of similarities for control pairs (see Hypothesis 1). To test this hypothesis, orthographic similarity was calculated for N=969 error pairs and for N=969 control pairs, and a chi-square test of independence was performed on the resulting 2 X 11 contingency table (e.g., error/control X 11 similarity ranges). With overall N=1938, the chi-square contingency test at a=.01 had greater than 99% power to detect effect sizes larger than w=.20.[39]

To establish whether lexical similarity represented a significant risk factor for being involved in a look- or sound-alike medication error (see Hypothesis 2), an unmatched case control study was conducted with N=969 cases (i.e., error pairs) and N=969 controls.[40] Relative risk was estimated with the odds ratio computed from a 2 X 2 contingency table (e.g.,

exposure/no exposure X case/control). Significance of the odds ratio was tested via the chi-square statistic with 1 degree of freedom.[40] For n-gram methods, exposure was defined as similarity greater than or equal to 0.10 by the Dice coefficient. For Levenshtein distance, exposure was defined as distance less than or equal to 10 edit operations. Assuming an exposure rate of 1% among controls, a=.01, and N=969 for both cases 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 using automated measures of lexical similarity to predict which pairs of names were error pairs and which were controls. Receiver operator characteristic (ROC) curves were plotted for each measure of lexical similarity described above. The curves were plotted by systematically varying the cutoff value 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 false positives. Error pairs that failed to exceed the threshold were false negatives. Control pairs that failed to exceed the threshold were true negatives. Predictive accuracy was defined as (true positives + true negatives)/(true positives + true negatives + false positives + false negatives). Accuracy was measured and plotted at various thresholds. Sensitivity was defined as the true positives/(true positives + false negatives). Specificity was defined as true negatives/(true negatives + false positives).[32] At each cutoff value, sensitivity (i.e., the true positive rate) was plotted against 1-specificity (i.e., the false positive rate). The resulting ROC curves were used to select the optimal cutoff value for each test. With N=969 cases and controls, it was possible to estimate 99% confidence intervals for sensitivity and specificity of ±5%.[29]

Positive predictive value was defined as the probability that a pair was an error pair, given a positive test. Positive predictive value 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 was defined as the probability that a pair was a control pair, given a negative test. Negative predictive value 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 at various prior probabilities at the optimal cutoff for each measure of similarity.[29,32]

Results

Frequency distributions. The frequency distributions of similarity scores for error pairs and control pairs, for each measure of similarity (i.e., bigram, trigram, and Levenshtein distance), are given in Figures 1-3. 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), simple visual inspection reveals that the similarity scores for error pairs were skewed to the high end of the scale, while the scores for control pairs were skewed to the low end of the scale. For the Levenshtein distance measure (Figure 3), scores for error pairs were skewed to the low end of the scale, while scores for control pairs were skewed to the high end of the scale. In each case, the chi-square test of independence was highly significant: for bigram string similarity, chi-square=1281.08, df=10, p<.00000; for trigram string similarity, chi-square= 1000.34, df=10, p<.00000; for Levenshtein distance, chi-square= 1573.02, df=10, p<.00000.

Figures 1-3 about here.

Relative risk. For bigram and trigram string similarity measures, exposure was defined as similarity greater than or equal to 0.1. For Levenshtein distance, exposure was defined as

distance less than or equal to 10. These cutoffs were chosen somewhat arbitrarily, although they were intended to be conservative values (i.e., relatively low similarity, relatively large distance). Contingency tables for bigram, trigram and Levenshtein distance at these exposure levels are given in Tables 3-5 respectively. For each measure of similarity, relative risk was approximated by the odds ratio.[40] For bigram string similarity, the odds ratio was substantially greater than 1 (OR=34.01, 95%CI=(25.16, 45.98), chi-square=797.06, df=1, p<.00000), as it was for trigram similarity (OR=67.27, 95%CI=(47.04, 96.19), chi-square=974.48, df=1, p<.00000) and Levenshtein distance (OR=275.81, 95%CI=(145.52, 522.76), chi-square=1105.33, df=1, p<.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.

Tables 3-5 about here.

Predictive accuracy. The predictive accuracy of each measure at various thresholds is given in Figures 4-6. The test with the highest overall accuracy (94%) was based on Levenshtein distance with a threshold of distance less than or equal to 6 (see 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 (see Figure 5). The trigram test with the highest accuracy classified 84% of the pairs correctly and was based on a threshold of similarity ≥ 0.1 (see Figure 5).

Figures 4-6 about here.

Sensitivity and specificity: ROC Curves. Figures 7-9 show the receiver operator characteristic (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 or prognostic test. The best point is at the "shoulder" of the ROC curve, 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 (see Figure 7). For the bigram measure, the best threshold was similarity ≥ 0.2, where the test achieved 58.6% sensitivity and 99% specificity (see Figure 8). For the Levenshtein distance measure, the best threshold was distance ≤ 5, where the test achieved 84% sensitivity and 98.8% specificity (see Figure 9).

Figures 7-9 about here.

Positive and negative predictive value. Estimates of positive and negative predictive value help clinicians interpret the results of diagnostic and prognostic tests. Positive predictive value is the probability that a pair was an error pair, given a positive test. Negative predictive value is the probability that a pair was a control pair, given a negative test. Both positive and negative predictive value were dependent on the prior probability of the event being predicted. In this context, 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 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.

Figures 10-15 about here.

Discussion

Usefulness of the prognostic tests. In order for a prognostic test to be viable, three conditions, in increasing order of difficulty, must be met. First, cases and controls must be distributed differently with respect to the prognostic measure. Second, the prognostic measure must be a significant risk factor for the condition being predicted. Third, 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 above. 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 being an error. In at least one case (i.e., Levenshtein distance) it was possible to construct a test with quite high values of sensitivity and specificity and with 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. Using this cutoff in screening tests, one would expect to correctly identify 84% of all true error pairs and 98.8% of all non-error pairs. Generally speaking, the rarer a condition, the more specific a test must be for it to be practically useful.[29] If a test for a rare condition is not sufficiently specific, an unacceptably high number of false positives will be reported. Since medication errors were assumed to be rare events, with an incidence of less than 1%, a more specific test was identified as best in this report. However, in the final analysis, the precise placement of a cutoff value depends on a careful consideration of the societal costs of false positives versus false negatives.

Regulatory implications. Given the results reported above, regulatory agencies that approve new drug nomenclature would be well-advised to explore the integration of automated similarity tests 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. If the similarity between a candidate name and an existing name exceeded some established threshold (e.g., Levenshtein distance ≤5), the candidate name would be refused, or perhaps contingently accepted with appropriate precautions. If the similarity between the candidate name and existing names never exceeded the threshold, then the name would be approved. Analogously, every pair of names in the existing pharmacopeia could be screened in an effort to identify pairs that are confusingly similar. 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 as part of the name approval process, discussion is required about how to set the threshold. This debate should focus on the costs associated with false positives and false negatives. In this context, a false positive would result in prohibiting a name that was not, in fact, likely to cause an error. The cost here would primarily be an opportunity cost for the company desiring a particular name. Not being able to use the name would mean forgoing the profits associated with one name while being forced to use another, ostensibly less desirable, though potentially safer, name. A false negative, in this context, would result in approving a name for a medication that was, in fact, likely to contribute to look- or sound-alike errors. The costs here would be the human costs associated with patient suffering and the monetary costs associated with avoidable hospitalizations as well as malpractice and/or liability litigation. Thus, the setting of a cutoff point becomes an important policy question deserving of public discussion and debate.

Conflicting concerns in medication naming. The effort to design error-resistant drug nomenclature is complicated by the need for new medication names to simultaneously satisfy commercial, professional, and safety-related concerns. New names must be reasonably safe and

free from confusion, but the terms must also be meaningful and memorable to physicians, nurses, pharmacists, and patients. Names must be distinct, but medications that share an indication, mechanism of action or chemical constituent are often intentionally given the same prefix or suffix.[9] Slight variants of certain medications are often intentionally 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 surface when it comes to developing error-resistant nomenclature. Representatives of industry, the 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 in large part 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 in psychology, human factors, engineering, or computer science. With respect to look- and sound-alike medication errors, success will depend on having an accurate and comprehensive understanding of the mental representations and processes that underlie language production and comprehension.

Beyond medication errors. The general principles that undergird the tests developed here are applicable to other domains where confusing nomenclature causes errors. For example, similar techniques should prove useful in identifying confusing pairs of medical procedure names and names of diagnoses. Even more generally, the procedures outlined here could be used to estimate the likelihood of trademark confusion in any domain, not just medications. 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 quite small compared to the total possible number of confusing pairs, and the published lists of error pairs were gathered by ad hoc methods that may have reflected selection biases. The cases (i.e., error pairs) were drawn, in effect, from referral centers (e.g., the Medication Error Reporting Program, the Institute for Safe Medication Practices, FDA MedWatch). As such, the cases were likely to be more vividly similar than an unreported pair of confusing names. The sensitivity of a prognostic tests tends to be exaggerated under these circumstances.[29] At the same time, though, control pairs were selected at random and were not known to be "disease free." It is possible that some of the control pairs used here could have been involved in unreported medication errors. This fact would tend to cause specificity to be underestimated.

Even though a prognostic test was being evaluated, the research design was retrospective. Rather than saying test results accurately predicted errors, it would be more accurate to say 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, due to well-known problems in error reporting and error surveillance (i.e., low incidence rates and underreporting), such an investigation would be difficult to carry out.

The cases and controls were not matched for frequency of prescribing, an important variable to be considered 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 here ignored similarity in product labeling and packaging, dosing, indication, and physical appearance of the dosage form.

International variations in spelling were ignored. Finally, the experiments lacked any concept of error severity; all errors were viewed as equally severe.

The phonological dimension of similarity was captured only indirectly by the orthographic measure used. The majority of 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, where pronunciation is not a simple function of spelling, would not be captured by the present model. Relatedly, mispronunciations and regional variations in pronunciation could not be captured by the current measures. These investigations did not draw clear distinctions between perceptual modalities (e.g., visual or auditory) or communications media (e.g., handwriting, typewriting, fax, computer monitor, telephone, face-to-face dialogue, etc.). Nor were distinctions drawn between recall, recognition, 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 were ignored, as was the number of syllables in each medication name, although both of these features are known to contribute to confusion potential. Some abbreviations may actually mitigate errors. Unfortunately, research on this question constitutes a major unknown. Research on a model of phonological similarity is currently being undertaken in order to address these shortcomings.

Perhaps the most significant limitation to keep in mind is that these experiments treated medication errors as if they occurred in an abstract, decontextualized, psychological realm. In reality, medication errors occur within complex and dynamic systems of activity within equally complex physical layouts and organizational environments. Although orthographic and phonolgical similarity surely contribute to errors, a satisfactory model of medication errors

must encompass an understanding of 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, computer-based measures of orthographic (i.e., spelling) similarity can form the basis for a highly accurate, sensitive and specific tests of look- and sound-alike error potential. These tests are comprehensive, theory-based, reasonably inexpensive to conduct, objective, and reliable. The tests lack certain features of expert evaluation of error potential, especially with respect to potential similarity in indication, packaging, and dosing. The tests presently lack methods for directly assessing phonological (i.e., sound) similarity. Despite these limitations, the tests described here are already more objective, comprehensive, and reliable than current methods used to assure the safety of drug nomenclature. Therefore, in the interest of public health, it seems that regulatory agencies responsible for approving new drug nomenclature should move quickly to incorporate these or similar procedures into routine name-approval processes.

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; 52:374-379.

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

3. U. S. Pharmacopeia. USP Quality Review. Rockville, MD: U. S. Pharmacopeia, 1995; No. 49.

4. Boring D, Homonnay-Weikel AM, Cohen M et al. Avoiding trademark trouble at FDA. *Pharmaceutical Executive*. 1996; 16:80-8.

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; 32-37.

8. U. S. Pharmacopeia. USP DI, Vol. I: Drug information for the health care professional. Rockville, MD: U. S. Pharmacopeia; 1995.

9. U. S. Pharmacopeia. USP dictionary of USAN and international drug names. Rockville, MD: U. S. Pharmacopeia; 1996.

10. Imsmarq Home Page. [resource on World Wide Web]. URL:http://www.denpat.lu/denpat/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, Wilkinson R, eds. Zurich, Switzerland: Association for Computing Machinery; 1996:166-172.

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: Lawrence Erlbaum Associates; 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; 96:523-568.
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: John Wiley & Sons; 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; 1990:122-147.

22. Frauenfelder UH. Computational models of spoken word recognition. In: Computational psycholinguistics. Dijkstra T, de Smedt K, eds. Bristol, PA: Taylor & Francis; 1996:115-138.

23. Grainger J, Dijkstra T. Visual word recognition: Models and experiments. In: Computational psycholinguistics. Dijkstra T, de Smedt K, eds. Bristol, PA: Taylor & Francis; 1996:139-165.

24. Dell GS, Juliano C. Computational models of phonological encoding. In: Computational psycholinguistics. Dijkstra T, de Smedt K, eds. Bristol, PA: Taylor & Francis; 1996:328-359.

25. Emmorey KD, Fromkin VA. The mental lexicon. In: Linguistics, The Cambridge survey III: Language: Psychological and biological aspects. Newmeyer F, ed. Cambridge: Cambridge University Press; 1988:124-149.

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. *Journal of Memory and Language*. 1989; 28:462-485.

29. Hulley SB, Cummings SR, eds. Designing clinical research. Baltimore, MD: Williams and Wilkins; 1988.

30. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: A basic science for clinical medicine. Boston, MA: Little Brown; 1985.

31. Lilienfeld DE, Stolley PD. Foundations of epidemiology. New York: Oxford University Press; 1994.

32. Fletcher R, Fletcher SW, Wagner EH. Clinical epidemiology: The essentials. Baltimore, MD: Williams & Wilkins; 1996.

33. U. S. Food and Drug Administration. Medication errors. *FDA Medical Bulletin*. 1996; 26:3.

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; 1992:131-160.

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. *Journal of the ACM*. 1974; 21:168-173.

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

40. Schlesselman JJ. Case control studies. New York: Oxford University Press; 1982.

41. Leape L. Preventing adverse drug events. *Am J Health-Syst Pharm*. 1995; 52:379-382.

Table 1.

Example of twenty error pairs selected at random

dipyridamole	disopyramide
chlor-trimeton	chloromycetin
dimetane	dimetapp
cytotec	cytoxan
oracin	orasone
atarax	marax
docusate	doxinate
actacel	actimmune
enflurane	isoflurane
auralgan	ophthalgan
imfeon	intropin
voltaren	vontrol
pralidoxime	pyridoxine
citracal	citrucel
lopid	slo-bid
diphenatol	diphenidol
accubron	accutane
phos-flur	phoslo
percocet	percodan
catapres	combipres

Note. This subset was randomly selected from the full list of 969 error pairs. All names are printed in lower case to emphasize that comparisons were case-insensitive.

Table 2.

Example of twenty control pairs selected at random

novo-tamoxifen	carbinoxamine compound-drops
immun-aid	ascomp with codeine no.3
alaxin	beta-hc
aminophylline	technetium tc 99m hsa
sodium dichloroacetate	klerist-d
propoxyphene hydrochloride, aspirin, and caffeine	slow fe
promote with fiber	diarrest
potassium gluconate	phentermine hydrochloride
nitropress	anectine
myciguent	velosulin human
k-phos neutral	lipisorb
metaprel	amrinone lactate
6-mp	canesten
apo-metoclop	contac jr. children's cold medicine
preemie sma 20	minocycline hydrochloride
dalgan	atasol-8
diphen cough	citrolith
anti-thymocyte serum	infumorph
poliovirus vaccine inactivated enhanced potency	tarpaste

noxzema anti-acne pads regular strength	allerest children's

Note. This subset was randomly selected from the full list of 969 control pairs. All names are printed in lower case to emphasize that comparisons were case-insensitive.

Table 3.

Frequency of exposure to bigram similarity ≥ 0.1 for an unmatched sample of n=969 cases and n=969 controls

	Error	Control	Total
Similarity ≥ 0.1			
Yes	913 (a)	314 (b)	1227
No	56 (c)	655 (d)	711
Total	969	969	1938

Relative risk≈odds-ratio=(a*d)/(b*c)[29]

Table 4.

Frequency of exposure to trigram similarity ≥ 0.1 for an unmatched sample of n=969 cases and n=969 controls

	Error	Control	Total
Similarity ≥ 0.1			
Yes	705 (a)	37 (b)	742
No	264 (c)	932 (d)	1196
Total	969	969	1938

Relative risk≈odds-ratio=(a*d)/(b*c)[29]

Table 5.

Frequency of exposure to Levenshtein distance ≤ 10 for an unmatched sample of n=969 error pairs and n=969 control pairs

	Error	Control	Total
Distance ≤ 10			
Yes	959 (a)	250 (b)	1227
No	10 (c)	719 (d)	711
Total	969	969	1938

Relative risk≈odds-ratio=(a*d)/(b*c)[29]

Figure Captions

Figure 1. Histogram of bigram string similarities for error pairs (N=969) and control pairs (N=969). Vertical axis is on logarithmic scale. Light bars represent errors. Dark bars represent controls. 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 error pairs (N=969) and control pairs (N=969). Vertical axis is on logarithmic scale. Light bars represent errors. Dark bars represent controls. 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 error pairs (N=969) and control pairs (N=969). Vertical axis is on logarithmic scale. Light bars represent errors. Dark bars represent controls. 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 is a distance measure, not a similarity measure, so errors pairs are skewed to the low end, and controls are skewed to the high end of the distance scale.

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.

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

Figure 7. ROC curve for bigram string similarity in the prediction of look- and sound-alike medication errors. The bigram similarity values of chosen cutoff points are in parentheses.

Figure 8. ROC curve for trigram string similarity in the prediction of look- and sound-alike medication errors. The trigram similarity values of chosen cutoff points are in parentheses.

Figure 9. ROC curve for edit distance in the prediction of look- and sound-alike medication errors. The Levenshtein distance values of chosen cutoff points are in parentheses.

Figure 10. Positive predictive value of prognostic test based on bigram string similarity with 73% sensitivity and 99% specificity. The cutoff for this test was similarity greater than or equal to 0.3. Positive predictive value is the probability that a pair is an error pair, given a positive test.

Figure 11. Positive predictive value of prognostic test based on trigram string similarity with 59% sensitivity and 99% specificity. The cutoff for this test was similarity greater than or equal to 0.2. Positive predictive value is the probability that a pair is an error pair, given a positive test.

Figure 12. Positive predictive value of prognostic test based on Levenshtein distance with 84% sensitivity and 98.8% specificity. The cutoff for this test was edit distance less than or equal to 5. Positive predictive value is the probability that a pair is an error pair, given a positive test.

Figure 13. Negative predictive value of prognostic test based on bigram string similarity with 73% sensitivity and 99% specificity. The cutoff for this test was similarity greater than or equal to 0.3. Negative predictive value is the probability that a pair is a control pair, given a negative test.

Figure 14. Negative predictive value of prognostic test based on trigram string similarity with 59% sensitivity and 99% specificity. The cutoff for this test was similarity greater than or equal to 0.2. Negative predictive value is the probability that a pair is a control pair, given a negative test.

Figure 15. Negative predictive value of prognostic test based on edit distance with 84% sensitivity and 98.8% specificity. The cutoff for this test was edit distance less than or equal to 5. Negative predictive value is the probability that a pair is a control pair, given a negative test.

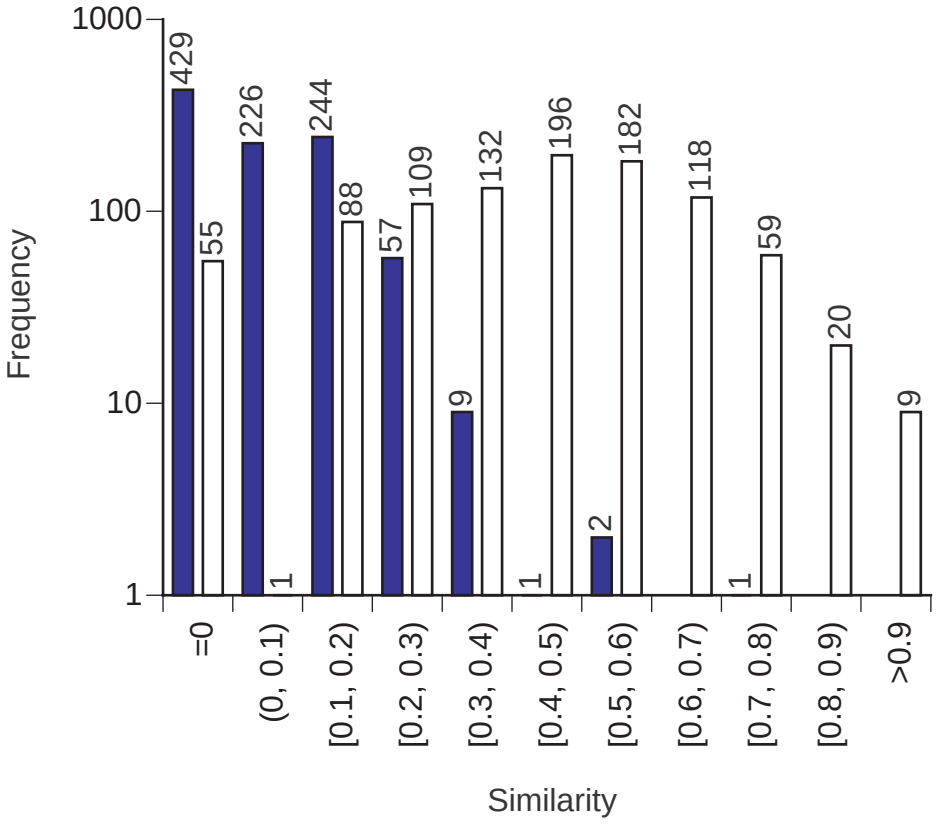
1000
100
10
1
Frequency
429
55
226
1
244
88
57
109
9
132
1
196
2
182
118
1
59
20
9
=0
(0, 0.1)
[0.1, 0.2)
[0.2, 0.3)
[0.3, 0.4)
[0.4, 0.5)
[0.5, 0.6)
[0.6, 0.7)
[0.7, 0.8)
[0.8, 0.9)
>0.9
Similarity

Figure 1

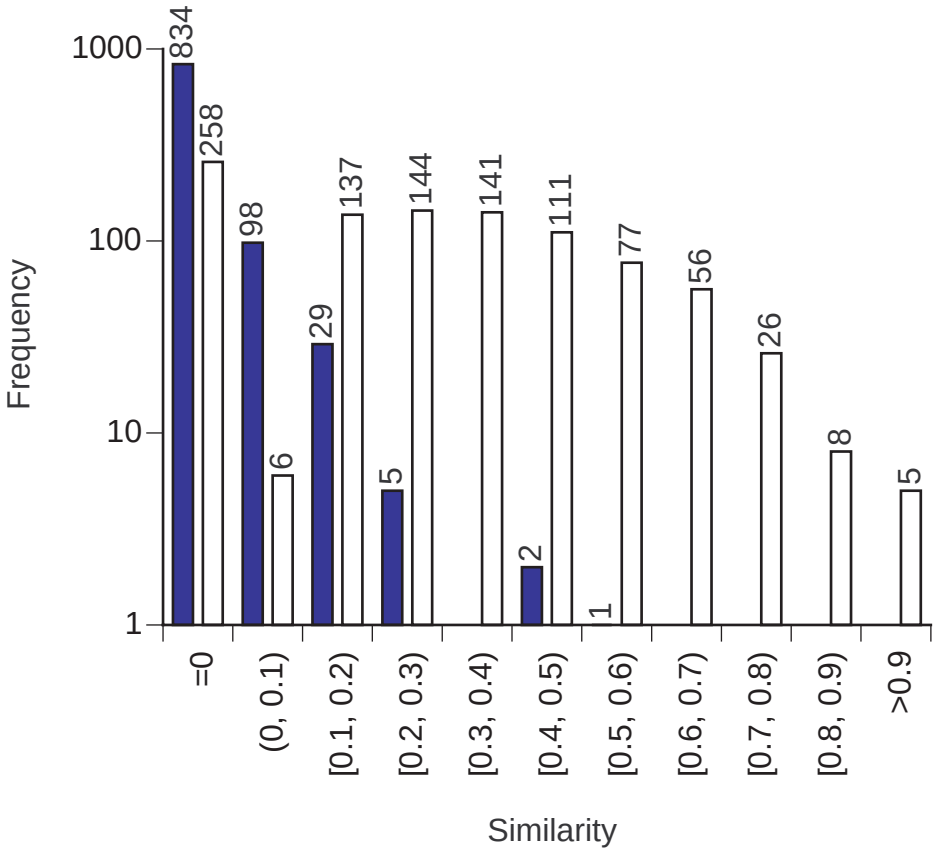
Frequency
1000
100
10
1
834
258
98
6
29
137
5
144
141
2
111
1
77
56
26
8
5
=0
(0, 0.1)
[0.1, 0.2)
[0.2, 0.3)
[0.3, 0.4)
[0.4, 0.5)
[0.5, 0.6)
[0.6, 0.7)
[0.7, 0.8)
[0.8, 0.9)
>0.9
Similarity

Figure 2

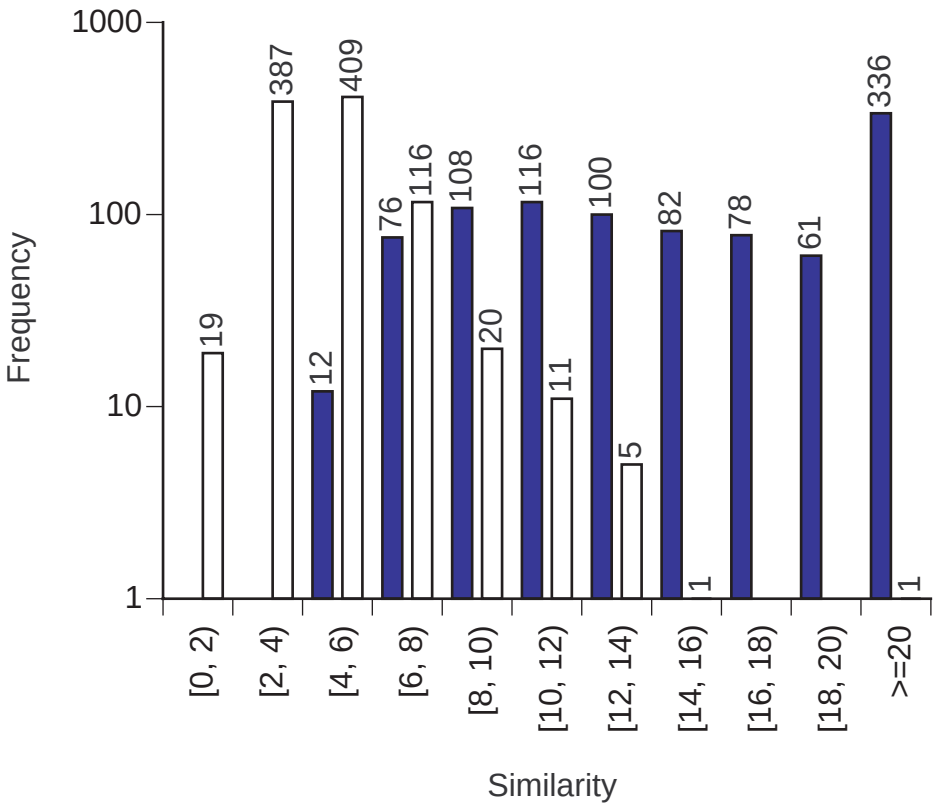
1000
100
10
1
Frequency
19
387
12
409
76
116
108
20
116
11
100
5
82
1
78
61
336
1
[0, 2)
[2, 4)
[4, 6)
[6, 8)
[8, 10)
[10, 12)
[12, 14)
[14, 16)
[16, 18)
[18, 20)
>=20
Similarity

Figure 3

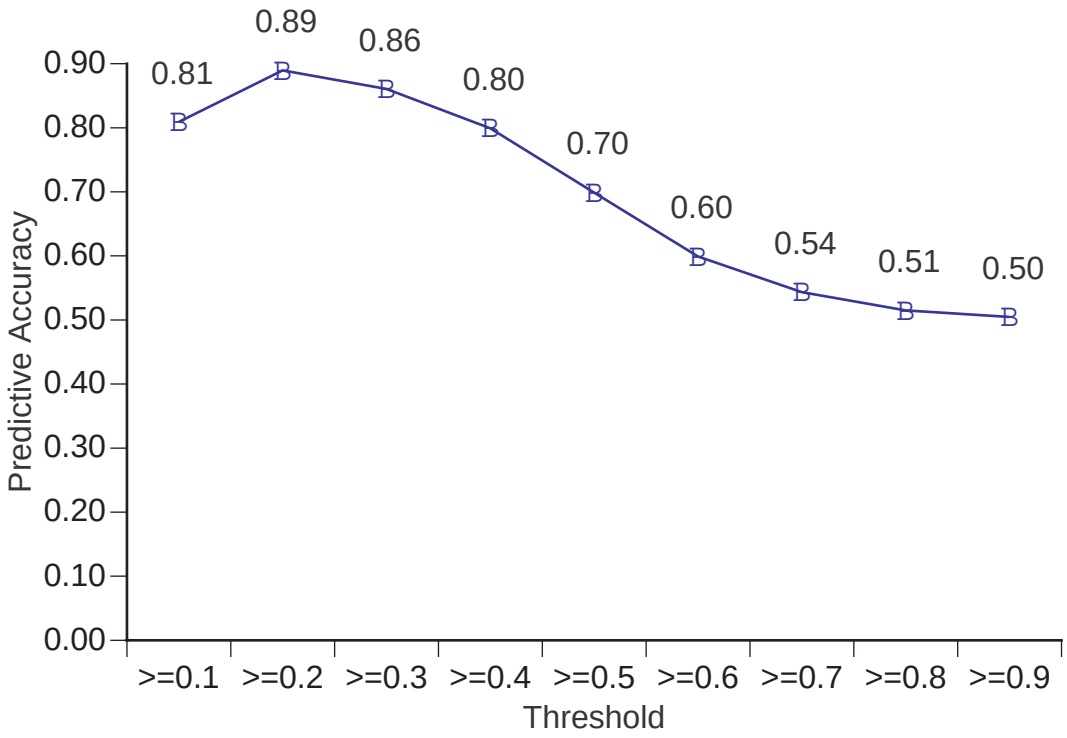
0.89
0.86
0.81
0.80
0.70
0.60
0.54
0.51
0.50
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Predictive Accuracy
>=0.1
>=0.2
>=0.3
>=0.4
>=0.5
>=0.6
>=0.7
>=0.8
>=0.9
Threshold

Figure 4

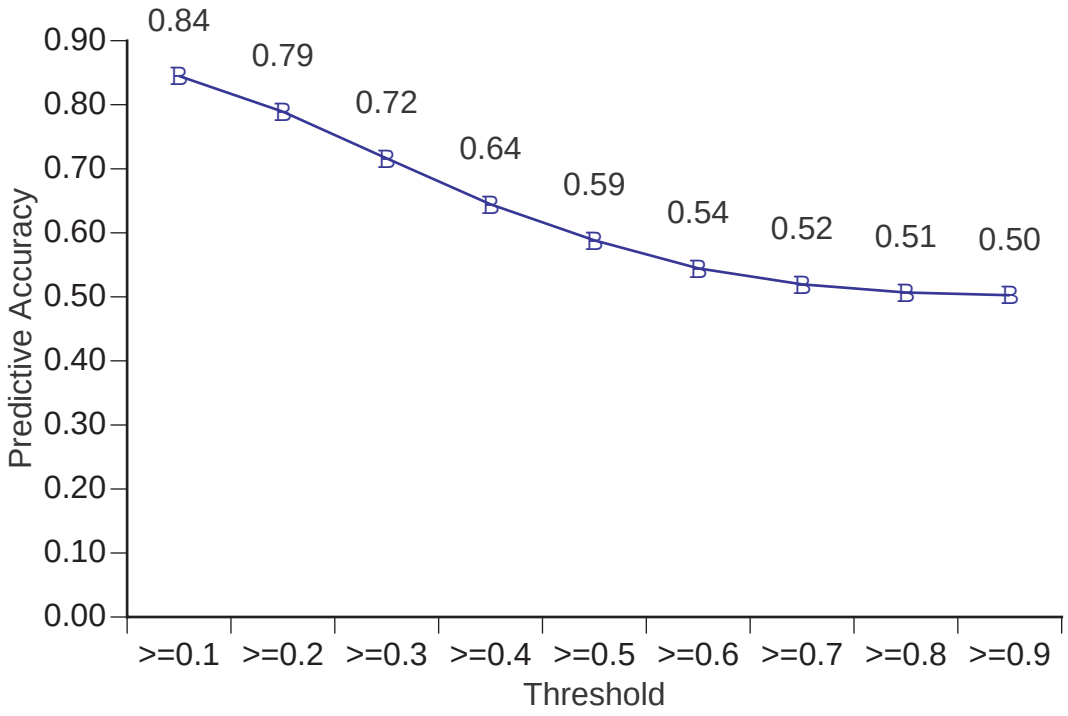
0.84
0.79
0.72
0.64
0.59
0.54
0.52
0.51
0.50
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Predictive Accuracy
>=0.1
>=0.2
>=0.3
>=0.4
>=0.5
>=0.6
>=0.7
>=0.8
>=0.9
Threshold

Figure 5

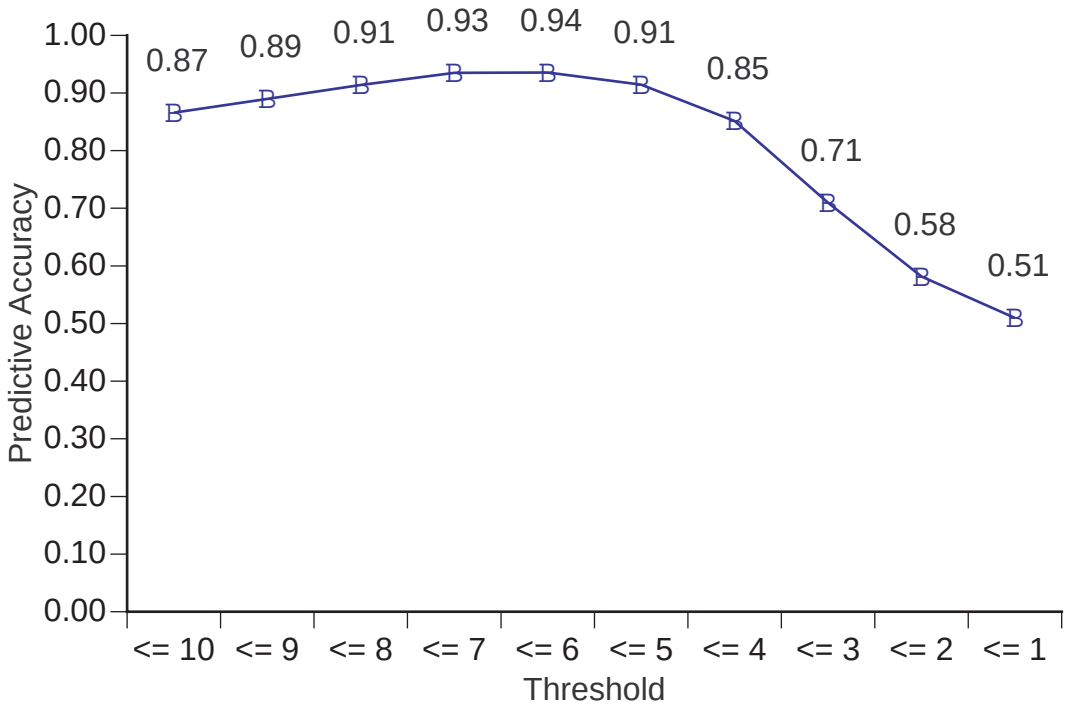
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Predictive Accuracy
0.87
0.89
0.91
0.93
0.94
0.91
0.85
0.71
0.58
0.51
<= 10
<= 9
<= 8
<= 7
<= 6
<= 5
<= 4
<= 3
<= 2
<= 1
Threshold

Figure 6

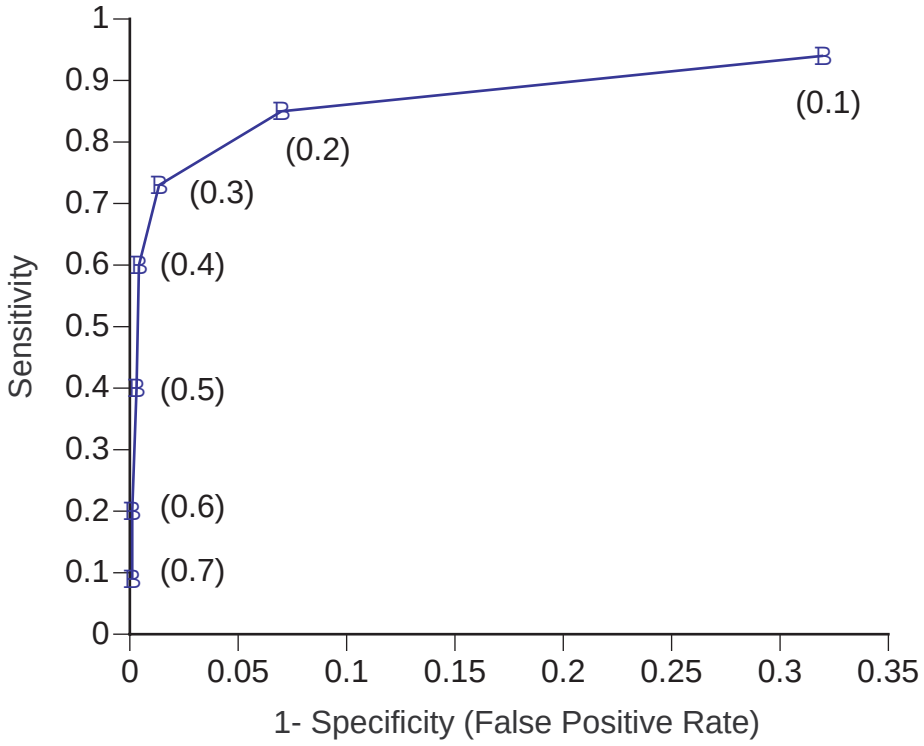
Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
1- Specificity (False Positive Rate)
(0.1)
(0.2)
(0.3)
(0.4)
(0.5)
(0.6)
(0.7)

Figure 7

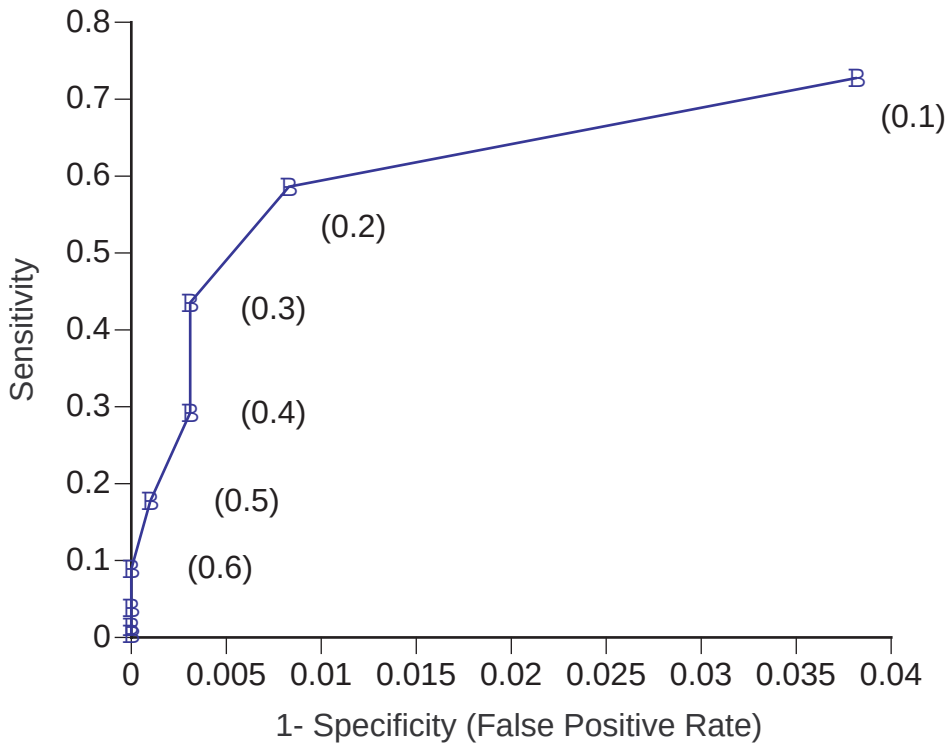
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Sensitivity
(0.1)
(0.2)
(0.3)
(0.4)
(0.5)
(0.6)
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
1- Specificity (False Positive Rate)

Figure 8

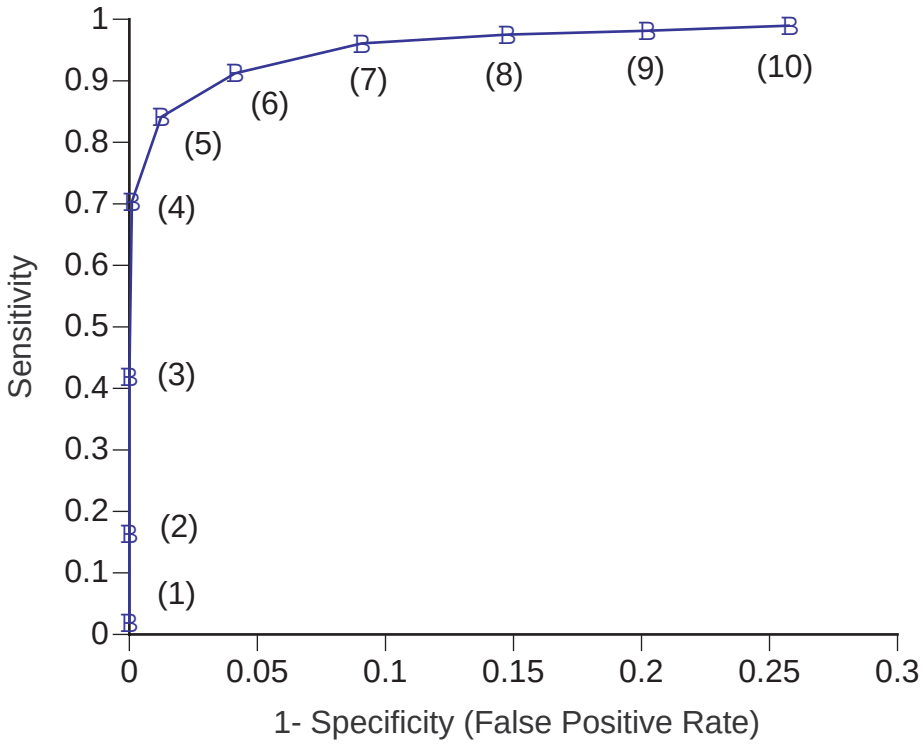
Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
1- Specificity (False Positive Rate)
(1)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
(9)
(10)

Figure 9

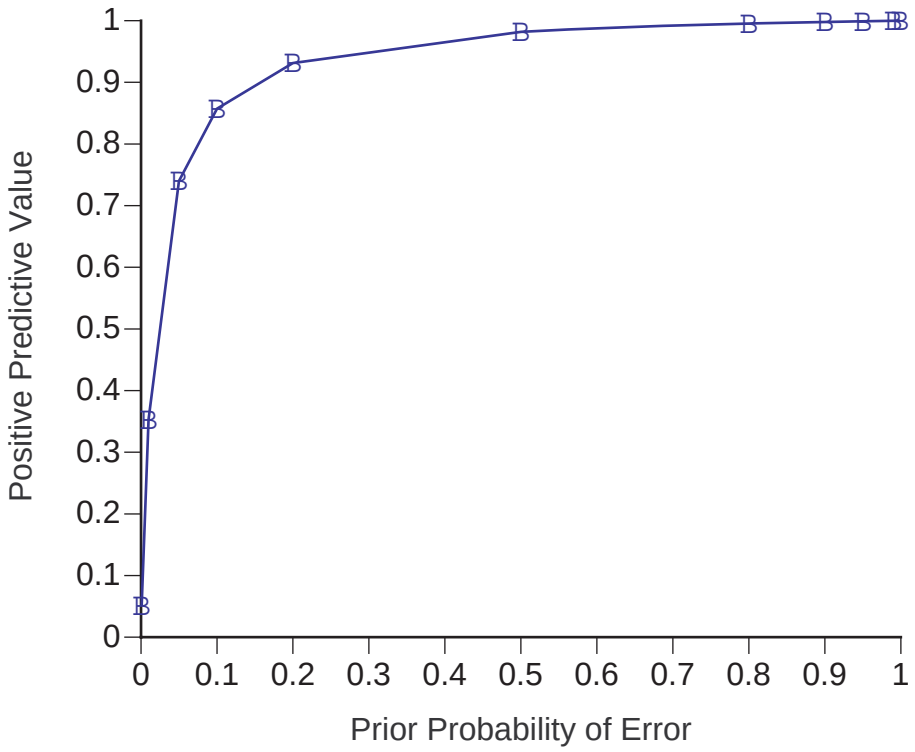
Positive Predictive Value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Prior Probability of Error

Figure 10

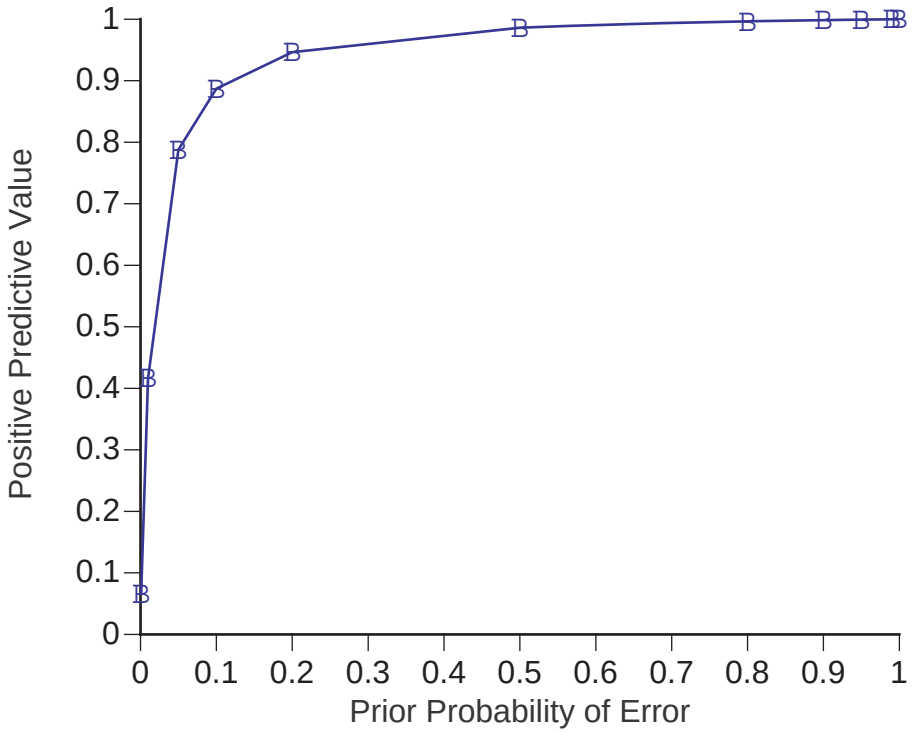
Positive Predictive Value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Prior Probability of Error

Figure 11

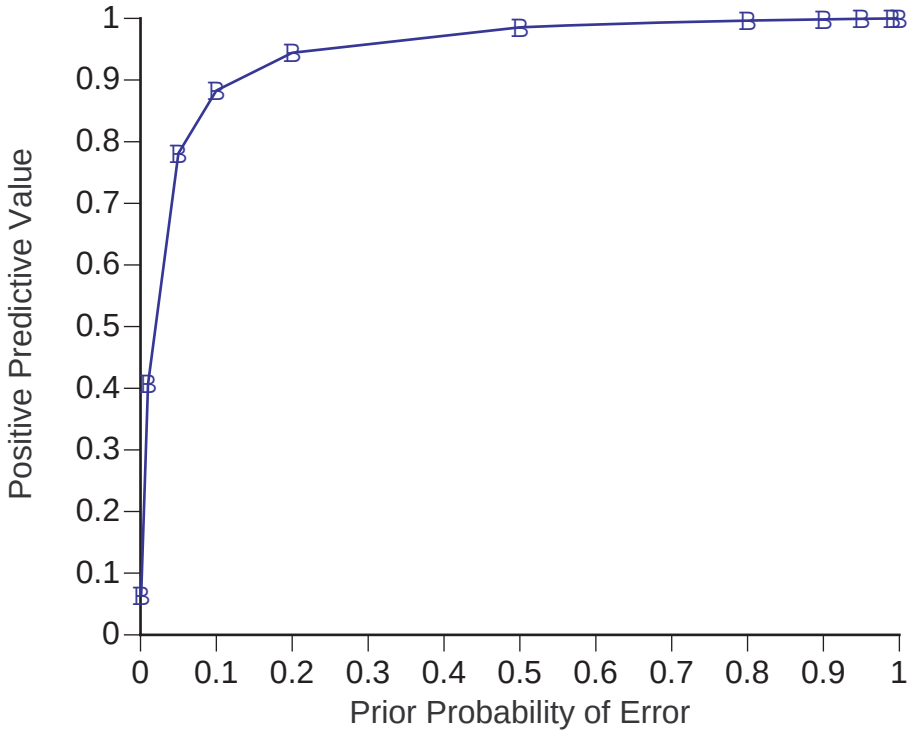
Positive Predictive Value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Prior Probability of Error

Figure 12

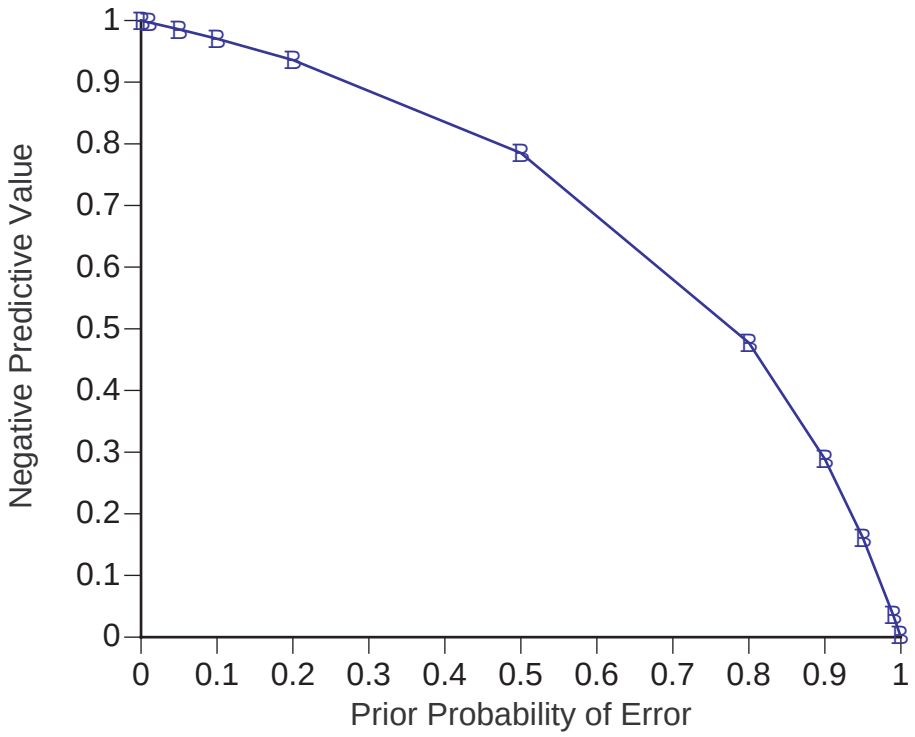
Negative Predictive Value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Prior Probability of Error

Figure 13

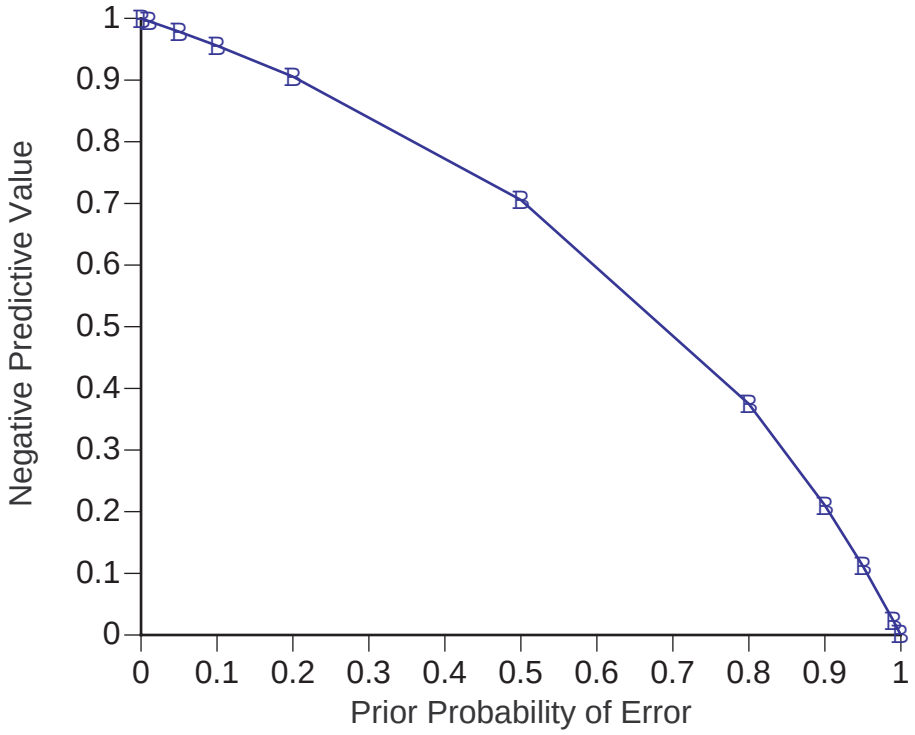
Negative Predictive Value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Prior Probability of Error

Figure 14

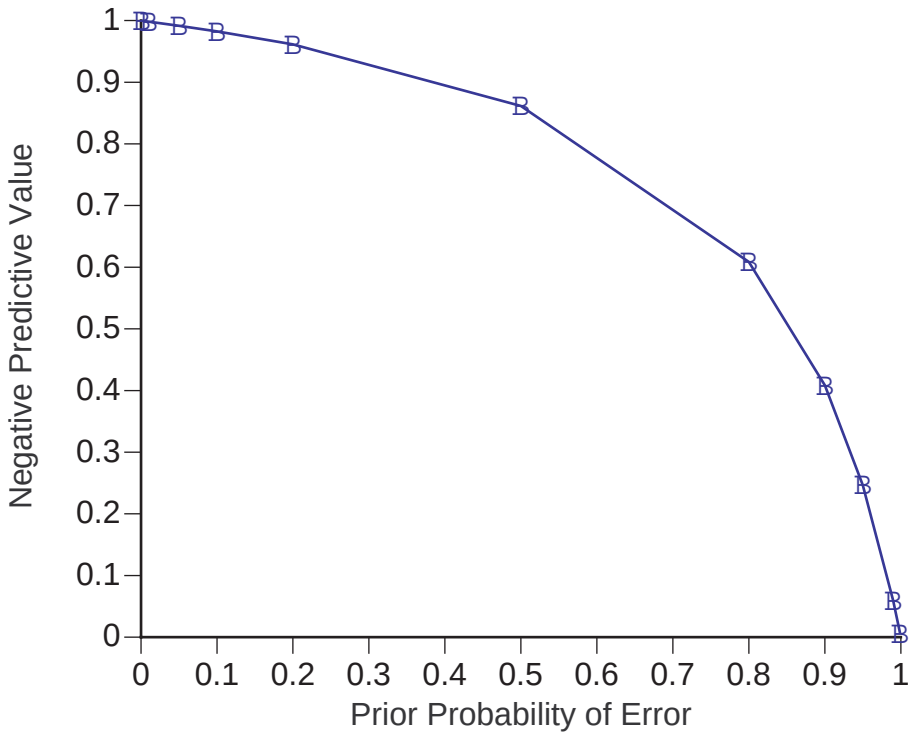
Negative Predictive Value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Prior Probability of Error

Figure 15