Detection and Prediction Limits for Identifying Highly Confutable Drug Names from Experimental Data

ARTICLE *in* JOURNAL OF BIOPHARMACEUTICAL STATISTICS · MAY 2015

Impact Factor: 0.59 · DOI: 10.1080/10543406.2015.1052481 · Source: PubMed

4 AUTHORS, INCLUDING:

Bruce L Lambert

Northwestern University

97 PUBLICATIONS **1,325** CITATIONS

SEE PROFILE

Runa Bhaumik

University of Illinois at Chicago

33 PUBLICATIONS **672** CITATIONS

SEE PROFILE

Dulal K Bhaumik

University of Illinois at Chicago

85 PUBLICATIONS **1,442** CITATIONS

SEE PROFILE

Detection and Prediction Limits for Identifying Highly Confusable Drug Names from Experimental Data

Bruce L. Lambert[1]

Runa Bhaumik[2]

Weihan Zhao[2,$]

Dulal K. Bhaumik[2,3,4,*]

[1]Department of Communication Studies

Center for Communication and Health

Northwestern University

710 Lake Shore Drive, Rm. 414

Chicago, IL 60611

[2]Biostatistical Research Center

Department of Psychiatry

[3]Division of Epidemiology and Biostatistics

1601 West Taylor Street

University of Illinois at Chicago

Chicago, IL 60612-4300

[4] Cooperative Studies Program Coordinating Center (151K)

Hines VA Hospital

5000 South 5th Avenue, Building 1

Hines, IL 60141-3030

* email: dbhaumik@psych.uic.edu

$ Current address: AbbVie Inc., 1 North Waukegan Road, North Chicago, IL 60064

This project was supported by grant number U19HS021093 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

SUMMARY

Confusions between drug names that look and sound alike are common, costly, harmful, and difficult to prevent. One prevention strategy is to screen proposed new drug names for confusability before approving them. Widespread acceptance of pre-approval tests of confusability is compromised by the lack of experimental designs and statistical methods to support valid inferences about whether a proposed new name is unacceptably confusing. One way of identifying confusing names is to conduct memory and perception experiments on a set of drug names which would include both the new name and a set of control names (e.g., names already on the market). The experiment would yield an observed error rate for every name. Inferences about the acceptability of the new name can be made by comparing the error rate of the new name to the distribution of error rates of the control names. We describe four memory and perception experiments on drug names, carried out using clinicians as participants. Each experiment included drug names designated as test and control names. We demonstrate how to use a combination of logistic regression, Poisson prediction limits, and highly assured credible intervals to identify and apply a threshold for identifying unacceptably confusing names. Our models show an excellent fit to the data. These experimental designs and analytic methods should be useful in the pre-approval testing of proposed new drug names and in similar regulatory scenarios where it is necessary to draw inferences about the comparative safety or effectiveness of new versus old products.

Keywords: drug name confusion, medication errors, Poisson prediction limits, detection limits, patient safety

1. Introduction

Drug names that look and sound alike are a leading cause of medication errors (*e.g.*, , diazepam and diltiazem, hydroxyzine and hydralazine, *Paxil* and *Taxol*, fomepizole and omeprazole, *Foradil* and *Toradol*) (Lambert, 1997, 2008; Lambert et al., 2001, 2003, 1999). Depending on the context, drug name confusion can cause prescribing errors, transcription errors, dispensing errors, administration errors, and consumer health product selection errors. The U.S. Pharmacopeia published a comprehensive review of name confusion errors from two large databases of spontaneous error reports covering the years 2003-2006. They identified 26,604 look-alike/sound-alike errors involving $3,170$ confusing pairs of drug names, 1.4% of which caused patient harm (Hicks et al., 2008).

Non-intercepted wrong drug errors in both inpatient and outpatient pharmacy account for between 8% and 11% of all observed errors (Cina et al., 2006; Flynn et al., 2003). Observational studies of dispensing in inpatient and outpatient pharmacies suggest that the rate of wrong drug errors — the type most likely to be the result of name confusion — is roughly one per thousand prescriptions (Cina et al., 2006; Flynn et al., 2003). With roughly 4 billion prescriptions dispensed (Bartholow, 2010), that translates to 4 million wrong drug errors per year in the U.S. If 6.5% were clinically significant (Flynn et al., 2003), that would mean potential harm to roughly $260,000$ people annually. Wrong drug errors are the most common source of malpractice claims against pharmacists (University of Florida College of Pharmacy and PMC Quality Commitment I, 2003). Despite advances in technology, policy and practice, and more than a decade of focused effort, preventing drugs with similar names from being confused by clinicians and patients remains an elusive goal. Given their frequency and potential for harm, preventing wrong drug medication errors such as drug name confusions is a major public health priority.

One way that manufacturers and regulators attempt to minimize the risk of name confusion errors is to subject proposed new names to pre-approval tests to determine their confusability (Lambert et al., 2005; U.S. Food and Drug Administration, 2008). Pre-approval strategies strive to prevent confusing new drug names from entering the marketplace. Pre-approval tests include database searches for existing similar names or products (Lambert et al., 2004), soliciting expert judgments about confusability (Medical Error Recognition and Revision Strategies, 2013), doing psycholinguistic tests on memory and perception (Lambert et al., 2001, 2003), and observing error rates during simulated ordering, dispensing,

and administration tasks (U.S. Food and Drug Administration, 2003b,a).

The widespread acceptance of pre-approval tests of confusability is compromised by the lack of experimental designs and statistical methods that would support inferences about whether a proposed new name is unacceptably confusing. One way of identifying confusing names in the pre-approval setting is to conduct memory and perception experiments on a set of drug names which would include both the proposed new name and a large set of control names (e.g., previously approved names already on the market in a given country). The experiment would yield an observed error rate for each and every name. Inferences about the acceptability of the proposed new name would be made by comparing the error rate of the proposed new name to the mean error rate of the control names. Existing literature in medication safety and experimental psychology provides little guidance on how to carry out the statistical analyses required to make the desired inferences in this scenario.

As a matter of fact, statistical literature remains silent for this critically important area. Inferences are drawn for dispensing errors in general, and for drug confusions in particular based on summary statistics (Hicks et al., 2008; Cina et al., 2006; Flynn et al., 2003; Cohen, 2007; Barker et al., 2002). There are two fundamental drawbacks of this approach, namely (i) effects of drug related fixed covariates are ignored, and (ii) between-drug variation is not incorporated. As a result, inferences drawn from the existing statistical analysis of drug name confusion data are much more susceptible to errors.

The purpose of this paper is to develop and demonstrate rigorous statistical methods that will be used to support inferences about the confusability of proposed new drug names in experimental designs. We describe four different memory and perception experiments on drug names, carried out using pharmacists, physicians and nurses as participants. Each experiment included drug names designated as test names and others designated as controls, the purpose being to draw inferences about the relative confusability of the two sets of names. In Section 2 we discuss all four experiments in detail. In Section 3, we perform several statistical procedures designed to identify any test drug names with unusually high error rates compared to the error rates of control drug names. We discuss the findings and their implications in Section 4. Much of the raw data and detailed descriptions of the methods can be found at http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/34122/version/1. The Superlab code to run the experiments can be downloaded from http://community.cedrus.com/forumdisplay.php?f=9.

2. Design of Experiments

All of the experiments described below were conducted as part of an ongoing effort to develop and test procedures for the pre-approval safety testing of proposed new drug names (Lambert et al., 2005). The idea was to subject each name to a series of tests which would assess confusability, in visual perception (Lambert et al., 2003), auditory perception (Lambert et al., 2010) and short term memory (Lambert et al., 2001, 2003), of new names compared to control names.

2.1 EXPERIMENT 1: PROGRESSIVE DEMASKING

Design and Task: We used a cross-sectional, observational design to study clinicians' ability to correctly identify drug names presented visually on a laptop computer screen. The task that subjects engaged in is known as progressive demasking because it involves identifying a visual stimulus as it is progressively revealed from behind an obscuring mask of #'s. This task is an well-accepted method for assessing accuracy in visual word perception (Dufau et al., 2008; Carreiras et al., 1997; Grainger and Dijkstra, 1996; Grainger and Jacobs, 1996).

Participants: Participants were recruited from among the staff at a large children's hospital in Ottawa, Ontario. A total of $n = 66$ people completed Experiment 1 over a two-day period, 30 on day one and 36 on day two. Because of equipment malfunction, data from Day 1 were discarded. Only results from Day 2 participants are reported below. Participants were primarily female ($n = 31$, 86%), nurses ($n = 16$, 44%) and pharmacists ($n = 11$, 30.6%) along with a few physicians, pharmacy technicians and pharmacy students. Average age was 38.5 ($SD = 11.1$), with 13.3 ($SD = 11.3$) years of experience (see Table 1).

Insert Table 1 here

Stimulus Materials: This experiment used the names of 105 drugs, biologics and natural products taken from the Canadian Drug Product Database (DPD) (Health Canada, 2008) and the Canadian Licensed Natural Health Product Database (LNHPD) (Health Canada, 2013) (see Table 2). We began by selecting five categories of health product names to be included as test names: an injectable prescription drug product, an oral solid prescription drug product, a biological product, an over the counter (OTC) product, and a natural heath product. Because one of the objectives was to determine how the new screening procedure would perform on confusing names, we selected three of the names

(the injectable, the biologic, and the NHP name) from among names that had previously appeared in published reports of drug name confusion errors. These were identified using a combined list of names involved in published lists of look-alike/sound-alike errors. This list had previously been constructed as part of a separate project on drug name confusion (Lambert et al., 1999). Other than the three "error" names, all other test names and control names were selected as stimuli only if they did not appear in the database of published confusing names.

Insert Table 2 here

Test names: Selecting the names was a multi-step process. Within each category, first the test names were chosen at random. Membership in a given category of drugs (*e.g.*, , Rx, OTC, biologic, oral, injectable, natural health product) was determined by codes in the DPD or by inclusion in the LNHPD. For example, to select an injectable name that had previously appeared in a published list of LASA errors, we took all injectable drugs from the DPD and identified the subset of those that were also on the list of published LASA errors. We then selected randomly from among the LASA injectable names, finally selecting *Taxotere*, which had previously been confused with *Taxol*. We followed a similar procedure for the biologics, leading to the selection of *Neulasta* (which had been confused with *Neumega*), and for NHP names, selecting *Aquasol E* (previously confused with *Aquasol A*). We selected test names from the bottom two-thirds of products ranked by Canadian sales volume, according to data provided by IMS Health. By selecting the names from the lower two-thirds of the sales volume list, we insured that the test names had relatively low familiarity and would therefore be more like the unfamiliar new names that are typically presented to regulators for approval.

Control names: For each test name, we selected 20 additional control names from the same category, matched as closely as possible on the number of letters in the test name. For Rx, OTC and biologic products, control names were defined as any name that met two criteria: (1) on the market for at least 5 years (prior to January 2010), according to the earliest "marketed (notified)" date in the DPD; (2) did not appear in previously published error reports. By insisting that the name be on the market for at least 5 years, we reduced the chance that a name would not appear in our error database simply because of its novelty. This was not a perfect definition of control names. Many LASA errors go unreported, so it is possible that some of our control names may have been involved in unreported errors. Nevertheless, this approach likely succeeded in eliminating most known confusing names from our list of controls.

Practice names: The first 10 trials of the experiment were used to familiarize participants with the task. The practice names were chosen at random (*Iressa, Mucaine, Suprane, Eligard, Glycemin, Septopal, Pregvit, Minirin, Tazorac, Antherpos*). Reponses to these names were not scored.

Final set: Because the OTC part of the experiment was dropped due to equipment failure, each participant responded to a total of 94 names: 10 practice names and 84 experimental trials. The 84 experimental names were comprised of 4 test names, each matched with 20 different controls (see Table 2).

Procedure: We used the PDM software package (freely available at www.up.univ-mrs.fr/wlpc/pdm) to conduct the experiments (Dufau et al., 2008). The screen size was set to 1024×768 pixels, color depth to 32 bits. Each trial began with a 500 millisecond fixation point. The experiment was conducted in full-screen mode with a refresh rate of 60Hz. There were 7 refresh cycles for every mask-target pair. Table 3 displays the duration of exposure for targets and masks on each of the 7 cycles in a given trial. Since each screen refresh lasts 1/60th of a second, each 7-refresh cycle lasted 116.8 milliseconds ($7 \times 1/60th = 116.7ms$). The total duration of each trial was 816.67 ms, but after 600 ms (583.3 + 16.17) the target was on the screen constantly with no mask. There was a 1.5 second interval between trials, and a user-controlled pause after 10 practice trials and after the 50th experimental trial. Stimuli appeared in a plain 12-point Arial font. The masking characters (#) also appeared in a plain 12-point Arial font. The 84 stimulus names appeared in a different random order for each participant.

Insert Table 3 here

Scoring: Verbatim typed responses were extracted from the program output and scored as correct if they exactly matched the stimulus name for a given trial. Any deviation from exact matching, even minor spelling errors, was scored as incorrect .

Results: Table 1 summarizes the results of Experiment 1. The 36 participants responded incorrectly to 1337 of the 3024 trials, an error rate of $44.2\%(SE = 0.9\%)$.

Discussion: None of the 4 test names in this experiment differed from the control names in terms of accuracy. How one evaluates this result depends in part on whether on accepts that the test names are, in fact, genuinely more confusing in the real world than control names are. Recall that we selected the test names because each had appeared in at least one published report of a name confusion error. Most errors are spontaneously and voluntarily reported through one or more reporting mechanisms, *e.g.*, , published case reports in clinical journals, FDA MedWatch, ISMP national medication error

reporting program, etc. Even a single occurrence may prompt a report. Thus, the mere appearance of a name in published error report may not be a valid indicator that is an abnormally confusing name. If one regards the test names as genuinely more confusing than most approved names, then the failure of Experiment 1 to identify them as such is a false negative error. Such an error may be due to the general lack of sensitivity of the task or the lack of sensitivity of the task to the underlying type of similarity or confusability that affects this particular name. For example, a visual perception task such as that used in Experiment 1 might not be sensitive to sound-alike similarity between a name and its confusing alternative. If, however, one doubts the validity of spontaneous reports, and therefore doubts that the test names are any more confusing than the average name on the market, then the failure of Experiment 1 to detect any difference between test names and control names is a true positive finding and a challenge to spontaneous error reports as valid indicators of confusability. As we will note in the general discussion below, the methods proposed here require further validation against a better gold standard of confusability. Error rates assessed by large pharmacy systems or by direct observation of prescribing and dispensing might offer such a gold standard.

2.2 EXPERIMENT 2: VISUAL PERCEPTUAL IDENTIFICATION ("PICK-FROM-PAIR")

Design and Task: We used a cross-sectional, observational design to study participants' ability to correctly select a target drug name from a pair of similar drug names after a brief visual presentation of the target on a computer monitor. We refer to this task as "pick-from-pair".

Participants: Participants were recruited from among the staff at a large children's hospital in Ottawa, Ontario. The project was approved by the ethics review board of the hospital. Participants were paid an honorarium in exchange for their participation, and the hospital was paid a fee in exchange for allowing the experimenters to set up an on-site data collection facility in a hospital meeting room. A total of $n = 54$ participants completed Experiment 2 over a two-day period. Table 1 shows demographic characteristics of the $n = 54$ participants in Experiment 2. Participants were primarily female ($n = 46$, 85%), nurses ($n = 21$, 38.9%) and pharmacists ($n = 17$, 31.5%) along with a few physicians, pharmacy technicians and pharmacy students. Average age was 39.2 ($sd = 10.4$), with 13.2 ($sd = 11.5$) years of experience.

Target names: The stimulus names were the same as those described in Experiment 1, except OTC names were included in this experiment. The final set included 105 names (see Table 2).

Nearest neighbor names: Each target name was paired with its nearest neighbor name. For a given target name, the nearest neighbor was defined as the name with the highest similarity to the target name (other than the target itself). To identify nearest neighbors, we computed the similarity between each target and every name in the combined DPD/LNHP database from Health Canada, excluding all but human drugs from the DPD. The name with the highest similarity score was paired with the target. To compute this similarity score, we used the Editex similarity measure (Lambert et al., 1999; Zobel and Dart, 1996). Table 4 gives the target names, nearest neighbors and Editex distance for all the non-OTC names in Experiment 2.

Insert Table 4 here

Editex: Editex is a method for computing a numerical distance or dissimilarity between to strings (i.e., sequences of alphabetic characters). It is a variant on the well-known edit distance algorithm which computes the minimum number of insertions, deletions or substitutions (i.e., "edits") required to transform one sequence into the other. Each additional edit operation increases the edit distance between two sequences. Once two strings are optimally aligned, using a process called dynamic programming, the aligned characters are compared for equality. In a simple edit distance measure, exact matching adds zero to the edit distance, and any mismatch requires an edit, typically increasing the distance by one. Editex groups similar looking or similar sounding letters into equivalence groups (e.g., m and n are in the same group; c and k are in the same group; all vowels are in the same group). With Editex, edit distance is computed as usual, but the cost of a letter substitution depended on the letter groups. If two letters were the same, the cost was 0. If two letters were in the same Editex letter group, the cost was 1. Otherwise, the cost of an insertion, deletion, or substitution was 2. So Editex is just edit distance using variable substitution costs, where the substitution cost is inversely related to similarity between characters.

Table 5 gives the targets, neighbors and distances for all the OTC names in Experiment 2.

Insert Table 5 here

Procedure: We used SuperLab stimulus presentation software (version 4.07b) to conduct the experiments (Cedrus, 2011). The experiment began with 10 practice trials and was followed by 84 non-OTC trials and then 21 OTC trials. Each trial began with a fixation point (+) which remained on the screen

for 2 seconds in 14 point Tahoma font at the center of the screen. Then the target name appeared on the screen at the same location as the fixation point for 40 milliseconds (ms). The target name was then replaced by a string of 16 Xs in 12 point Tahoma font in the center of the screen. The row of Xs remained on the screen for 1 second. Then a pair of names appeared at the center of the screen. One of the names was the target and one was a name similar to the target (*i.e.*, , its nearest neighbor). Whether the target appeared on the left or the right side of the screen was determined at random. Participants indicated which name they thought was the target name by pressing the left or right button on a button box (Cedrus model RB-730) connected to the laptop via a USB cable.

Scoring: The correct location of the target name (either left or right) was pre-programmed into SuperLab for each name pair. SuperLab automatically scored each participant's response as correct or incorrect.

Non-OTC Product Names: Table 1 summarizes the results of Experiment 2 for non-OTC product names. The 54 participants responded incorrectly to 515 of the 4536 trials, for a mean error rate of 11.4% ($SE = 4.3\%$).

OTC Product Names: Table 1 summarizes the results of Experiment 2 for OTC product names. The 54 participants responded incorrectly to 208 of the 972 trials, for a mean error rate of 21.4% ($SE = 1.3\%$).

Discussion: Only one of the 5 test names in this experiment (the over-the-counter drug **Relievol Allergy Sinus Caplets Extra Strength**) differed from the control names in terms of accuracy. As with Experiment 1, the significance of this result hinges on whether the test names can be regarded as, in fact, more confusing than most names already on the market. If the test names are actually more confusing in the real world than the control names, then Experiment 2 produced a false positive result, caused perhaps by lack of sensitivity or lack of external generalizability (i.e., perhaps the experimental task was not enough like real perceptual tasks that cause this name to be confusing in actual practice). But again, if spontaneously reported name confusions are, on their own, not a valid indicator that a name is abnormally confusing, then Experiment 2's failure to detect any difference between test and control names is correct. In the case of **Relievol Allergy Sinus Caplets Extra Strength**, which was chosen at random to be a test name and which was not previously reported to be confusing, Experiment 3 suggests that it may be more confusing than names of other over-the-counter medicines. As with the results from Experiment 1, the ambiguity in interpreting the results highlights the need for more

validation of these methods against a better gold standard of confusability.

2.3 EXPERIMENT 3: AUDITORY PERCEPTUAL IDENTIFICATION (IN NOISE)

Design and Task: We used a cross-sectional, observational design to study clinicians' ability to correctly identify a spoken drug name played back over headphones against a background of multi-speaker babble. This task is known as auditory perceptual identification, and it is widely accepted as a method for assessing the auditory confusability of words (Gernsbacher, 1994; Lively et al., 1994; Luce and Pisoni, 1998; Lambert et al., 2005).

Participants: We recruited participants from among the staff at a large children's hospital in Ottawa, Ontario. The project was approved by the ethics review board of the hospital. Participants were paid an honorarium in exchange for their participation, and the hospital was paid a fee in exchange for allowing the experimenters to set up an on-site data collection facility in a hospital meeting room. A total of $n = 42$ participants completed Experiment 3 over a two-day period. Participants were primarily female ($n = 37$, 88%), nurses ($n = 16$,38%) and pharmacists ($n = 8$, 19%) along with physicians, pharmacy technicians and pharmacy students. Average age was 39 ($SD = 12$), with 16 ($SD = 11$) years of experience (see Table 1).

Stimulus Materials: The stimulus materials were a 69-name subset of those used in Experiment 2 (see Table 2), except 5 additional names were added (*Altocor, amrinone, Kapidex, Omacor and Reminyl*). These additional names had all been removed from the market in the US due to concerns about drug name confusion errors. They were included in this experiment to serve as negative controls. Names were digitally recorded by a male speaker. All names were spoken in a sentence context. The name portion of the sentence was isolated and extracted from the recording. Then each audio file (containing one name) was normalized and the dB of each file was equated to the dB of the quietest file. These last two steps were carried out using the free Praat audio editing software. Then each name file was mixed with background noise (Auditec, 2005) at a fixed signal-to-noise ratio of +8 dB (Lambert et al., 2005).

Procedure: We recruited and scheduled participants in advance via email and posted announcements. Participants arrived at the testing room in groups of four. Each filled out a brief demographic

questionnaire and read a simple consent form before entering the room.

Scoring: The correct location of the target name in the pick list was pre-programmed into SuperLab for each target name, and SuperLab automatically scored each participant's response as correct or incorrect.

Non-OTC Names: The fourth column of Table 1 summarizes the results of Experiment 3 for non-OTC names. Technical malfunctions caused us to discard results from two participants. The 40 participants responded incorrectly to 926 of the 2760 trials, for a mean error rate of 33.6% ($SE = 0.9\%$).

OTC Names: The fourth column of Table 1 summarizes the results of Experiment 3 for OTC names. Technical malfunctions caused us to discard results from two participants. The 40 participants responded incorrectly to 176 of the 520 trials, for a mean error rate of 33.9% ($SE = 2.1\%$).

Discussion: Recognizing the need for a better gold standard of confusability, we added 5 new test names to Experiment 3 (**Altocor**, **amrinone**, **Kapidex**, **Omacor**, and **Reminyl**). Each of these names had been removed from the US market (post-approval) because the FDA judged them to be unacceptably confusing. These names should therefore serve as better gold standards of confusability than the other test names which qualified as test names merely by virtue of appearing in at least one published report of name confusion. Using logistic regression, Poisson prediction limits, and highly assured credible intervals, we detected in Experiment 3 that test names **Altocor**, **Kapidex**, **Neulasta**, **Omacor**, **Taxotere**, and **Relievol Allergy Sinus Caplets Extra Strength** were more confusing than the controls but detected no such difference between any of the other test names and control names (see Tables 6-8). The ability of Experiment 3 to detect that these 6 names were more confusing than the control names provides partial support for the validity of the task in Experiment 3 (auditory perceptual identification) as a method for pre-approval screening of proposed new drug names. Perhaps, had these names been run through these tests prior to approval, they would not have been allowed on the market in the first place. It is noteworthy that some statistical techniques were more sensitive at detecting differences than others, with the credible interval method (see below and Table 8) being the most sensitive (detecting differences in 6 out of 10 test names) and Poisson prediction limits the least sensitive (detecting a difference in only 1 out of 10 test names). Less sensitive methods, e.g., logistic regression, could be made more sensitive by increasing the size of the confidence interval on the prediction limit.

2.4 EXPERIMENT 4: RECOGNITION MEMORY

Design and Task: We used a cross-sectional, observational design to study clinicians' ability to correctly remember a drug name after it is briefly displayed on a computer screen. This recognition memory task is an accepted method for assessing the confusability of words in short term memory (Lambert et al., 2001). Participants were the same as those in Experiment 3.

Stimulus Materials: The stimulus materials were the a 69-name subset of those used in Experiment 2 (see Table 2), except 5 additional names were added (*Altocor*, *amrinone*, *Kapidex*, *Omacor* and *Reminyl*). These additional names had all been removed from the market in the US due to drug name confusion problems. They were included in this experiment to serve as negative controls.

Procedure: Participants had been recruited and scheduled in advance via email and posted announcements. Participants arrived at the testing room in groups of four, each filling out a brief demographic questionnaire and reading a simple consent form before entering the room.

Scoring: The correct location of the target name in the pick list was pre-programmed into SuperLab for each target name, and SuperLab automatically scored each participant's response as correct or incorrect.

Statistical hypothesis: The statistical test was similar to that used in Experiments 1 and 2. The goal was to determine whether the observed error rate for the test names was significantly different than the observed error rate for the controls.

Non-OTC names: The fourth column of Table 1 summarizes the results of Experiment 4 for non-OTC names. The 42 participants responded incorrectly to 500 of the 2898 trials, for a mean error rate of 17.3% ($SE = 0.7\%$).

OTC names: The fourth column of Table 1 summarizes the results of Experiment 3 for non-OTC names. The 42 participants responded incorrectly to 269 of the 756 trials, for a mean error rate of 35.6% ($SE = 1.7\%$).

Discussion: Experiment 4 detected that **Kapidex** and **Relievol Allergy Sinus Caplets Extra Strength** were significantly more confusing than controls. For **Kapidex**, this difference was detected with all three statistical approaches. For **Relievol Allergy Sinus Caplets Extra Strength**, it was detected by the Poisson prediction limits and the credible interval approach but not by logistic regression.

3. Statistical Methods

In what follows we lay down the steps of how to detect new drug names that are more confusing compared to the existing drug names or in other words, new drugs that have "undesirable names". An undesirable drug name is quantified by its error rates. In each experiment, the error rate of a drug name in determined by the misspecification of its correct name by the participants. With the error rate (the complement of the error rate will be called as the accuracy rate) of a new drug name alone, we cannot decide whether it has an undesirable name. In order to determine the status of an experimental drug name, we compare its accuracy rate with those of existing drugs (referred to as control drugs) available in the market. The procedure of our comparison is based on a threshold value or a limit determined by using the distribution of accuracy rates of control drugs. Using the control drugs, we construct three different types of limits based on (i) mixed-effects logistic regression models, (ii) Poisson Distributions, and (iii) Bayesian Analysis. Next, we use those limits as thresholds to detect new drugs that have undesirable names. In what follows we discuss each of these three procedures.

3.1 Mixed-Effects Logistic Regression Models

Denote the response from the kth responder on a control drug i for the experiment j by $y_{ijk}, i = 1, 2, ..., m_i; j = 1, 2, 3, 4; k = 1, 2, ..., n_k$, where $m_i = 80, 80, 60, 60$ and $n_k = 36, 54, 42, 42$ for Experiment 1, 2, 3, 4, respectively. The values of y_{ijk} are denoted by 1 for a correct response, and 0 for an incorrect response. The corresponding probability of a correct response is denoted by p_{ijk}. We use a random intercept logistic regression model to analyze such dichotomous data collected from experiments. The assumption of independence of responses nested within the same responder is justified by the very short duration of each experiment. Carry over effects of such experiments are negligible. The variation of accuracy rates from one drug to the other in the control group is incorporated by the random intercept of the model. We fit experiment-based logistic regression models with random intercept to control drugs. Thus a total of four logistic regression models will be used for four experiments. In our models, we use drug name length (x_{1ij}), familiarity (x_{2ij}) (participants' subjective familiarity on a scale of 1 to 5), years of experience of the responder who is taking the test (x_{3ij}), and age of the responder (x_{4ij}) as fixed covariates. We propose the following mixed-effects logistic regression model to analyze our outcomes.

$$\text{logit}(p_{ijk}) = \alpha_{0j} + \alpha_{ij} + \beta_{1j}x_{1ijk} + \beta_{2j}x_{2ijk} + \beta_{3j}x_{3ijk} + \beta_{4j}x_{4ijk}. \quad (1)$$

In model (1), α_{0j} is the fixed intercept for the jth experiment and α_{ij} is the random intercept (or latent score (LS)) of the ith control drug in the jth experiment that deviates from the mean intercept α_{0j} for the ith control drug. The random intercept is accounted for between control drugs variation. β_{1j}, β_{2j}, β_{3j}, and β_{4j} are fixed parameters associated with four covariates. We assume that α_{ij} follows a normal distribution with mean zero and variance σ_j^2. σ_j^2 measures the between-drug variability on the logit scale for the jth experiment. Estimates of posterior means of the LS α_{ij} is the deviation of the ith drug from the mean in the jth experiment. Note that p_{ijk} is an increasing function of α_{ij}. Thus a small value of the LS of a drug indicates that the corresponding probability of a correct response is small and hence the drug name may be confusing.

3.1.1 Estimation of Parameters

There are several procedures available in the literature for estimating parameters of mixed-effects models for count data. A robust approach that provides consistent estimates together with their standard errors is the marginal maximum likelihood estimation (MMLE) method. We are not using the EM algorithm as its convergence rate is slow. In the MMLE method, we used Gauss-Hermite quadrature for numerical integration, and Newton-Raphson for optimization. SAS version 9.2 is used for numerical computations. Our analysis shows that except the length of drug names, no other covariate becomes significant, i.e. β_{2j}, β_{3j}, and β_{4j} remain insignificant in all four experiments. This implies that responder's specific information plays no significant role to estimate the probability of a correct response for any control drug names. Hence the estimated probability $\hat{p}_{ijk}$ does not depend on k (*i.e.*, responder). Using this information, we estimate p_{ij} from the following expression.

$$\hat{p}_{ij} = \frac{exp(\hat{\eta}_{ij})}{1 + exp(\hat{\eta}_{ij})}, \quad \text{where } \hat{\eta}_{ij} = \hat{\alpha}_{0j} + \hat{\alpha}_{ij} + \hat{\beta}_{1j}x_{1ij}. \quad (2)$$

Using $\hat{p}_{ij}$ we can classify drug names as abnormally confusing or not. However, fitting a distribution on it will provide a better way to determine its percentile points. In what follows, we use a smoothing technique on $\hat{p}_{ij}$ using a beta distribution B(γ_{1j}, γ_{2j}), where γ_{1j} and γ_{2j} are two parameters of the beta distribution for the jth experiment. We estimate these two parameters by using $\hat{p}_{ij}$s. Finally we

determine the $5th$ percentile of the estimated beta distribution and denote it by B_{5j}. These percentile values will be called LB thresholds.

Our strategy is if the observed probability of correctness $\hat{p}_{i*j}$ of an experimental (new) drug $i*$ obtained from jth experiment, is less than B_{5j} (or B_{10j} depending on the situation), then it will get a red flag (i.e. we will call it a confusing or undesirable name). This procedure first identifies 5% (or 10%) worse performing control drugs and compares a test drug with this group of worse performing control drugs. It also means that 95% (or 90%) of the existing control drugs have better rates of correctness than a test drug with a red flag. The 5th (or 10th) percentiles of the fitted beta distributions of four experiments are provided in Table 6.

Insert Table 6 here

Inspecting Table 6, we find that prescription test drug **Altocor** has an inferior performance for Experiment 3, and **Kapidex** has also an inferior performance for both Experiments 3 and 4 based on both 5th and 10th percentiles. The over the counter test drug **Relievol Allergy Sinus Caplets Extra Strength** has also an inferior performance with respect to the 5th percentile in Experiment 2, and with respect to the 10th percentile in Experiments 2, 3, and 4. Our strategy is to put a red flag on an experimental drug if its performance is unsatisfactory in any of the four experiments. Hence the final result is that drugs **Altocor, Kapidex**, and **Relievol Allergy Sinus Caplets Extra Strength** are getting red flags in our analysis.

The current procedure does not borrow any strength from experimental drugs mainly because of two reasons; (i) the number of experimental drugs is far less compared to that of control drugs, and hence the added information from the experimental drugs will not be significant, and (ii) in a realistic situation, when the correct response distributions of these two groups of drugs are different, estimation of parameters from these two combined groups will not reflect the true nature of the control group, and hence the comparison will be dubious. In case the number of experimental drugs is large, we can fit a different logistic-beta model and compute accuracy rates of each experimental drug to compare with the lower percentile of that of control drugs. This procedure provides a better picture regarding the disassociation of two groups than comparing proportions of two groups.

3.2 Poisson Prediction Limit (PPL) for Detecting Confusing Drug Names

Prediction limits that we have constructed in the previous section are adjusted for lengths of control drug names. However, lengths of experimental drugs did not play any role in the construction of such prediction limits. This is mainly because we have only a few experimental drugs. Moreover, in practice experimental drugs are not pre-determined, those will be added in the list gradually over time, and hence their lengths are not known at the time of construction of prediction limits. Parameter estimation and consistency of results will be questionable if we use two different models, one for experimental drugs and the other one for control drugs. This practical difficulty raises the concern of using model-based results for the current problem. In order to address this issue properly, we now construct prediction limits based on Poisson distributions without using any covariates. Let Y_{ij} be the total number of correct responses for the ith control drug from the jth experiment. It is better to fit the distribution for the rates instead of numbers when denominators for control drugs and experimental drugs are different. Let n_j be the number of individuals who participated in the jth experiment to evaluate drug names, and $\bar{y_j} = \sum_{i=1}^{n_j} y_{ij}/n_j$ be the mean number of correct responses. The large sample based formula for the lower Poisson prediction limit is

$$PPL = \bar{y_j} - z_\alpha \sqrt{\bar{y_j}(1 + \frac{1}{n_j})}, \tag{3}$$

where z_α is the $\alpha 100$th percentile point of a standard normal distribution. The idea is to put the red flag on an experimental drug i^* of the jth experiment if $\bar{y}_{i^*j}$ is less than the PPL. Inspecting Table 7, we see that prescription drug name **Kapidex** and over the counter drug name **Relievol Allergy Sinus Caplets Extra Strength** have inferior performance in both Experiments 3 and 4. These two drug names were also detected for their under performance by the logistic regression model.

Insert Table 7 here

3.3 Highly Assured Credible Intervals

In this section we will classify drug names based on the posterior distributions of α_{ij}. An experimental drug name with a smaller number of responders (in real world the number of responders will not be the same for all drug names) is more likely to be classified in the wrong category by chance due to the higher variability of the estimate. Our goal is to classify an experimental drug name to

the undesirable category with a very high accuracy. We try to achieve this goal by making sure that the probability of a new drug being classified to the undesirable category is larger than a pre-specified threshold value. The threshold value is generally taken as high as 90%, and the classification is performed based on the $5th$ percentile of the distribution of accuracy rates of control drugs. Note that p_{ij} is an increasing function of α_{ij}, hence the order of p_{ij} will be maintained if the classification theory is based on α_{ij}. Mathematically, we can write this concept as follows:

$$Pr(\hat{\alpha}_{ij} \leq \xi_{5th}) \geq \gamma, \tag{4}$$

where γ is the pre-specified assurance level, and ξ_{5th} is the score's $5th$ percentile. The procedure has two steps. In the first step we construct $(2\gamma-1)100\%$ credible interval for α_{ij} under the assumption that the corresponding posterior distribution is symmetric. This credible interval will be specific to the ith drug in the jth experiment. Next, we determine the $5th$ percentile of the upper bounds of all credible intervals. By construction, the probability of scores less than the $5th$ percentile of the upper bounds of credible intervals is γ. To implement this concept, we assume that experimental and control drugs have the same distribution and the $5th$ percentile is determined by combining both experimental and control drugs. Note that the Poisson and logistic prediction limits are not based on this assumption. To estimate parameters, we now implement the full Bayes (FB) approach by using the WinBUGS package. We used non-informative prior distributions since we do not have specific prior knowledge about the parameters. Prior distributions for the fixed parameters α_{ij} are chosen to be N(0,100), and priors for the variance parameters σ are $\Gamma(0.01, 0.01)$. The model we fit to the data is a mixed-effects logistic regression model with random intercepts. We updated two Markov chains with different starting values. We discard the first 5000 iterations for burn-in and further updated 5000 iterations to obtain the posterior sample. Convergence diagnostics is performed using several different approaches. First, we determine whether the Markov chain in the model updates reaches the stabilizing distribution by examining the trace plot of every parameter, and look for nice bell-shaped curves on the density plots of all parameters. Next, for each parameter, the within-chain auto-correlation with lags not greater than zero is verified. In addition, we examined the Brooks-Gelman-Rubin (BGR) statistic, which is a ratio of the between-chain and within-chain variation. Upon convergence, the closeness of BGR statistic to one is verified. Figure 1 shows the trace, density, and BGR plots of some of the parameters.

Insert Figure 1 here

Parameter estimates are based on 5000 posterior samples obtained upon convergence of the chains. We fix γ at .90. We obtain posterior means, standard deviations, and $90th$ percentiles for each drug. The $90th$ percentiles are playing the roles of upper bounds. Based on these upper bounds, we obtain the $5th$ and $10th$ percentile thresholds. Thus the thresholds are not drug specific even though they depend on experiments. In Table 8, we provide $5th$ and $10th$ percentile scores and identify both prescription and over the counter drugs. In addition to **Altocor** and **Kapidex**, we identify two new prescription drugs **Omacor** and **Taxotere** that are getting red flags by this method. Over the counter drug **Relievol Allergy Sinus Caplets Extra Strength** is also identified as a confusing drug name by this method.

Insert Table 8 here

4. Model Validations

In this section we are proposing two different methods for model validation. The first method compares predicted accuracy rates with observed accuracy rates by computing a special type of correlation. A high value of this correlation indicates a strong agreement. The second method evaluates our model based results by cross validation which is frequently used in machine learning literature. These methods are implemented to both logistic regression and Poisson models. In what follows we will see that results obtained using Poisson model are weaker than those obtained by the Logistic regression model.

4.1 Agreement Between Estimated and Observed Rates

In order to justify the proposed methods for detecting drug names, it is desirable to have a strong agreement between estimated and observed probabilities of accuracy. We incorporate Lin's[16] concordance correlation coefficient(CCC) to compare predicted and observed probabilities of correctness. In our context the CCC measures the agreement between observed and predicted values as opposed to the Pearson correlation coefficient which measures their linear relationship. Let us denote two commensurable variables by X and Y, their means by μ_X, and μ_Y respectively, their variances by σ_{XX}, and σ_{YY},

and their covariance by σ_{XY}. The CCC denoted by ρ_c is defined as

$$\rho_c = \frac{2\sigma_{XY}}{\sigma_{XX} + \sigma_{YY} + (\mu_X - \mu_Y)^2}. \tag{5}$$

The CCC is the product of Pearson correlation coefficient and a bias correction factor (BCF). In fact Lin (1989) showed that $\rho_c = \rho BCF$, where BCF=$[(v + 1/v + u^2)/2]^{-1}$. In BCF, $v = \sqrt{\sigma_{XX}/\sigma_{YY}}$ measures the scale shift, and $u = (\mu_X - \mu_Y)/\sqrt{\sigma_{XX}\sigma_{XX}}$ measures the location shift relative to the scale, and ρ is the Pearson correlation coefficient. Here $0 < BCF \leq 1$, and it measures the accuracy, or amount of deviation from a 45^0 line. The deviation becomes zero when the line between the observed and predicted values passes through the origin and makes a 45^0 angle with the horizontal line and in that case $BCF = 1$.

The CCC takes values between -1 and 1. $CCC = 1$ indicates a high correlation between the observed and predicted values and it is achieved when the line passes through the origin and makes a 45^0 angle with the horizontal line. The estimator of CCC proposed by Lin is consistent and it has asymptotic properties, i.e. it follows, asymptotically, a normal distribution. We compute the CCC for all experiments and for both prescription and over the counter drugs. In this study the smallest value of CCC is .909 for prescription drugs in Experiment 2, and the largest value is .991 for over the counter drugs in both Experiment 2 and 3 (see Table 9). Generally, CCC with a value of more than .80 indicates a very high agreement. Thus the proposed random intercept logistic regression model fits well for the current data. Note that computation of estimated accuracy rates utilize observed probabilities, and hence the CCCs are inflated. If the accuracy rates of new drugs are estimated on the basis of their characteristics only (i.e., drug length), the corresponding CCC value is expected to be smaller than what we see in Table 9.

Insert Table 9 here

4.2 Cross Validation Approach

To evaluate the proposed logistic regression model together with the smoothing technique by the beta distribution (call it by LR-BD), we use a k-fold cross validation approach which is widely accepted in data mining and machine learning community. Cross validation is a statistical method for evaluating

and comparing learning algorithms by dividing data into two segments: one segment used to learn or train a model and the other segment used to validate the model. In a k-fold cross-validation, the data is first partitioned into k equally (or nearly equally) sized segments or folds. Subsequently k iterations of training and validation are performed such that within each iteration, a different fold of the data is held out for validation while the remaining $k-1$ folds are used for learning. Next, k results obtained from folds are averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. Generally a 10-fold cross-validation is used in practice. In our experiment, we divided the control drugs into training and validation sets and performed a 10-fold cross validation for prescription drugs and a 5-fold cross validation for over the counter drugs (due to the limited number of control drugs). The LR-BD threshold was determined using the training data set. The observed probability of each drug in the validation set was compared with the LR-BD threshold and classified the drug as confusing or not. Knowing the ground truth of each drug in the validation set (i.e., whether actually the drug is confusing or not is known), we are able to determine whether the drug is correctly classified or not. We measured the sensitivity as the percentages of correctly identified drugs in the validation set. Table 10 shows the partition of the data for each experiment and the sensitivity of the model. Inspecting Table 10 we find that the sensitivity for most of the experiments are more than 90.00%. The sensitivity of Experiment 2 for prescription drugs is 79.00% and Experiment 3 for OTC drugs is 79.00%. Similarly, the specificities are more than 90.00% in Experiment 2 and Experiment 3 for prescription drugs and Experiment 2 for OTC drugs. The specificity of other experiments are also more than 70.00%. This strong result validates the use of LR-BD approach to identify experimental drug names that are confusing compared to control drug names.

Insert Table 10 here

5. Discussion

Drug name confusions are a common type of medication error, and they often have harmful effects. Preventing such errors is an important priority in patient safety research and practice. Subjecting new names to a series of memory and perception tests prior to approval is one method that has been

proposed for identifying names with a high likelihood of being confused once they entered the market. One obstacle to the acceptance of these pre-approval testing methods has been the absence of validated experimental designs and statistical frameworks for making the types of inferences that are critical, i.e., inferences about whether a proposed new name is any more confusing than a valid sample of existing names. In this paper we illustrated experimental designs (e.g., memory and perception experiments with test names and controls), and we developed a valid statistical framework. The statistical framework supports two types of inferences: (i) is the error rate for the test name different than that observed for the controls; and (ii) does the test name fall in the extreme range of the distribution of controls. Application of these analytic methods to the data demonstrated that the methods could identify some names that were known to be confusing (i.e., some of those which were previously removed from the market due to confusability) and could identify extreme areas of the distribution of error or accuracy rates that could be used as thresholds for the acceptability or unacceptability of proposed new drug names. We also demonstrated that the models provided an excellent fit to the observed data.

Although we demonstrated the usefulness of these methods in the specific domain of newly proposed drug names, we believe they may have general utility in a variety of regulatory scenarios. Whenever a new drug or device or regulated product has a measured characteristic (e.g., a failure rate) and there exists a population of previously approved products in the same category on which this characteristic can be or already has been measured, then it should be possible to use the methods described here to make valid inferences about the characteristic of the new product compared to that observed in the population of approved products. Consider the case of a new medical device such as an artificial hip or knee joint. Using the methods described here, the failure rate of a new device in clinical studies, or in the first year of real-world use, could be compared to the failure rate observed in the population of all other devices in the same category, allowing one to make inferences about the comparative safety or effectiveness of the new product compared to similar existing products.

6. Limitations

We used only a small set of drug names on a relatively small, non-representative sample of clinicians from one hospital. Not all of the drug names known to be involved in previous errors were found to be more confusing than the controls in our experiments. There are several possible explanations.

The experiments may not perfectly assess real world confusability. The experiments may have lacked the power to detect small differences between test and control names. Some of the names which were withdrawn from the American market may not have been equally confusing in the Canadian market, where the population of drug names is different. Some drugs identified as confusing because they appeared in published error reports may, in fact, not be any more confusing than the control names. There is one additional possibility. The absolute number of errors that come to light in the marketplace once a drug is approved and in use is a function of (i) the inherent confusability of the name (which should be captured in experiments like ours) and (ii) the number of opportunities for error, which is primarily a function of the dispensing frequency of a given drug (a fact not captured in our experiments). The experiments used here, although valid measures of confusability in visual perception, auditory perception, and short term memory, are very low fidelity simulations of real world clinical practice, in that the tasks do not use realistic prescriptions, labels, packages or order entry systems.

7. Conclusion

Prediction limits can be used to support inferences about the relative confusability of drug names in controlled experiments of visual perception, auditory perception and short term memory. These experimental designs and analytic methods may be of use in the pre-approval safety testing of proposed new drug names, and they may also be useful in similar regulatory scenarios where it is useful to draw inferences about the comparative safety or effectiveness of new and old products.

References

Auditec (2005). Multitalker (20-speaker) [audio recording]. *Auditec, St. Louis, MO*.

Barker, K. N., Flynn, E. A., Pepper, G. A., Bates, D. W., and Mikeal, R. L. (2002). Medication errors observed in 36 health care facilities. *Archives of Internal Medicine*, 162:1897–1903.

Bartholow, M. (2010). Top 200 prescription drugs of 2009. *http://www.pharmacytimes.com/publications/issue/2010/May2010/RxFocusTopDrugs-0510*, Accessed April 21, 2011.

Carreiras, M., Perea, M., and Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:857–871.

Cedrus (2011). Superlab. *http://www.cedrus.com/*, Accessed April 25, 2011.

Cina, J. L., Gandhi, T. K., Churchill, W., Fanikos, J., McCrea, M., Mitton, P., Rothschild, J. M., Featherstone, E., Keohane, C., Bates, D. W., and Poon, E. G. (2006). How many hospital pharmacy medication dispensing errors go undetected. *Journal on Quality and Patient Safety*, 32:73–80.

Cohen, M. R. (2007). *Medication errors*. American Pharmaceutical Association, Washington, DC.

Dufau, S., Stevens, M., and Grainger, J. (2008). Windows executable software for the progressive demasking task. *Behavior Research Methods*, 40:33–37.

Flynn, E. A., Barker, K. N., and Carnahan, B. J. (2003). National observational study of prescription dispensing accuracy and safety in 50 pharmacies. *Journal of the American Pharmaceutical Association*, 43:191–200.

Gernsbacher, M. A. (1994). *Handbook of psycholinguistics*. Waltham, MA: Academic Press.

Grainger, J. and Dijkstra, T. (1996). *Visual word recognition: Models and experiments*. Taylor & Francis, Bristol, PA.

Grainger, J. and Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, 103:518–565.

Health Canada (2008). Drug product database. *http://www.hc-sc.gc.ca/dhp-mps/prodpharma /databasdon /index-eng.php/*, Accessed October, 2008.

Health Canada (2013). Licensed natural health products database. *http://www.hc-sc.gc.ca/dhp-mps/prodnatur/applications/licen-prod/lnhpd-bdpsnh-eng.php*, Accessed June 3, 2013.

Hicks, R., Becker, S., and Cousins, D. (2008). Medmarx data report: A report on the relationship of drug names and medication errors in response to the institute of medicineâĂŹs call for action. *US Pharmacopedia, Rockville, MD*.

Lambert, B. L. (1997). Predicting look- and sound-alike medication errors. *American Journal of Health-System Pharmacy*, 54:1161–1171.

Lambert, B. L. (2008). Recent developments in the prevention and detection of drug name confusion. In Hicks, R. W., Becker, S. C., and Cousins, D. D., editors, *Recent developments in the prevention and detection of drug name confusion*. Rockville, MD: US Pharmacopeia.

Lambert, B. L., Chang, K. Y., and Gupta, P. (2003). Effects of frequency and similarity neighborhoods on pharmacists' visual perception of drug names. *Social Science and Medicine*, 57:1939–1955.

Lambert, B. L., Chang, K. Y., and Lin, S. J. (2001). Effect of orthographic and phonological similarity on false recognition of drug names. *Social Science and Medicine*, 52:1843–1857.

Lambert, B. L., Dickey, L. W., Fisher, W. M., Gibbons, R. D., Lin, S. J., Luce, P. A., McLennan, C. T., Senders, J. W., and Yu, C. T. (2010). Listen carefully: the risk of error in spoken medication orders. *Social Science and Medicine*, 70:1599–1608.

Lambert, B. L., Lin, S. J., Gandhi, S. K., and Chang, K. Y. (1999). Similarity as a risk factor in drug name confusion errors: The look-alike (orthographic) and sound-alike (phonological) model. *Medical Care*, 37:1214–1225.

Lambert, B. L., Lin, S. J., and Tan, H. K. (2005). Designing safe drug names. *Drug Safety*, pages 495–512.

Lambert, B. L., Yu, C., and Thirumalai, M. (2004). A system for multi-attribute drug product comparison. *Journal of Medical Systems*, 28:29–54.

Lively, S. E., Pisoni, D. B., and Goldinger, S. D. (1994). *Spoken word recognition*. Waltham, MA: Academic Press.

Luce, P. A. and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, 19:1–36.

Medical Error Recognition and Revision Strategies (2013). *www.med-errs.com*, Accessed January 14, 2004.

University of Florida College of Pharmacy and PMC Quality Commitment I (2003). CQI compliance guide for Florida pharmacists. *University of Florida, Gainesville, FL*.

U.S. Food and Drug Administration (2003a). Drug safety and risk management advisory committee meeting. *http://www.fda.gov/ohrms/dockets/ac/03/slides/4007s1.htm*, Accessed January 14, 2003.

U.S. Food and Drug Administration (2003b). Evaluating drug names for similarities: Methods and approaches (public meeting). *http://www.fda.gov/cder/meeting/drugNaming.htm*, Accessed January 14, 2004.

U.S. Food and Drug Administration (2008). PDUFA pilot project proprietary name review. *Federal Register*, 74.

Zobel, J. and Dart, P. (1996). *Phonetic string matching: Lessons from information retrieval: Lessions from information retrieval*. New York: Association for Computing Machinery.

Table 1. Demographics of Participants

Characteristic	Exp1 (n=36) Mean (SD)	Exp2 (n=54) Mean (SD)	Exp3 and Exp4 (n=42) Mean (SD)
Age	38.5 (11.1)	39.2 (10.4)	39.12 (11.7)
Experience	13.3 (11.3) (n=35)	13.2 (11.5)	16.22 (11.0)
Gender	N (%)	N (%)	N (%)
Male	5 (13.9)	8 (14.5)	5 (11.9)
Female	31 (86.1)	46 (85.2)	37 (88.1)
Language			
English	27 (75.0)	39 (72.2)	31 (73.8)
French	5 (13.9)	8 (14.8)	9 (21.4)
Other	3 (8.3)	5 (9.3)	1 (2.4)
Missing	1 (2.7)	2 (3.7)	1 (2.4)
Licensed			
Yes	29 (80.6)	43 (79.6)	31 (73.8)
No	7 (19.4)	11 (20.4)	4 (9.52)
NA		7 (16.7)	
Profession			
MD	2 (5.6)	5 (9.3)	5 (11.9)
RN	16 (44.4)	21 (38.9)	16 (38.1)
RPh	11 (30.6)	17 (31.5)	8 (19.0)
Pharmacy Tech.	4 (11.1)	8 (14.8)	10 (23.8)
Pharmacy Student	3 (8.3)	3 (5.6)	2 (4.8)
Consumer		1 (2.4)	
Degree			
MD	1 (2.8)	4 (7.4)	5 (11.9)
RN	14 (38.9)	20 (37.0)	14 (33.3)
RPh	4 (11.1)	6 (11.1)	3 (7.14)
PharmD	4 (11.1)	5 (9.3)	5 (11.9)
BS	7 (19.4)	9 (16.7)	1 (2.4)
Other	5 13.9)	8 (14.8)	9 (21.4)
None	1 (2.8)	3 (5.6)	5 (11.9)
Dominant Hand			
Right	34 (94.4)	52 (96.3)	37 (88.1)
Left	2 (5.5)	2 (3.7)	5 (11.9)
Prior Participant			
Yes			16 (38.1)
No			26 (61.9)

Prior participant: participants in Experiment 3 or 4 who had participated in Experiment 1 and 2. Both sets of experiments were conducted at the same institution but on separate occasions several weeks apart.

Table 2. Stimulus Names for Experiments 1 and 2

Biologics	Natural Health Products	OTC	Rx Oral	Rx Injectable
Neulasta[1,2]	Aquasol E[1,2]	Relievol Allergy Sinus Caplets Extra Strength[1]	Parsitan[1]	Taxotere[1,2]
Act-Hib	Amygdales	Balminil DM + Decongestant + Expectorant	Apo-Sulin	Baciject
Amevive	Arthron 5	Balminil DM + Expectorant Extra Strength	Apo-Verap	Betaject
Betaseron	Artichaut	Children's Coltalin Fruit Flavor Chewable Cold	Cystadane	BSS Plus
Center-Al	Atropinum	Coricidin II Extra Strength Cold And Flu	Dalmacol	Cetrotide
Dukoral	Blueberry	Denti-Care Prophylaxis Paste With Fluoride	Dexasone	Diazemuls
Immucyst	Carthamex	Dristan Long Lasting Nasal Spray Mentholated	Euglucon	Diphenist
Intron A	Cetecal D	Extra Strength Head Cold And Sinus Caplets	Formulex	Epipen Jr
Neisvac-C	Dangguisu	Extra Strength Sinus Medication Daytime Relief	Hydergine	Extraneal
Ovidrel	Eurocal D	Hot Lemon Relief For Symptoms Of Cough And Cold	Lectopam	Fortaplex
Pegasys	Glonoinum	Magnesia's Pellegrino Type Mmmm Effervescent	Lescol XL	Hemabate
Pegetron	Glutamine	Multigenics Intensive Care Formula Without Iron	Mazepine	Kidrolase
Pregnyl	Hepatinum	Muscle And Back Pain Relief Extra Strength	Norventyl	Quelicin
Priorix	Homeoslim	Nasal Decongestant Spray With Moisturizers	Nu-Cimet	Remodulin
Pulmozyme	Melaton-3	Preservative-Free Thera Tears Lubricant Eye Drops	Nu-Pindol	Robaximol
Repronex	Modu Chol	Rheuma Heilkrauter Tee (Rheumatism Herbal Tea)	Parvolex	Rogitine
Stemgen	Osteomate	Ricola Swiss Lemon-Mint Herb Cough Drops	Protylol	Serostim
Typherix	Pain Ease	Scott's Emulsion Of Cod Liver & Capelin Oil	Sensipar	Suprefact
Typhim Vi	Florabile	Sucrets Cough Control Extra Strength Lozenges	Tebrazid	Thyrogen
Varilrix	Ultra Efa	Timed Release Ultra Mega Gold Without Iron	Ulcidine	Valtaxin
Vivotif	Uristatin	Vicks Custom Care Chest Congestion/Cough	Yohimbine	Vascoray

[1]: test names; [2]: names previously involved in published name confusion errors.
OTC names were excluded from Experiment 1.

Table 3. Summary of Results for Experiments 1-4

Variable		Experiment			
		1: Progressive De-masking	2: Pick from Pair	3: Recognition Memory	4: Auditory Perception
OTC	**No. Participants**		54	42	40
	No. Names		18	18	13
	Control		17	17	12
	Test		1	1	1
	Error Rate (%)				
	Overall				
	Mean (SE)		21.4 (1.3)	35.6 (1.7)	33.9 (2.1)
	Range		0-50.0	11.1-66.7	7.7-76.9
	Control Names				
	Mean (SE)		20.2 (1.3)	34.2 (1.7)	32.5 ()
	Range		0-47.1	5.9-64.7	0-75
	Test Names				
	Mean (SE)		42.6 (6.8)	59.5 (7.6)	50.0 ()
	Range		0-100	0-100	0-100
Rx, NHP, and Bio	**No. Participants**	36	54	42	40
	No. Names	84	84	69	69
	Control	80	80	60	60
	Test	4	4	9	9
	Error Rate (%)				
	Overall				
	Mean (SE)	44.2 (0.9)	11.4 (0.5)	17.3 (0.7)	33.6 (0.9)
	Range	7.1-90.5	0-40.5	1.4-55.1	5.8-60.9
	Control Names				
	Mean (SE)	44.4 (0.9)	11.5 (0.5)	16.6 (0.7)	34.3 (1.0)
	Range	7.5-92.5	0-41.3	1.7-50	5-61.7
	Test Names				
	Mean (SE)	39.6 (4.1)	8.3 (1.8)	21.4 (2.1)	28.7 (2.4)
	Range	0-100	0-50	0-88.9	0-55.6

OTC=Over the counter, Rx=prescription, NHP=natural health product, Bio=biological.

Table 4. Target Names, Nearest Neighbor Names and Editex Distance (Experiment 2)

Target	Neighbor	Editex	Target	Neighbor	Editex	Target	Neighbor	Editex
act-hib	actos	10	extraneal	estrogel	9	pegasys	pegalax	7
amevive	amatine	8	florabile	florasil	6	pegetron	protrin	10
amygdales	amygdeel	8	formulex	formule a	6	pregnyl	pronal	7
apo-sulin	apo-gain	8	fortaplex	formulex	8	priorix	protrin	8
apo-verap	apo-peram	5	glonoinum	glycerinum	10	protylol	propofol	8
aquasol e	atasol 8	9	glutamine	glutapure	8	pulmozyme	pulminum	10
arthron 5	arthron	6	hemabate	hemoban	8	quelicin	quetiapine	11
artichaut	artechol	9	hepatinum	hepaton-s	8	remodulin	robitussin	12
atropinum	abrotanum	8	homeoslim	homeocap	10	repronex	reparagen	11
baciject	betaject	7	hydergine	hytrin	11	robaximol	robaxin	8
betaject	baciject	7	immucyst	imuran	11	rogitine	reactine	8
betaseron	betaxin	10	intron a	ironol	12	sensipar	senior	8
blueberry	bilberry	8	kidrolase	karesse	13	serostim	serofin	8
bss plus	betullus	9	lectopam	lycopus	10	stemgen	sialgen	8
carthamex	cardamom	9	lescol xl	lescol	9	suprefact	suplevit	9
center-al	control	10	mazepine	mycamine	9	taxotere	taxol	11
cetecal d	cical	14	melaton-3	melatonin	6	tebrazid	tofranil	11
cetrotide	choroide	9	modu chol	monurol	11	thyrogen	thyrocsin	7
cystadane	cetacaine	10	neisvac-c	nevanac	11	typherix	thyplex	10
dalmacol	dalmane	8	neulasta	neuleptil	10	typhim vi	typherix	14
dangguisu	digest	12	norventyl	nervosyl	10	ulcidine	urixin	11
dexasone	depakene	8	nu-cimet	nu-nifed	8	ultra efa	ultra mega	6
diazemuls	diazepam	11	nu-pindol	nu-indo	6	uristatin	urixin	11
diphenist	dipentum	11	osteomate	osteocit	8	valtaxin	voltaren	7
dukoral	doloral	5	ovidrel	oxytrol	9	varilrix	varizig	8
epipen jr	epipen	9	pain ease	pain aid	9	vascoray	vasotec	10
euglucon	eupion	11	parsitan	parnate	10	vivotif	vivotif l	6
eurocal d	euro d	9	parvolex	pariodex	6	yohimbine	yasmin	13

Table 5. Target Names, Nearest Neighbor Names and Distance for All OTC Names (Experiment 2)

Target	Neighbor	Editex
balminil dm + decongestant + expectorant	balminil codeine + decongestant + expectorant	16
balminil dm + expectorant extra strength	balminil dm + expectorant	44
children's coltalin fruit flavor chewable cold	children's tylenol cold chewable	61
coricidin ii extra strength cold and flu	coricidin ii cold and flu	42
denti-care prophylaxis paste with fluoride	dr. ken toothpaste with fluoride	54
dristan long lasting nasal spray mentholated	dristan long lasting nasal mist	41
extra strength head cold and sinus caplets	extra strength tylenol allergy sinus caplets	34
extra strength sinus medication daytime relief	extra strength cold medication (daytime rel)	24
hot lemon relief for symptoms of cough and cold	hot lemon relief for symptoms of cough	26
magnesia's pellegrino type mmmm effervescent	magnesia s. pellegrino	62
multigenics intensive care formula without iron	multigenics intensive care formula	34
muscle and back pain relief extra strength	muscle and back pain relief regular strength	14
nasal decongestant spray with moisturizers	no7 soft and sheer tinted moisturizer spf	66
preservative-free thera tears lubricant eye drops	preservative-free cosopt	75
relievol allergy sinus caplets extra strength	relievol sinus caplets extra strength	21
rheuma heilkrauter tee (rheumatism herbal tea)	red maple naturals prenatal formula	77
ricola swiss lemon-mint herb cough drops	ricola swiss herb cough drops	33
scott's emulsion of cod liver & capelin oil	seven seas cod liver oil	69
sucrets cough control extra strength lozenges	sucrets extra strength cherry lozenges	54
timed release ultra mega gold without iron	timed release ultra mega gold	34
vicks custom care chest congestion/cough	vicks custom care nasal congestion/cough	13

Table 6. Identification of Confusing Drug Names using Mixed-Effect Logistic Regression Models

Prescription Drugs									
	Prediction	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
Drug	Limit	5th=.30	10th=.35	5th=.80	10th=.82	5th=.72	10th=.75	5th=.40	10th=.46
Altocor		-	-	-	-	.71*	.71*	.60	.60
Amrinone		-	-	-	-	.85	.85	.82	.82
Aquasol E		.69	.69	.90	.90	.85	.85	.92	.92
Kapidex		-	-	-	-	.61*	.61*	.30*	.30*
Neulasta		.72	.72	.87	.87	.78	.78	.60	.60
Omacor		-	-	-	-	.76	.76	.82	.82
Parsitan		.47	.47	.88	.88	.85	.85	.72	.72
Reminyl		-	-	-	-	.85	.85	.85	.85
Taxotere		.52	.52	1.0	1.0	.76	.76	.77	.77

Over-the-Counter Drugs									
	Prediction	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
Drug	Limit			5th=.58	10th=.63	5th=.36	10th=.43	5th=.46	10th=.51
Relievol Allergy Sinus Caplets Extra Strength		-	-	.57	.57	.40	.40*	.5	.5

*: Accuracy rate less than the prediction limit.

-: Not assessed in experiment.

Table 7. Identification of Confusing Drug Names using Poisson Prediction Limits

Prescription Drugs

Drug	PPL#	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
		5th=12.6	10th=14.2	5th=36.3	10th=38.9	5th=25.2	10th=27.4	5th=17.8	10th=19.7
Altocor		-	-	-	-	30	30	24	24
Amrinone		-	-	-	-	36	36	33	33
Aquasol E		25	25	49	49	36	36	37	37
Kapidex		-	-	-	-	26	26*	12*	12*
Neulasta		26	26	47	47	33	33	24	24
Omacor		-	-	-	-	32	32	33	33
Parsitan		17	17	48	48	36	36	29	29
Reminyl		-	-	-	-	36	36	34	34
Taxotere		19	19	54	54	32	32	31	31

Over-the-Counter Drugs

Drug	PPL#	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
				5th=32.0	10th=34.5	5th=18.7	10th=20.7	5th=18.1	10th=20.1
Relievol Allergy Sinus Caplets Extra Strength		-	-	31	31	17*	17*	20	20*

#: PPL: Poisson prediction limit.

*: Mean accuracy less than the PPL.

-: Not assessed in experiment.

Table 8. Identification of Confusing Drug Names by the Credible Interval Method

Prescription Drugs

Drug	EAS#	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
		5th=-0.480	10th=-0.381	5th=-0.331	10th=-0.155	5th=-0.486	10th=-0.213	5th=-0.804	10th=-0.549
Altocor		-	-	-	-	-0.5351*	-0.5351*	-0.3098	-0.3098
Amrinone		-	-	-	-	0.1055	0.1055	0.6592	0.6592
Aquasol E		0.4753	0.4753	0.09432	0.09432	0.09991	0.09991	1.275	1.275
Kapidex		-	-	-	-	-0.8799*	-0.8799*	-1.389*	-1.389*
Neulasta		0.5757	0.5757	-0.1497	-0.1497	-0.2271	-0.2271*	-0.302	-0.302
Omacor		-	-	-	-	-0.3407*	-0.3407*	0.6595	0.6595
Parsitan		-0.3061	-0.3061	-0.03141	-0.03141	0.1089	0.1089	0.1928	0.1928
Reminyl		-	-	-	-	0.1095	0.1095	0.7929	0.7929
Taxotere		-0.1211	-0.1211	0.8656	0.8656	-0.3389	-0.3389*	0.4222	0.4222

Over-the-Counter Drugs

Drug	EAS#	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
				5th=-0.954	10th=-0.677	5th=-0.517	10th=-0.515	5th=-0.714	10th=-0.245
Relievol Allergy Sinus Caplets Extra Strength		-	-	-1.129*	-1.129*	-1.011*	-1.011*	-0.5944	-0.5944*

#: EAS: Estimated accuracy score.

*: EAS less than the credible limit.

Table 9. Concordance Correlation Coefficient between Observed and Predicted Accuracy

Types of Drugs	**Exp1**	**Exp2**	**Exp3**	**Exp4**
Prescription Drugs	0.978	0.909	0.929	0.985
Over-the-counter Drugs	N/A	0.991	0.991	0.966

Table 10. Results of 10-fold Cross Validation for Prescription Drugs and 5-fold Cross Validation for OTC Drugs

Prescription Drugs (10 fold cross validation)					
	# of Control Drugs	# of Training Drugs	# of Validation Drugs	Sensitivity	Specificity
Experiment1	80	72	8	90.00%	71.00%
Experiment2	80	72	8	79.00%	100.00%
Experiment3	60	54	6	91.10%	90.00%
Experiment4	60	54	6	93.00%	78.00%

Over-the-Counter Drugs					
	# of Control Drugs	# of Training Drugs	# of Validation Drugs	Sensitivity	
Experiment2	17	14	3	94.00%	100.00%
Experiment3	17	14	3	82.00%	75.00%
Experiment4	12	10	2	92.00%	77.00%

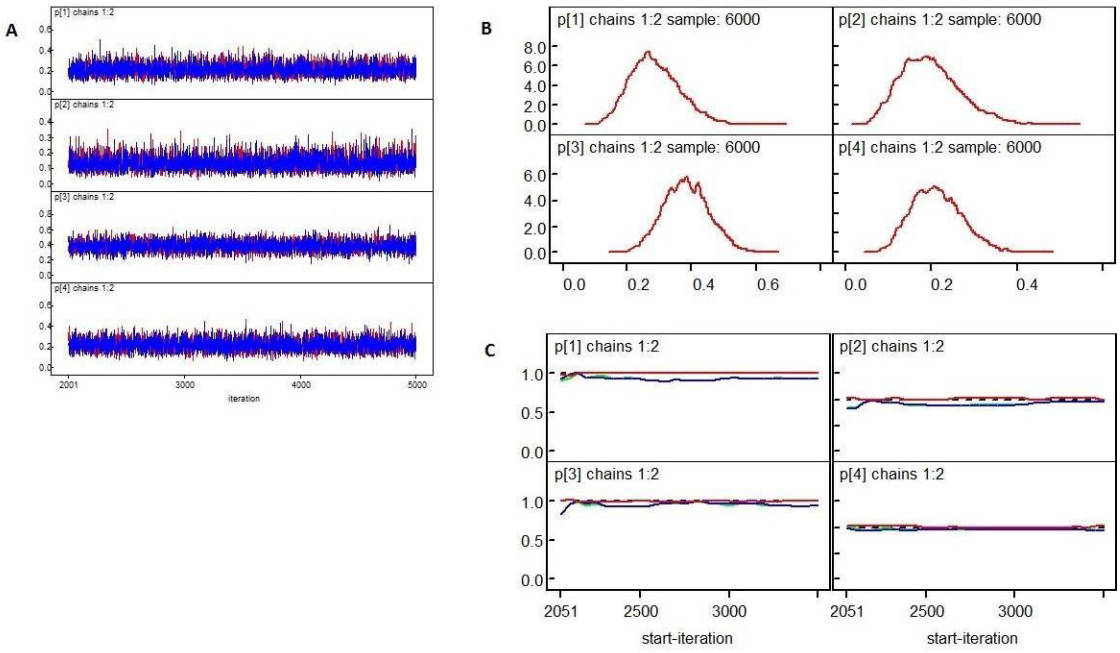

Figure 1. Covergence diagnostics: trace (A), density (B), and BGR (C) plots.