Descriptive Analysis of the Drug Name Lexicon

Bruce L. Lambert, Ph.D.[a, b]

Ken-Yu Chang, B. Pharm., MPH[a]

Swu-Jane Lin, B. Pharm., MPH[a]

[a]Department of Pharmacy Administration

[b]Department of Pharmacy Practice

University of Illinois at Chicago

2/8/00 1:10 PM

Short Title: Drug Name Lexicon

Address: Department of Pharmacy Administration,

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

Phone: 312-996-2411

Fax: 312-996-0868

Email: lambertb@uic.edu

Word Count: 2212 (main body text), 3955 (entire manuscript)

Acknowledgements

This research was supported in part by the National Patient Safety Foundation, the Latiolais Leadership Grant Program, and the Campus Research Board of the University of Illinois at Chicago. The authors acknowledge the helpful assistance of Dan Boring, Mike Cohen and the staff of the Institute for Safe Medication Practices, Ruta Freimanis, Sanjay Gandhi, Keith Johnson, David Lambert, Robert Lee, and Don Rucker. Requests for reprints should be sent to Dr. Lambert.

Abstract

The complexity of the drug use process is managed in part by developing systematic nomenclature for drugs. This nomenclature is cataloged in a variety of drug information databases. However, answers to simple questions about the whole population of brand and generic drug names are not easily obtained. This paper provides a descriptive analysis of the drug name lexicon, with a primary (though not exclusive) emphasis on drugs marketed in the United States. Using the techniques of computational lexicography, one large database of trademark names (the U. S. Patent and Trademark database) and one large database of non-proprietary names (the *USP Dictionary of USAN and International Drug Names*) were analyzed. Results describe a variety of distributional characteristics of drug names, including the number of characters per name, the number of syllables per name, and the number of words per name. Distributions of pairwise similarity and distance scores for a large sample of names are provided, as are lists of the 25 most common initial and terminal bigrams and trigrams. The information should be of interest to trademark attorneys, patient safety advocates, regulators, and students of drug nomenclature.

Key words—drug nomenclature, medication errors, similarity, description, trademark, generic

INTRODUCTION

The process of discovering, designing, developing, approving, marketing, dispensing, and administering drugs is complex and prone to error (1,2). One strategy for managing complexity and minimizing error has been to develop standard nomenclature for drugs. Drug nomenclature falls into two broad categories: proprietary (i.e., brand, trademark), and non-proprietary (i.e., generic) (3). These names are listed in a variety of familiar references, including (among many others) the U.S. Pharmacopeia's *Drug Information for the Health Care Professional, Vol. I*, the *U. S. Pharmacopeia Dictionary of USAN and International Drug Names*, and international category 5 of the U. S. Patent and Trademark Office's (USPTO) Trademarks Registered database (4-6).

As useful as these and other references are, however, a great deal of important descriptive information about the drug lexicon is still not readily available. In fact, when we began our research on drug name confusion errors several years ago, there were several questions about the drug name lexicon that were surprisingly difficult to answer. Basic questions involved the number of existing brand and generic names, the number of letters and syllables in an average name, the similarity characteristics of the names, and so on. In some cases, we could find no answers to these questions at all. In other cases, there were multiple, contradictory answers. We have now had the opportunity to work intensively with several drug name databases, and have begun to answer some of these questions ourselves. The purpose of this study was to produce a descriptive analysis of the drug name lexicon, with a limited (though not an exclusive) focus on names used in the United States. In particular, we sought to answer the following research questions:

RQ1: How many brand and generic drug names are there in the US?

RQ2: How similar are drug names to one another?

RQ3: How long is the average drug name? How many letters are in the average drug name? How many syllables and words are in the average drug name?

RQ4: How do brand names compare to generic names in terms of their numbers, similarity, lengths, etc.?

RQ5: What are the most common two- and three-letter subsequences at the beginning and ending of drug names?

Although it was not possible to produce unequivocal answers to all of these questions, the results reported here should be useful to others interested in drug nomenclature.

MATERIALS AND METHODS

Broadly speaking, the techniques of computational lexicography were used to answer the questions posed above. That is, we used a computer to analyze large, electronic databases of drug names.

Drug Name Databases

Non-proprietary names were drawn from an electronic version of the 1998 *USP Dictionary of USAN and International Drug Names* (4). The *USP Dictionary* is distributed in Dialog B database format. Using Unix®-based text processing tools, we extracted the name and date fields from the overall database and saved the resulting file as plain text. Brand names were taken from the February 1999 update of the USPTO's Trademarks Registered database, available on CD-ROM (7). Within the trademark database, we only analyzed names from US category 018 (Medicines and Pharmaceutical Preparations) and international category 005 (Pharmaceuticals) (6). From these categories, we extracted the word mark and registration date fields from the CD-ROM database and saved the resulting file as plain text (ASCII). Some names appeared in a

database more than once (e.g., *Feen-a-Mint*® appeared in the trademark database 3 times). Duplicates were deleted before further analyses were done.

Analysis Plan

An initial data file was created using Lisp computer programs written by the first author. For each name, this file contained a source identifier (USP Dictionary, USPTO category 018 or international category 005), the year of registration, the number of letters per name, the number of syllables per name, and the number of words per name. This data file was subsequently analyzed by SPSS-PC and Microsoft Excel.

Length

Although it was straightforward to compute the number of characters (i.e., letters) in each name, the programs for computing the number of syllables in a name and for computing the number of words in a name deserve further description. In English, the number of syllables in a word is closely (but not perfectly) related to the number of vowels in a word. A simple program that predicts the number of syllables in a word based on the number of vowels would probably achieve 70%-80% accuracy. The program we used counted vowels, and took into consideration a wide variety of exceptions that fool the vowel-counting strategy (e.g., double and triple vowel sequences, silent vowels, “y” as a vowel, etc.). The program to compute the number of words in a name was also simple in the default case, where words are separated by spaces, but the task of separating names into distinct words was complicated by the presence of several non-alphabetic delimiters (e.g., hyphens, slashes, etc.). The program we used was designed to handle the default and the exceptional cases.

Similarity

When examining the distribution of pairwise similarity scores, we randomly selected 16,641 pairs of one-word brand names and 16,641 pairs of one-word generic names. With this number of pairs, estimates of the percentage of pairs at any given similarity level had 99% confidence intervals of ± 1%. To compute similarity, we used measures which have been described in detail elsewhere (8,9). Specifically, we used n-gram and normalized edit distance measures. N-gram measures compute similarity by breaking words down into n-letter subsequences of adjacent letters and then examining the number of common subsequences between two words (8,10). To accomplish this, first the unique n-grams (i.e., n-letter subsequences) in each name were generated. For example, for the drug *Tylenol*®, the unique trigrams were {*--t, -ty, tyl, yle, len, eno, nol*}. In this case, two spaces were added to the beginning of each word to increase sensitivity to similarity at the beginning of words (9). For the drug *atenolol*, the unique trigrams were {*--a, -at, ate, ten, eno, nol, olo, lol*}. Trigram string similarity was defined by the Dice coefficient:

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

where A was the number of unique trigrams in the first word, B was the number of unique trigrams in the second word, and C was the number of common trigrams between the two words (11). Thus, the trigram string similarity between the names *Tylenol*® and *atenolol*, which have two trigrams in common (*eno* and *nol*), was (2 * 2)/(8 + 7) = .27.

Edit distance refers to the number of edits (i.e., insertions, deletions, and substitutions) required to transform one word into another (10,12). For example, to transform *Ambien*® into *Amen*®, one must delete the *b* and the *i*, so the edit distance between *Ambien*® and *Amen*® was equal to 2. Normalized edit distance takes into account the length of the words being compared. Normalized edit distance is equal to the edit distance divided by the length of the longer of the two words being

compared (9). The normalized edit distance between *Ambien*® and *Amen*® is 2/6 = 0.33. In other words, 33% of the letters in *Ambien*® need to be changed in order to transform it into *Amen*®.

Common Beginnings and Endings

To determine the most common initial and terminal bigrams and trigrams among brand and generic names, we simply identified all of the unique two- and three-letter initial and terminal subsequences in the USAN and USPTO databases and tallied the frequency of occurrence of each one.

Results

All Names

Table 1 provides basic descriptive statistics for the two databases we analyzed. There were 8,712 distinct names in the USAN Dictionary and 32,748 distinct names in the combined USPTO database (including international category 005 and US category 018). The modal US drug name had 8 letters, three syllables, and one word. Generic drug names had more letters and more syllables than brand names.

Insert Table 1 about here.

Figure 1 displays the distribution of letters per name. Figure 2 shows the percentage of names with a given number of syllables for both brand and generic names. The percentage of names with a given number of words-per-name is illustrated in Figure 3. Figure 4 shows how many brand and generic drug names were registered per year during the twentieth century. When interpreting Figure 4, it is important to note that 5351 names in the USP Dictionary had no associated date. Names without dates were either the titles of USP monographs or International Nonproprietary Names (INNs) (4).

Insert Figure 1-4 about here.

One-Word Names

Table 1 and Figures 1-4 describe the whole range of brand and generic names, but the data are somewhat misleading in that they contain many multi-word names and even slogans (e.g., *Epicure Sports Cream®*, *A little drop that does a whole lot®*, *Arm & Hammer: The standard of purity®*). It is also important to understand the properties of prototypical, one-word drug name. Thus, several of the descriptive analyses were repeated, this time including only one-word names. Table 2 provides descriptive statistics for 21,687 one-word brand names and 5331 one-word generic names. Figure 5 charts the percentage distribution of normalized edit distances for 16,641 randomly selected pairs of one-word USAN and USPTO names. Figure 6 gives the percentage distribution of trigram similarity scores (with two spaces added to the beginning of each word) for 16,641 randomly selected pairs of one-word USAN and USPTO names. The majority of pairs of USAN names had normalized edit distances greater than 0.8 (i.e., more than 80% of the letters in one name would need to be changed in order to transform it into another name). In terms of trigram similarity (with two spaces added to the start of each word), more than 80% of name-pairs had similarity scores of zero. The distributions of similarity scores for brand names closely mirrored the pattern observed for USAN names, with brand names being, on average, somewhat less similar to one another.

Insert Figure 5-6 about here.

Tables 3 and 4 give the most common initial and terminal bigrams and trigrams for generic and brand names respectively.

Insert Tables 3-4 about here.

Limitations

Those wishing to draw conclusions from the analyses presented here should keep in mind several limitations. First, only two databases of names were examined, and coverage of brand names was limited to trademarks registered in the U.S. Common names (e.g., AZT) were not included, nor were abbreviations. With regard to the number of new names appearing per year, it is important to note that many of the names in the *USP Dictionary* had no date associated with their first appearance. Thus the number of generic names appearing per year has almost certainly been underestimated. The similarity score distributions pertain only to look-alike similarity. Measures of sound-alike similarity have been developed, but they were not used here (9).

Discussion and Conclusions

This project was motivated by our own inability to provide answers to simple questions about the drug name lexicon. The answers to many of those questions have now been provided in the charts and tables above. There are roughly 33,000 trademark names and 9,000 generic names registered in the U. S. The modal one-word trademark drug name in the U. S. has 8 letters and 3 syllables. The modal generic name has 10 letters and 4 syllables. The number of brand names registered each year (3038 in 1998) is increasing rapidly. The number of generic names registered per year is relatively constant (approximately 100) and an order of magnitude smaller than the number of brand names. Compared to brand names, generic names showed much more redundancy in their initial and terminal ngrams, as would be expected from the use of USAN's standardized stem system (4). For example, fully 35% of all generic names end in *–ne*. IN

contrast, the most common terminal bigram in among brand names (-*in*) occurred in only 5% of the names.

The effect of the stem system on overall pairwise similarity is also evident, with generic names exhibiting greater similarity than brand names (see Tables 5 and 6). The greater similarity of USAN names suggests that generic names may be more confusing, on average, than brand names. The costs and benefits of the stem system, which increases average pairwise similarity between generic names, should be considered in light of these findings.

In general, drug names have very few similar 'neighbors'. In a recent study, Lambert found that the trigram similarity score (with two spaces added to the beginning of words) was the best predictor of drug name confusion errors. In that study, confusion errors were predicted whenever the trigram similarity score between two names exceeded 0.11 (9). Fewer than 10% of pairs of brand or generic names had similarity scores exceeding this threshold. So, contrary to some impressions that the drug lexicon is getting too crowded, the evidence presented here suggests that most pairs of drug names are not similar to one another (at least using measures of orthographic or spelling similarity). This suggests that, with appropriate screening, it should be possible to coin new names that have few, if any, dangerously similar neighbors. It would be useful to have similar analyses for a larger database of international trademark names, but in the meantime, these data can be used as a baseline and frame of reference in ongoing discussions about drug nomenclature.

References

1. Cohen M. *Medication errors*. Washington, DC: American Pharmaceutical Association, 1999.
2. Corrigan J, Kohn L, Donaldson M. *To err is human: building a safer health system*. Washington, DC: Institute of Medicine, 1999.
3. Boring D. *The development and adoption of nonproprietary, established, and proprietary names for pharmaceuticals*. Drug Inf J 1997;31:621-34.
4. U. S. Pharmacopeia. *USP dictionary of USAN and international drug names*. Rockville, MD: U. S. Pharmacopeia; 1998.
5. U. S. Pharmacopeia. *USP DI, Vol. I: Drug information for the health care professional*. Rockville, MD: U. S. Pharmacopeia; 1995.
6. U. S. Department of Commerce--Patent and Trademark Office. *Trademark CD-Rom User's Guide*. Washington, DC: Author; 1994.
7. U.S. Patent and Trademark Office. *Trademarks Registered*. 1999 [cited 1999 June 15, 1999]; Available from: URL: http://www.uspto.gov/web/offices/ac/ido/oeip/catalog/tmcassis.htm#TMregistered.
8. Lambert B. *Predicting look- and sound-alike medication errors*. Am J Health-Syst Pharm 1997;54:1161-71.
9. Lambert BL, Lin S-J, Gandhi SK, Chang K-Y. *Similarity as a risk factor in drug name confusion errors: The look-alike (orthographic) and sound-alike (phonological) model*. Med Care 1999;37:1214-25.
10. Stephen GA. *String searching algorithms*. River Edge, NJ: World Scientific; 1994.

11. Frakes WB. *Stemming algorithms*. In: Frakes WB, Baeza-Yates R, eds. *Information retrieval: Data structures and algorithms*. Englewood Cliffs, NJ: Prentice-Hall; 1992, p. 131-60.

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

Table 1

Descriptive statistics for brand (USPTO) and generic (USAN) drug names

	Overall (N = 41,460)			USAN (Generics) (N = 8,712)			USPTO (Brands) (N = 32,748)		
	Letters	Sylls.	Words	Letters	Sylls.	Words	Letters	Sylls.	Words
Mean	11.20	3.90	1.66	14.44	5.27	1.49	10.34	3.53	1.70
Median	9	3	1	12	4	1	9	3	1
Mode	8	3	1	10	4	1	8	3	1
Std. Dev.	6.90	2.30	1.13	6.50	2.13	0.80	6.74	2.20	1.19
Min	1	0	1	3	0	1	1	0	1
Max	80	31	41	77	24	21	80	31	41

Note: Sylls. is an abbreviation for syllables. Some words with no vowels were counted as having zero syllables. The maximum field length for word marks in the USPTO database is 80 characters. Some trademarks were longer than 80 characters, but they were truncated to fit in the USPTO database. These very long ‘names’ were typically slogans and other non-name trademarks. See text for details.

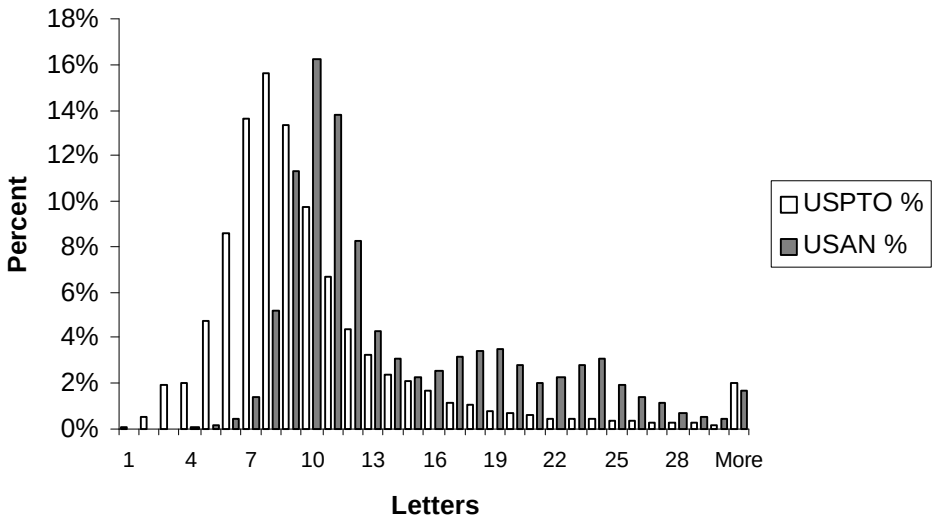


Figure 1. Distribution of name lengths (by number of characters) for brand (USPTO) and generic (USAN) drug names.

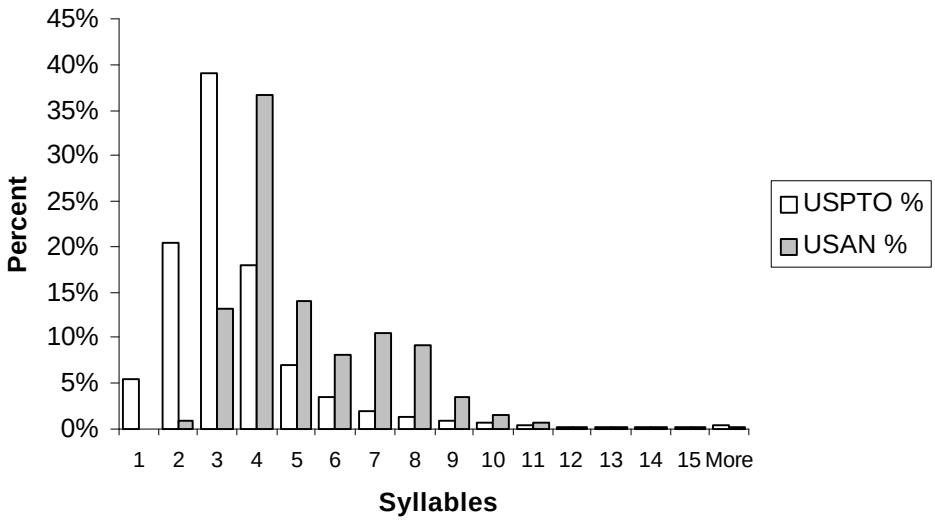


Figure 2. Distribution of name lengths (by number of syllables) for brand (USPTO) and generic (USAN) drug names

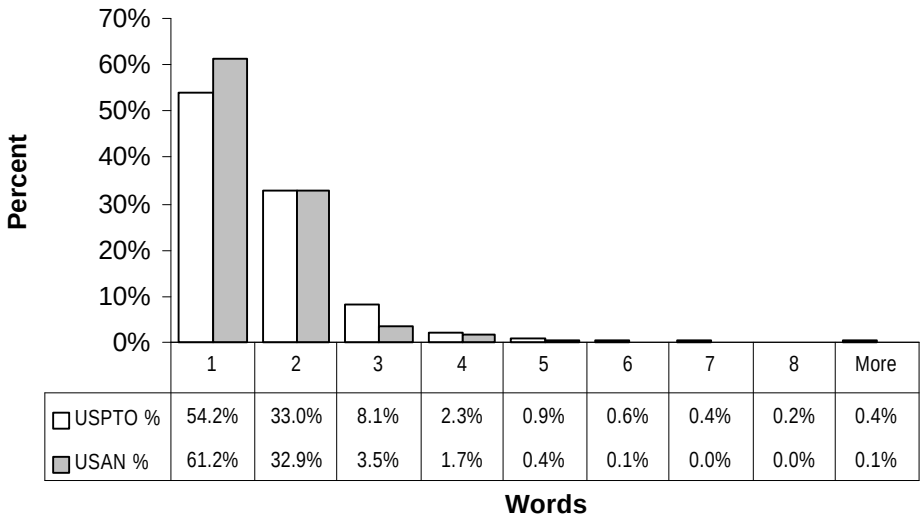


Figure 3. Distribution of name lengths (by number of words per name) for brand (USPTO) and generic (USAN) drug names

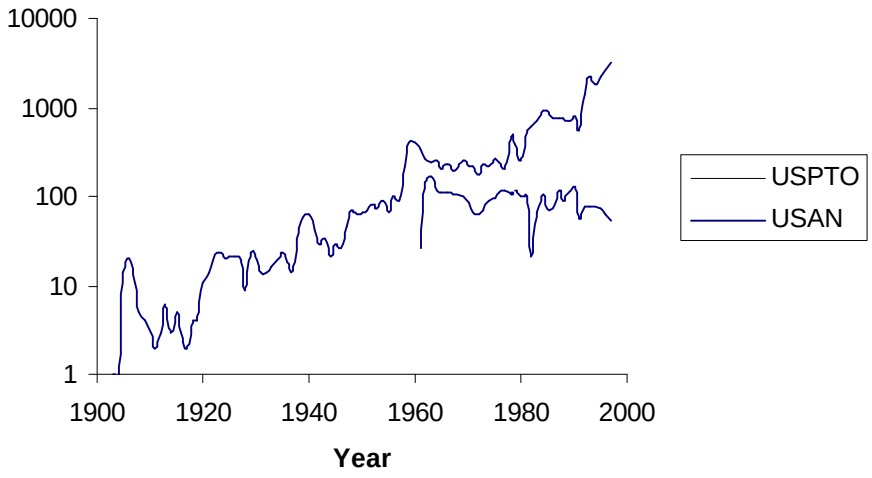


Figure 4. Number of new brand (USPTO) and generic (USAN) drug names per year. The y-axis is on a logarithmic scale. Not every name in the USP dictionary is a USAN name. Many are INN names and some are the titles to USP monographs. INN names and USP monograph titles do not have dates associated with them in the USP dictionary and so they are not represented in this graph.

Table 2

Descriptive statistics for one-word brand (USPTO) and generic (USAN) drug names

	Overall (N = 27,018)		USAN (N = 5331)		USPTO (N = 21,687)	
	Letters	Sylls.	Letters	Sylls.	Letters	Sylls.
Mean	8.14	3.08	10.46	4.02	7.57	2.85
Median	8	3	10	4	8	3
Mode	8	3	10	4	8	3
Std. Dev.	2.36	1.07	1.97	0.83	2.08	0.99
Min	1	0	3	1	1	0
Max	27	11	26	11	27	10

Note: Sylls. Is an abbreviation for syllables. Some words with no vowels were counted as having zero syllables.

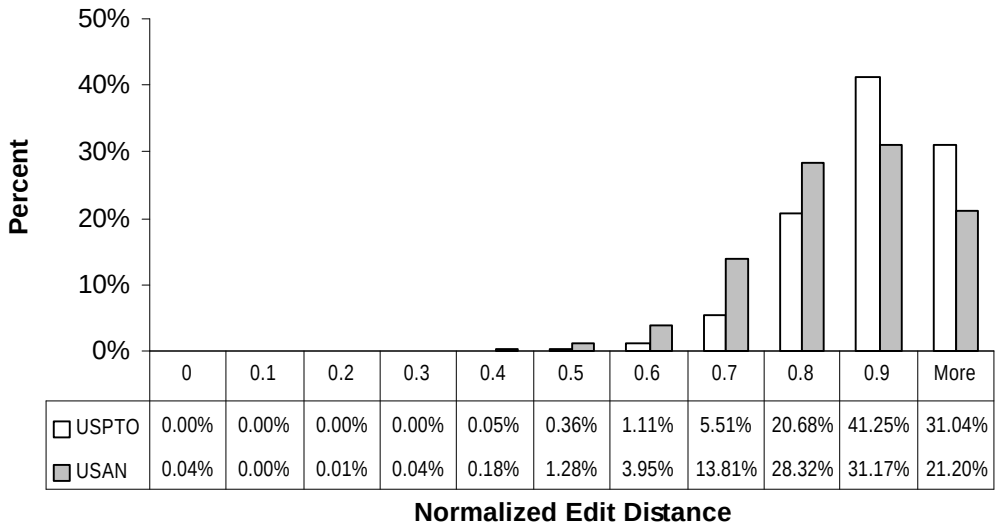
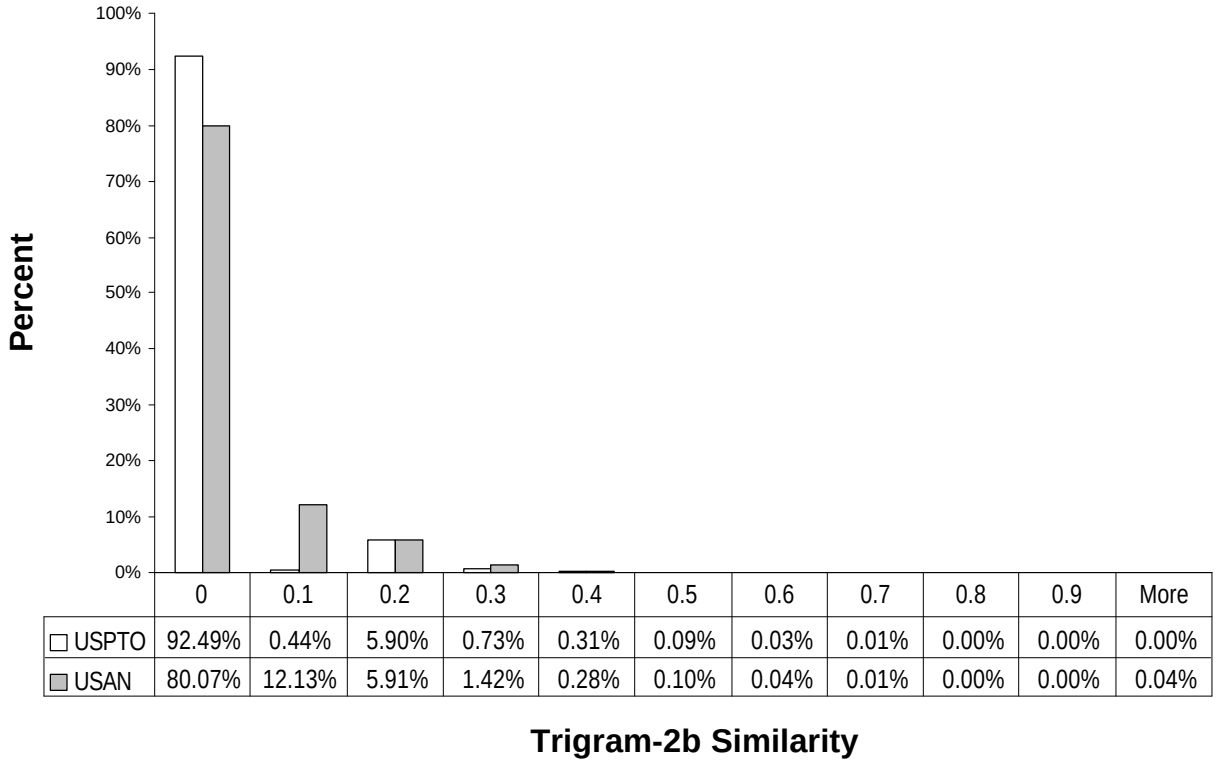


Figure 5. Histogram of pairwise normalized edit distances for 16,641 randomly selected pairs of one-word brand (USPTO) and generic (USAN) drug names. For this histogram, a value is counted within a particular bin if it is equal to or less than the bin value but greater then the previous bin value. For example, 5.51% of USPTO pairwise distances were greater than 0.6 and less than or equal to 0.7.



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	More
□ USPTO	92.49%	0.44%	5.90%	0.73%	0.31%	0.09%	0.03%	0.01%	0.00%	0.00%	0.00%
▒ USAN	80.07%	12.13%	5.91%	1.42%	0.28%	0.10%	0.04%	0.01%	0.00%	0.00%	0.04%

Figure 6. Histogram of trigram similarities for 16,641 randomly selected pairs of one-word brand (USPTO) and generic (USAN) drug names. For this histogram, a value is counted within a particular bin if it is equal to or less than the bin value but greater then the previous bin value. For example, 5.90% of USPTO pairwise similarities were greater than 0.1 and less than or equal to 0.2.

Table 3. Most common initial and terminal bigrams and trigrams in one-word generic (USAN) names (N = 5331)

Initial						Terminal					
Trigram	Freq.	%	Bigram	Freq.	%	Trigram	Freq.	%	Bigram	Freq.	%
Met-	107	2.01%	Me-	228	4.28%	-ine	1191	22.34%	-ne	1844	34.59%
Sul-	90	1.69%	Di-	154	2.89%	-ide	553	10.37%	-in	626	11.74%
Flu-	90	1.69%	Pr-	126	2.36%	-one	504	9.45%	-de	560	10.50%
Clo-	84	1.58%	Su-	121	2.27%	-ate	234	4.39%	-ol	478	8.97%
Pro-	74	1.39%	Ni-	121	2.27%	-ole	222	4.16%	-te	243	4.56%
Tri-	68	1.28%	Tr-	117	2.19%	-cin	198	3.71%	-le	227	4.26%
Ben-	62	1.16%	Pi-	115	2.16%	-nol	110	2.06%	-il	163	3.06%
Car-	57	1.07%	Fl-	115	2.16%	-lin	94	1.76%	-am	128	2.40%
Fen-	54	1.01%	Cl-	115	2.16%	-ene	89	1.67%	-an	120	2.25%
Cef-	49	0.92%	Be-	110	2.06%	-lol	85	1.59%	-en	117	2.19%
Pen-	42	0.79%	De-	109	2.04%	-rol	80	1.50%	-st	73	1.37%
But-	42	0.79%	Ci-	101	1.89%	-tin	69	1.29%	-on	62	1.16%
Bro-	42	0.79%	Ca-	98	1.84%	-fen	65	1.22%	-al	61	1.14%
Phe-	38	0.71%	Am-	95	1.78%	-ane	59	1.11%	-se	58	1.09%
Chl-	36	0.68%	Ti-	92	1.73%	-rin	55	1.03%	-el	42	0.79%
Ami-	36	0.68%	Te-	91	1.71%	-pam	53	0.99%	-ac	35	0.66%
Pir-	35	0.66%	Al-	87	1.63%	-dol	49	0.92%	-me	34	0.64%

Initial						Terminal					
Trigram	Freq.	%	Bigram	Freq.	%	Trigram	Freq.	%	Bigram	Freq.	%
Ace-	35	0.66%	Pe-	86	1.61%	-ast	46	0.86%	-at	34	0.64%
Tol-	34	0.64%	Et-	85	1.59%	-dil	43	0.81%	-im	30	0.56%
Lev-	33	0.62%	Bu-	84	1.58%	-ase	40	0.75%	-ab	29	0.54%
Dex-	33	0.62%	Fe-	83	1.56%	-ril	39	0.73%	-nt	28	0.53%
Cin-	33	0.62%	Ox-	73	1.37%	-tol	38	0.71%	-ex	27	0.51%
Nif-	32	0.60%	Le-	71	1.33%	-sin	31	0.58%	-ir	25	0.47%
Nic-	32	0.60%	Ce-	71	1.33%	-nin	30	0.56%	-ox	24	0.45%
Dim-	30	0.56%	Mi-	69	1.29%	-mab	29	0.54%	-id	19	0.36%

Table 4. Most common initial and terminal bigrams and trigrams in one-word brand (USPTO) names (N = 21,687)

Initial						Terminal					
Trigram	Freq.	%	Bigram	Freq.	%	Trigram	Freq.	%	Bigram	Freq.	%
Pro-	302	1.39%	Pr-	487	2.25%	-ine	395	1.82%	-in	993	4.58%
Bio-	232	1.07%	Co-	394	1.82%	-one	214	0.99%	-ne	881	4.06%
Car-	119	0.55%	Ca-	343	1.58%	-ide	194	0.89%	-ex	606	2.79%
Tri-	110	0.51%	Re-	333	1.54%	-ate	192	0.89%	-ol	565	2.61%
Vit-	108	0.50%	De-	318	1.47%	-ite	168	0.77%	-on	562	2.59%
Pre-	105	0.48%	Me-	315	1.45%	-lex	166	0.77%	-an	508	2.34%
Nut-	101	0.47%	Bi-	295	1.36%	-rin	165	0.76%	-te	488	2.25%
Ult-	92	0.42%	Vi-	256	1.18%	-rol	136	0.63%	-en	366	1.69%
Con-	90	0.41%	Di-	246	1.13%	-gen	135	0.62%	-al	365	1.68%
Per-	87	0.40%	Tr-	243	1.12%	-lin	133	0.61%	-er	328	1.51%
Com-	85	0.39%	Ma-	243	1.12%	-cin	133	0.61%	-il	304	1.40%
Cal-	85	0.39%	St-	239	1.10%	-ard	132	0.61%	-re	290	1.34%
Der-	83	0.38%	Al-	234	1.08%	-are	121	0.56%	-el	285	1.31%
Met-	78	0.36%	Pe-	222	1.02%	-tin	120	0.55%	-st	279	1.29%
Med-	76	0.35%	Se-	219	1.01%	-ent	117	0.54%	-de	268	1.24%
Opt-	75	0.35%	Su-	213	0.98%	-erm	113	0.52%	-ic	267	1.23%
Sta-	73	0.34%	Mi-	211	0.97%	-rex	111	0.51%	-ax	250	1.15%

Initial						Terminal					
Trigram	Freq.	%	Bigram	Freq.	%	Trigram	Freq.	%	Bigram	Freq.	%
San-	70	0.32%	Pa-	201	0.93%	-ron	110	0.51%	-se	242	1.12%
Mic-	70	0.32%	Ch-	192	0.89%	-ase	108	0.50%	-ar	229	1.06%
Res-	68	0.31%	He-	191	0.88%	-ene	106	0.49%	-or	228	1.05%
The-	66	0.30%	In-	186	0.86%	-est	105	0.48%	-ac	220	1.01%
Tra-	66	0.30%	Sa-	183	0.84%	-gel	99	0.46%	-id	219	1.01%
Sup-	65	0.30%	Li-	183	0.84%	-ion	98	0.45%	-et	197	0.91%
Col-	63	0.29%	Te-	168	0.77%	-sol	92	0.42%	-nt	181	0.83%
Max-	62	0.29%	Ne-	168	0.77%	-max	91	0.42%	-rm	175	0.81%