Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at Work

Adam Wright PhD [1, 2, 3], David W. Bates MD, MSc [1, 2, 3]

[1]Brigham and Women's Hospital, Boston, MA, USA, [2]Harvard Medical School, Boston, MA, USA, [3]Partners HealthCare, Boston, MA, USA

Keywords

Data Mining, Electronic Health Records, Statistical Distributions

Summary

Background: Many natural phenomena demonstrate power-law distributions, where very common items predominate. Problems, medications and lab results represent some of the most important data elements in medicine, but their overall distribution has not been reported.
Objective: Our objective is to determine whether problems, medications and lab results demonstrate a power law distribution.
Methods: Retrospective review of electronic medical record data for 100,000 randomly selected patients seen at least twice in 2006 and 2007 at the Brigham and Women's Hospital in Boston and its affiliated medical practices.
Results: All three data types exhibited a power law distribution. The 12.5% most frequently used problems account for 80% of all patient problems, the top 11.8% of medications account for 80% of all medication orders and the top 4.5% of lab result types account for all lab results.
Conclusion: These three data elements exhibited power law distributions with a small number of common items representing a substantial proportion of all orders and observations, which has implications for electronic health record design.

Correspondence to:

Adam Wright, Ph.D.
Division of General Medicine and Primary Care
Brigham and Women's Hospital
1620 Tremont St.
Boston, MA 02120
Tel: (781) 416-8764
Fax : (617) 732-7072
E-mail: awright5@partners.org

Appl Clin Inf 2010; 1: 32–37
doi: 10.4338/ACI-2009-12-RA-0023
received: December 15, 2009
accepted: March 10, 2010
published: March 19, 2010

Citation: Wright A, Bates DW. Distribution of problems, medications and lab results in electronic health records: the pareto principle at work. Appl Clin Inf 2010; 1: 32–37.
http://dx.doi.org/10.4338/ACI-2009-12-RA-0023

Introduction

Many naturally occurring and man-made phenomena demonstrate a non-uniform exponential distribution whereby a small set of common elements in a class represent the bulk of all uses of the class [1]. This phenomenon has variously been referred to as the 80/20 rule [2], the Pareto principle (for continuous data) [3], Zipf's law (for discrete data) [4] and the power-law phenomenon [5]. For example, linguistic researchers studying a large corpus of English language text demonstrated that the word "the" constitutes 7% of all word uses in the corpus, while "to" and "of" each represent another 3%. Indeed, only 135 out of the 50,000 unique observed words are needed to account for half of all word uses, while nearly half of the words in the corpus are used only a single time [6]. This pattern, commonly called Zipf's law, has been observed in a variety of languages including American English, Chinese and the Latin of Plautus [7]. The same phenomenon has been observed in such disparate areas as the population of human settlements [8], distribution of wealth [9], movie rental patterns from Netflix and purchases from Amazon.com [1].

Electronic health records are increasingly widely used in the U.S. [10], although uptake has been limited by a number of issues including ease of use [11]. Improving usability depends on a number of factors, but understanding the distributions of key data elements is likely one of them. Furthermore, clinical data exchange will likely have many clinical benefits, but has been difficult to achieve [12]; enabling this will also require understanding these distributions.

Objectives

We hypothesized that many forms of clinical data are likely to exhibit these properties as well. Medications, problems and lab results represent central data elements in clinical practice and record keeping and they are also some of the most frequently documented and used elements in electronic health record systems. We tested this hypothesis by analyzing data from a widely deployed electronic health record system and characterizing the distributions of medications, problems and lab results.

Methods

We obtained electronic health record data for 100,000 patients seen at least twice in the outpatient setting at the Brigham and Women's Hospital, and with at least one visit between January 1, 2006 and December 31, 2007. The study was approved by the Partners HealthCare Human Subjects Committee. The electronic health record is used for both primary and specialty care, and the underlying sample was drawn from a total population of 839,300 patients (primarily adults, but some neonates). The data were used as recorded in the computer system – no manual validation or aggregation (e.g. into panels or classes) was undertaken.

The data extracted included medications, problems and lab results. We analyzed the data using Microsoft SQL Server and Excel (Microsoft Corporation, Redmond, WA) and used R (the R Foundation, Vienna, Austria) and the zipfR package [13] to characterize their distribution, estimate the distribution's parameters and assess the goodness of fit. For goodness of fit, we used the chi square goodness of fit test with the conservative modifications proposed by Baayen to account for non-constant variance of the distribution elements [14].

Results

Within this population of 100,000 patients in the electronic health record system, there were a total of 272,749 coded problems, 442,658 medications and 11,736,718 lab results for the 100,000 patients in the electronic health record system during this period.

Table 1 The ten most commonly recorded problems, medications and lab results. There were 272,749 coded problems, 442,658 medications and 11,736,718 lab results recorded for the 100,000 patients in our sample.

Problems		Medications		Lab Results	
Problem	**Proportion**	**Medication**	**Proportion**	**Result**	**Proportion**
Hypertension	5.90%	Ibuprofen	2.27%	Hematocrit	2.44%
Elevated Cholesterol	2.52%	Aspirin	2.01%	Potassium	2.40%
Depression	2.31%	Lisinopril	1.87%	Platelets	2.38%
Coronary Artery Disease	2.08%	Multivitamins	1.80%	Hemoglobin	2.37%
Hyperlipidemia	2.00%	Oxycodone	1.60%	White Blood Cell Count	2.37%
Asthma	1.98%	Atorvastatin	1.59%	Mean Corpuscular Volume	2.37%
Gastroesophageal Reflux Disease	1.79%	Albuterol	1.51%	Red Blood Cell Count	2.37%
Breast Cancer	1.58%	Omeprazole	1.49%	Red Blood Cell Distribution Width	2.37%
Diabetes Mellitus Type 2	1.45%	Levothyroxine	1.46%	Mean Corpuscular Hemoglobin Concentration	2.37%
Diabetes Mellitus	1.32%	Simvastatin	1.44%	Mean Corpuscular Hemoglobin	2.37%

(▶Table 1) shows the top ten problems, medications and lab results along with their proportion. Each proportion represents the proportion of total items in that category – i.e. 2.27% of all prescriptions are for ibuprofen. Since each patient may be on multiple medications (an average of 4.42 medications per patient in our dataset), this is distinct from the proportion of patients on ibuprofen (which is 10.1%). It is worth noting that this data is exactly as recorded in the EHR and may reflect imprecisions in coding by users – for example, most of the patients with "diabetes mellitus" on their problem list appear to have diabetes mellitus type 2; however, the recording clinician did not specify the type. Likewise, lab results appear to be a special case: nine of the top ten lab results are components of the complete blood count (results are filed in the EHR independently, even when ordered as part of a panel, which is why these components are presented individually rather than as a group).

(▶Figure 1) shows the cumulative distribution of the three data types. These distributions show the classic Zipfian form. The observed distributions are even more skewed than the classic 80/20 rule (where 20% of items account for 80% of observations). The 12.5% most frequently used problems account for 80% of all problems, while the top 11.8% of medications account for 80% of all medication orders. The distribution of lab results is even steeper: the top 4.5% of lab tests account for 80% of all lab results.

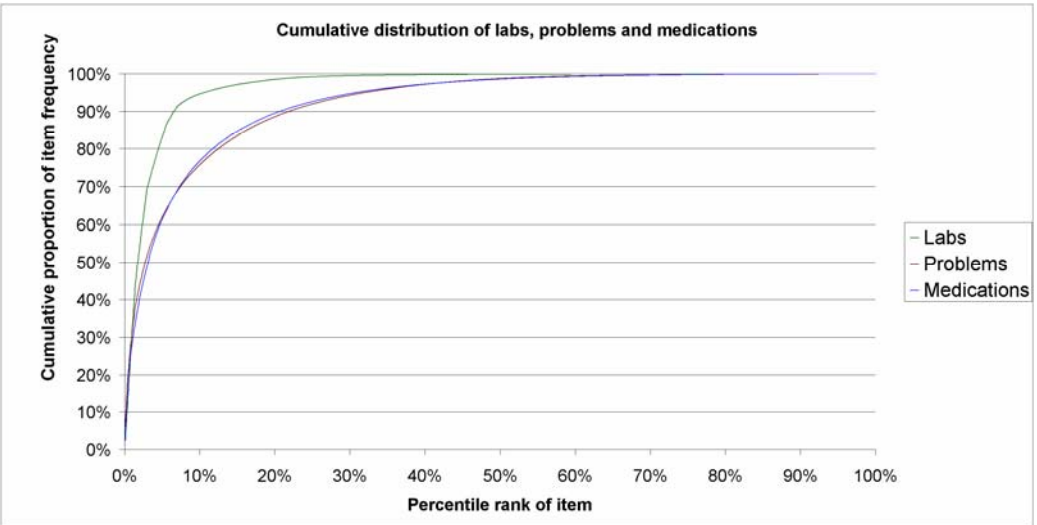

Figure 1 Cumulative distribution of item frequency.

We formally fitted the Zipf distribution to these data using the zipfR package in R and assessed the goodness of fit using the Baayen's modified chi square goodness of fit test. The chi square statistic for problems was 100.22, for medications 31.58 and labs 88.02. Lower values of this statistic indicate better fit, but its exact interpretation is controversial. The best norms come from word frequency analysis in linguistics. The chi square goodness of fit measure for a Zipfian fit of the distribution of words in Lewis Carroll's *Alice's Adventures in Wonderland* is 29.05, while a Zipf distribution fits Arthur Conan Doyle's *The Hound of the Baskervilles* with $\chi^2 = 227.84$ [14]. The Brown corpus (the source of the word frequency statistics in the introduction) can be fit with Zipf's law with chi square of 819.64. Compared to these sources (considered classically Zipfian), Zipf's law appears to fit our medication, lab and problem data quite well.

Discussion

As observed in many other areas, the distribution of problems, medications and lab results in our electronic medical record appears to follow a Zipfian distribution with frequently documented or ordered items accounting for a substantial majority of all observations and orders.

The distribution of problems, medications and lab results we have described has a number of implications, particularly in any setting where resources are constrained or space is limited. For example, physicians may want to preferentially stock handouts for common conditions or samples of common medications. It also has implications for designers of electronic health record systems. Screen space is often limited in these systems, so it may make sense to concentrate on displaying the most common choices. Building clinical decision support systems and order entry and documentation templates is time consuming and expensive [15]. System developers may consider focusing their efforts on building top quality interfaces for all of the most common problems, medications and lab results. They may also choose to develop specialized content for items in the less frequently used long tail, particularly where there are quality, safety or risk management issues (e.g. a potentially fatal drug interaction between two uncommonly used medications). For those building clinical data exchanges, this information is also important. If it is possible to begin by ensuring that reliable exchange of the most frequent data types, much of the value may be achieved. However, some infrequently collected data may be especially valuable, for example the results of cardiac catheterizations. In some instances, frequency may even be inversely associated with value (there is low clinical value, for example, of having a long string of MCHC's).

It is worth noting that fitting distributions to data is an inexact and subjective process. In many cases, including ours, a variety of distributions fit our data well, and some more generalized distri-

butions (such as the generalized Gamma distribution) actually fit the data better, although they have more parameters and are generally able to be fit to more kinds of data (indeed the generalized Gamma distribution is the parent of the exponential, chi-square, Erlang and Maxwell-Boltzman distributions, so its ability to fit a variety of data is quite extensive). We preferred the Zipfian distribution because it is relatively parsimonious (having only an exponent and size parameter), because there is a reasonable theoretical basis that suggests why problem, medication and lab result data might follow the distribution, and because it is widely used for assessing the distribution of rank data such as ours. However; even if another distribution were chosen, the implications remain the same: problems, medications and lab results are heavily skewed toward common items.

These results have limitations. They were drawn from a single electronic health record in a single region which has been used for some time, and the results might differ somewhat with a different record or in a different area. Further, the institution under study serves a primarily adult population is a tertiary care center (though it is not a specialty hospital and does provide a full spectrum of medical and surgical care), so the distribution of items used at other sites may differ. For example, an orthopedic hospital might have relatively fewer items in its universe (because of the limited scope of care it provides) but may use some items which are rare in a general hospital more frequently. A direction for future research might be to carry out the same analysis at other hospitals (and with other item types) to see if the pattern holds.

Conclusion

Problems, medications and lab results appear to follow a fairly steep Zipfian distribution, with a relatively small number of items predominating in each class. It may be useful to focus a variety of efforts on these most commonly occurring items when resources are constrained.

Clinical Implications

These results have important implications for both clinical and informatics practitioners. Since a relatively small number of distinct problems, lab results and medications comprise the large majority of usage, practitioners may consider concentrating effort and resources on these common items. This is likely to be particularly true for decision support, where content should be prioritized according to frequency, with content initially targeted at common items before progressing to relatively rarer items.

For example, Lovis *et al.* described creating an order parser whose knowledge base took into account order frequency [16] and in prior work we described an approach for automated development of order sets and corollary orders based on frequently co-occurring orders [17]. Similar approaches have also been used outside of informatics by, for example, toxicologists making recommendations on antidotes to stock in hospitals based on (among other factors) frequency of use [18] and space planners determining where to store small parts in a warehouse [19].

At the same time; however, focusing solely on common items, though efficient, may lead to unintended consequences, such as poor user experience when performing (or even inability to perform) uncommon actions. Also, for certain types of decision support, users may need the most support when carrying out uncommon and unfamiliar tasks, so rare items should not be entirely neglected. The optimal balance between the efficiency of focusing on common items and the importance of being comprehensive is difficult to strike and has not yet been studied empirically – validating these proposed implications will be an important area for future research.

Conflict of Interest

Neither Dr. Wright nor Dr. Bates have any conflicts to report. Dr. Wright had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Human Subjects Review

The study was reviewed and approved by the Partners HealthCare Human Subjects Committee.

References

1. Anderson C. The long tail. New York: Hyperion; 2008.
2. Trueswell RL. Some Behavioral Patterns of Library Users: The 80/20 Rule. Wilson Libr Bull. 1969;43(5):458-61, 69.
3. Bookstein A. Informetric distributions, part I: Unified overview. J Am Soc Inform Sci. 1990;41(5):368-75.
4. Reed WJ. The Pareto, Zipf and other power laws. Economics Letters. 2001;74(1):15-9.
5. Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology. Proc Conf on Applications, Technologies, Architectures, and Protocols for Computer Communication. 1999:251-62.
6. Kucera H, Francis WN. Computational analysis of present-day American English. Providence: Brown University Press; 1967.
7. Zipf GK. The psycho-biology of language; an introduction to dynamic philology. Boston,: Houghton Mifflin Company; 1935.
8. Gabaix X. Zipf's Law For Cities: An Explanation. Quarterly Journal of Economics. 1999;114(3):739-67.
9. Pareto V. Cours d'économie politique. Geneva: Librairie Droz; 1964.
10. DesRoches CM, Campbell EG, Rao SR, et al. Electronic health records in ambulatory care--a national survey of physicians. N Engl J Med. 2008 Jul 3;359(1):50-60.
11. Bates DW. Physicians and ambulatory electronic health records. Health Aff (Millwood). 2005 Sep-Oct;24(5):1180-9.
12. Adler-Milstein J, McAfee AP, Bates DW, Jha AK. The state of regional health information organizations: current activities and financing. Health Aff (Millwood). 2008 Jan-Feb;27(1):w60-9.
13. Evert S, Baroni M. zipfR: Word frequency distributions in R. Proc 45th Ann Meeting of the Association for Computational Linguistics. 2007:29-32.
14. Baayen RH. Word frequency distributions. Dordrecht ; Boston: Kluwer Academic; 2001.
15. Johnston D, Pan E, Walker J, Bates DW, Middleton B. Patient Safety in the Physician's Office: Assessing the Value of Ambulatory CPOE. Oakland, CA: California Healthcare Foundation; 2004.
16. Lovis C, Chapko MK, Martin DP, et al. Evaluation of a command-line parser-based order entry pathway for the Department of Veterans Affairs electronic patient record. J Am Med Inform Assoc. 2001 Sep-Oct;8(5):486-98.
17. Wright A, Sittig DF. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. AMIA Annu Symp Proc. 2006:819-23.
18. Dart RC, Borron SW, Caravati EM, et al. Expert consensus guidelines for stocking of antidotes in hospitals that provide emergency care. Ann Emerg Med. 2009 Sep;54(3):386-94 e1.
19. Bartholdi JJ, Hackman ST. Allocating space in a forward pick area of a distribution center for small parts. Institute of Industrial Engineers Transactions. 2008;40(11):1046-53.