# Correlating Web Usage of Health Information with Patient Medical Data

**Bradley A. Malin**
**Center for Automated Learning and Discovery, School of Computer Science**
**Carnegie Mellon University, Pittsburgh, Pennsylvania**

*The number of online websites providing health-related information to the general public increases daily. Yet, it is relatively unknown as to how individuals in the general population access information with respect to their own medical status. In this study, clickstream data from an online health information website is analyzed with respect to the user's health insurance claims. The relationships are assessed through the construction and study of intersecting sets of ICD-9 codes in visited web pages and claims made. Results demonstrate that approximately 15% of patients use health information on the web in exact agreement with their medical status. In addition, almost 40 codes were found to be indicative of temporal aspects in user behavior with respect to physician visits.*

## INTRODUCTION

The use of the World Wide Web for providing information with respect to health problems has grown dramatically in recent times. Initially in the health environment, the internet was utilized as a medium for disseminating information on specific disorders maintained by individual web users. However, the provision of health information has turned down a more commercial route. Currently, there exist over 70,000 websites distributing health-related information[1], ranging in scope of information content from disease support groups and homepages to large-scale information brokers, such as *WebMD*. While the number of websites continues to grow and information content on already existing pages expands, it is important to have an understanding of how this information is actually utilized by the general patient population with reference to their medical status.

A recent survey of the consumer health informatics society established several important research topics.[2] One such topic is the evaluation and education of consumer health information from online medical websites. Studies have been confined to the analysis of specific disease populations or analysis of the quality of information accessed. Key aspects of the relationships between actual patient medical status and web usage of health information have been neglected. For example, do patients search for information relevant to their own medical status? And if so, do patients attempt to gather related information before or after the date of diagnosis? The answers to such question can facilitate in understanding the online behavior of individuals searching for health information related to their medical status. This research attempts to examine several areas of user behavior in an online health environment. First, we determine which diagnoses are accessed in concordance with an individual's medical status. Second, the quantity of information accessed is analyzed with respect to the number of webpages that patients view. Finally, the temporal aspect of online behavior is analyzed to determine when users access information relative to their physician visits.

## BACKGROUND

Researchers believe that the World Wide Web possesses the potential to provide the public with unprecedented involvement in their own health care.[3] Prior research in the field has focused on specific disease cohorts, evaluation of the quality of health information websites, and offline surveys of patient populations. Encouraging results have been observed in several studies, confirming the ability of health information systems for educating users outside of the health-care profession. It was found by Gustafson, et. al., that the use of a computer-based support system can improve the quality of life for a HIV-positive population.[4] Patients with computer assistance reported an increased quality of life and less time in ambulatory care visits, making phone calls, and fewer hospitalizations. However, such systems are tailored to specific-disease groups and it is difficult to provide specialized systems to a patient population afflicted with a wide range of diseases.

Surveys, such as those conducted by the PEW Internet and American Life Project[5], have found that about fifty percent of individuals who seek health information on the web are better able to care for themselves and improve the way they get medical information. A larger proportion reported that the information online influenced how they treated a disease, questions to ask their physician, or seek a second opinion. Yet, these results are survey based and not objective analysis of actual information accessing.

One research community has focused on the evaluation of the quality and reliability of online health

information systems.[6,7] The findings of Hernandez-Borges et.al. suggest websites with considerable numbers of visitors and frequent updating indicate quality health information websites.[8] Yet, little has been done to quantify how an individual accesses information at a trusted information site or how information is accessed with respect to medical status.

Much of the difficulty in studying the relationships between web usage and patient health status stems from the inability to link the IP addresses of online users with actual identities. Fortunately, the commercial appeal of online health information provision has led to the collaboration of the two industries of online health websites and health insurance. It is from such a large venture that our data has been provided.

## METHODS

### Materials

Datasets used for this study were procured from a joint venture between a large health insurance provider and Personal Path Systems, Inc, an online medical information website. The data covers the three-year period 1999-2001 and consists of 66,422 individuals with health insurance coverage in a single state. The insurance claims and web usage data were stored in two separate databases, which we refer to as MED and WEB. The MED database contains 1,661,771 distinct claims, where each claim is a tuple that includes the fields $\{policy, dx_1,...,dx_k, date\}$, where *policy* is the insurance policy ID (unique to the policy), $dx_1,...,dx_k$ are a variable number $k$ of ICD-9 codes assigned to the policy at the date of claim *date*. The WEB database contains 3,930,651 webpage accesses for 17,348 webpages by the policy-holding population. The useful attributes of this database are $\{policy, docID, date\}$, where *docID* is a unique identifier for a webpage accessed at a particular *date*.

The webpages have been labeled with ICD-9 codes generalized to the first three characters of the code by resident experts. The ICD-9 codes in the MED database consist of four and five characters. Since ICD-9 codes have a coherent generalizable relationship up to two characters of generalization (i.e. xx*, where * is a wildcard), the ICD-9 codes of the MED database are generalized to three characters as well.

The clickstream and claims data was supplied in accordance to the safe harbor provisions specified by HIPAA.[9] Therefore, all sensitive information was removed and the data were de-identified. As such, this study serves as an example of how de-identified information can remain useful for research purposes.

### WC Vectors

From the MED and WEB databases, vectors of information were constructed for each policy. The vectors are represented as the sequence:

$$W_1C_1W_2C_2\ldots W_kC_k$$

where, $W_i$ is the vector $[<w_{i1}, time_{i1}>,\ldots,<w_{im}, time_{im}>]$, which is the ordering on *time* of the webpages accessed by the policy between two claim dates, and $w$ is $\{icd9_1,..,icd9_{1l}\}$, the set of generalized ICD-9 labels for this webpage. $C_i$ is $\{<c_{i1}, time_{i1}>,\ldots,<c_{in}, time_{in}>\}$, the set of claims for a claim dates between webuse. Henceforth, this data structure is referred to as a WC vector.

### Access With Respect to Medical Status

One of the first questions that we answer is "What diagnosis codes do individuals access that are directly indicative of their medical status as determined by a physician?" Furthermore, we need to determine how this information relates to the information that all users of the website access. To answer this question, we first determine the patients that possess a direct relationship between their insurance claim and webuse data. Such a relationship is defined as the intersection of ICD-9 codes in terms of the WC vector:

$$I=\left(\bigcup_i (icd9 \in W_i)\right)\cap\left(\bigcup_j (icd9 \in C_j)\right)$$

From the intersections we can count the number of times that each diagnosis code occurs. Each individual provides a maximum of one count for each diagnosis code. The rank ordering of the codes is constructed by ranking the code with the largest number of intersecting individuals highest. This rank ordering can be compared with the ordering of the codes accessed by all individuals utilizing the website. The counts in the latter ordering are provided all individuals accessing the codes, both individuals with and without the code in their intersection.

### Overall Use of Health Website

The second question we answer pertains to the quantity of information a typical user of the health information website accesses. For this analysis, information is defined as the number of distinct pages that an individual accesses. This analysis is performed for both the general population and the population of individuals with non-null intersections. Such an analysis helps to characterize how much of the website is actually being used by each individual.

### Temporal Relationships

The previous analyses provide insight into how much information the general user population accesses, as well as which diagnoses characterize the non-null intersecting users. However, when studying clickstream data, there is a temporal aspect to webpage

accesses, and as such, a finer granularity of how users access information can be established.

Patients that use the web in order to gather information with respect to their medical status may do so in one of several ways. We characterize these different methods of information accessing and gathering as temporal online user behaviors. There are two general ways in which an individual can search for online health-related information with respect to their medical status. The first user behavior is characterized by the user who searches for information related to their symptoms before they visit the family physician or hospital. This type of behavior we label diagnosis-predicting (DP). The other behavioral type is exemplified by the patient that follows up a doctor visit for confirmation and/or clarification of received diagnoses. This type of behavior we label as diagnosis-following (DF). The relationship between user type and the WC vector is depicted in Figure 1.

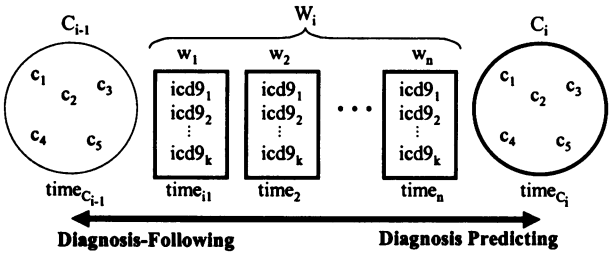


**Figure 1. First-order diagnosis-following ($W_i \rightarrow C_{i-1}$) and diagnosis-predicting web accessing ($W_i \rightarrow C_i$).**

To determine which diagnosis codes are specific to user behavior type, we compare the intersection sets at varying time periods from a physician visit. For each diagnosis, *d*, we compute the set of individuals with a non-null intersection at varying time points from each individuals claim dates. The dates chosen for this study were 1, 2, 3, 7, 14, 30, and 60 days from the claim date. The time period includes all intersections between the claim date to the beginning/end of the period. To distinguish between the DP and DF user type, the set of intersecting individuals is partitioned based on whether or not the web access for *d* occurred before or after the claim. Users with an access date prior to the claim are considered DF for *d* and users with an access after their claim are considered DP.

Hypothesis testing is performed for each equidistant time period (i.e. –1 vs. 1 or –60 vs. 60) to determine if there exists a significant difference in the number of individuals intersecting on the diagnosis. The proportion test is ideal to evaluate the relative proportions of two user types as it is based on the pooled decisions (to access before or to access after a claim) of individuals. The definition of the test follows. Let $n$ equal the number of individuals with an intersection for diagnosis *d* within a defined period of time. Let $p_p$ and $p_f$ be the proportion of intersecting individuals who reside in the DP and DF user types, respectively. The expected proportion $p$ is the average of $p_p$ and $p_f$, which is $(p_p+p_f)/2$. Our null hypothesis is that $p_p$ is equal to $p_f$ and the alternative is that $p_p$ is not equal to $p_f$. We compute the significance statistic as:

$$Z = \frac{p_p - p_f}{\sqrt{2p(1-p)/n}}$$

When $n$ is less than or equal to 40, the P-value is computed from the significance statistic using the t-distribution for $n$-1 degrees of freedom. Otherwise, the standard normal is employed. If multiple time periods are found significant for a particular diagnosis, the most significant time period is chosen.

## RESULTS

Of the 66,422 policies, non-null intersection sets were found in 9652 policies (~14.53%) with ICD-9 codes generalized to 3 character codes. Of the intersecting population, 483 of 881 accessed ICD-9 codes had a minimum of one individual intersecting on the code. The number of individuals intersecting for each code was computed and the top ten most accessed codes, by intersecting individuals, is provided in Table 1. For the most part, the disorders listed in Table 1 appear to be chronic diseases or severe disorders, such as problems related to the back, bone, heart, and cancer.

| Rank | Code | Description | # of Intersects |
|---|---|---|---|
| 1 | 401 | essential hypertension | 805 |
| 2 | 250 | diabetes mellitus | 684 |
| 3 | 272 | disorders of lipoid metabolism | 677 |
| 4 | 780 | general symptoms | 600 |
| 5 | 724 | other & unspecified disorders of back | 494 |
| 6 | 715 | osteoarthrosis & allied disorders | 487 |
| 7 | 414 | other forms of chronic ischemic heart disease | 445 |
| 8 | 733 | other disorders of bone & cartilage | 374 |
| 9 | 722 | intervertebral disc disorders | 343 |
| 10 | 185 | malignant neoplasm of prostate | 314 |

**Table 1. The top ten accessed codes by individuals with non-null intersections.**

The information in the intersection was compared with the overall ICD-9 accesses of the general population. This information was compared at different cutoff points in the ranks. For example, at rank 10, the number of codes that were found to be the same in both rank lists before the cutoff is provided as the fraction of

same codes. The analysis is provided in Table 2. Notice, that within the first ten most accessed codes for each population, one half of the codes are the same. These codes are prostate cancer, osteoarthritis, chronic heart disease, diabetes, and general symptoms. This result is not a confounding feature of the intersecting population, since the number of accesses of these and most codes in the general population is greater than a factor of four to the intersecting population. However, an interesting difference between these populations relates to the dissimilar codes. In the first 20 ranks for the general population, there are almost 10 codes accessed represents varying types of neoplasms, where in the intersecting population there are only two codes (breast and prostate).

| Rank | Fraction of Same Codes |
|---|---|
| 10 | 0.5 |
| 25 | 0.32 |
| 50 | 0.22 |
| 75 | 0.253333 |
| 100 | 0.35 |
| 125 | 0.488 |
| 150 | 0.6 |
| 200 | 0.66 |
| 300 | 0.686667 |
| 483 | 0.778468 |

**Table 2. Rank comparison of codes accessed by the intersecting population and the general population.**

In addition, it should be noted that as the number of codes considered increases, the number of codes that are the same in both lists decreases to half of the original number. This suggests that the codes at such ranks for the intersecting population have the most disproportionate accesses compared with the general population. Consider code 786, which represents "symptoms involving respiratory system & other chest symptoms". This code ranks as 14 for the intersecting population, but only at 128 in the general population.

Analysis of the number of pages accessed by the population is summarized in Figure 2. The graph demonstrates that there exists a log-log relationship between the number of individuals and the number of pages that they access. A linear fit of this plot results in a correlation coefficient of almost 0.9, which suggests that there exist a large percentage of individuals who access a small number of webpages. The fact that the leftmost tail is the least correlating suggests that individuals possess a similar likelihood to access any number of less than approximately five webpages. The pages that such individuals access, usually consist of general information and introductory pages that lie in close proximity to the front page of the website.

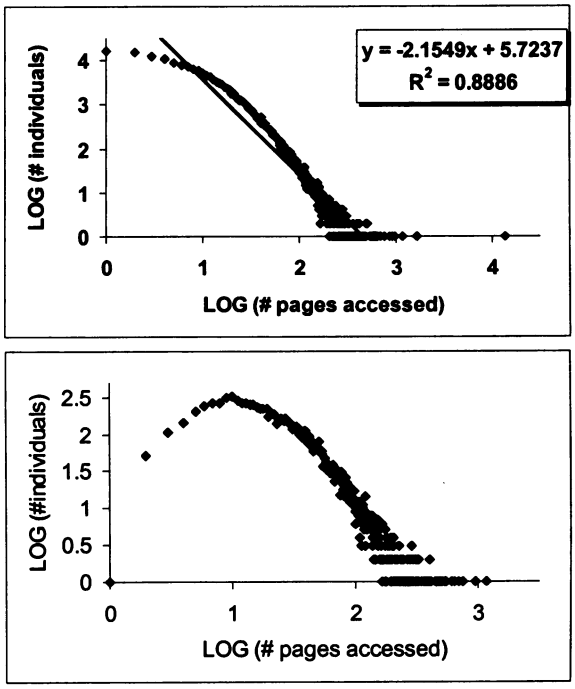


**Figure 2. Distribution for number of pages accessed by the general population (top) and the non-null intersection population (bottom).**

This distribution is similar to that of the intersecting population, however, there is one major difference, which is obvious by examining the section of the distribution between 0 and 1on the number of pages accessed axis. Notice that in the general population, there is a general decreasing trend in the number of individuals accessing a larger number of pages. Yet, in the intersecting population, the number of individuals actually increases up toward 1, or 10 pages (since we use a log scale of base 10). Thus, there appears to be a trend suggesting that individuals in the intersecting population tend to access more pages than the general population, closer to 10 versus less than 3 pages.

The final analysis performed was the proportion hypothesis test to determine if particular diagnoses were temporal. The temporal aspect considered was different time periods of the disease-predicting and disease-following user behaviors. There were 37 codes found to be indicative of temporal user behavior. 25 codes were indicative of the DP user type and 12 codes were indicative of the DF user type. A sample of the codes is provided in Table 3. Some codes, such as "normal pregnancy" had significance as close to the claim as two days prior to.

| Code | Description | User Type | Most Significant (period) | Earliest Significance (period) |
|---|---|---|---|---|
| 185 | malignant neoplasm of prostate | DP | 60 | 30 |
| 272 | disorders of lipoid metabolism | DF | 30 | 7 |
| 205 | myeloid leukemia | DP | 30 | 2 |
| 722 | intervertebral disc disorders | DP | 60 | 30 |
| 836 | dislocation of knee | DP | 60 | 30 |
| 608 | other disorders of male genital organs | DP | 30 | 30 |
| V22 | Normal pregnancy | DP | 14 | 2 |
| 255 | disorders of adrenal glands | DF | 60 | 60 |
| 441 | aortic aneurysm | DP | 60 | 30 |
| 606 | male infertility | DP | 30 | 30 |

**Table 3. Codes indicative of temporal behavior. "Most significant" is the time period with the largest significance score. "Earliest" is the period of earliest significance.**

## DISCUSSION

The results confirm that there exists a significant population of online users who access information corresponding to their own health status. There is a difference in the information that this population accesses compared to the general population. It is apparent that the accessed online diagnoses for the intersecting population have some disproportionate accessing compared to that of the general population. It appears that the general population has a propensity for searching for cancer and heart disease related diagnoses, while the intersecting population has a propensity for chronic disorders. Furthermore, it appears that the intersecting population tends to access more pages than the non-intersecting population. One possible explanation for this phenomenon is that individuals in the non-intersecting population tend to browse introductory pages and neglect pages with particular medical information.

It may be the case that diagnoses not equal to the physician-received diagnosis can be indicative of a specific accessing strategy. If so, then other concepts may be better indicators of certain diagnoses. Second, in this study we consider only the distinct set of diagnosis codes accessed, however, the number of times that an individual accesses an ICD-9 code may be indicative of a search strategy as well. Third, as the PEW study confirmed, a significant population searches for online health information regarding another individual. This may be an additional confounding feature.

There is also indication that specific diagnosis accessing occurs for diagnosis-predicting and diagnosis-following user behavior types. This temporal relationship is dependent on the time from the point of physician visit. It is encouraging to note that there exists a subset of diagnosis codes indicative of temporal user behaviors. For instance, one of the more interesting findings was that the generalized code V22, normal pregnancy, is an indicator of users searching for information before they make a visit to a doctor visit. This may not have been known, or expected, by the site content administrators, but by learning this information they may adapt the associated online information with respect to expected web users accordingly. Information on doctors or specialists in the area might be more useful for this user group than one that is following their doctor's diagnoses.

Acknowledgements

The author thanks Personal Path Systems for the use of their data. Additional thanks are extended to Latanya Sweeney for helpful suggestions. This work was supported by the Laboratory for International Data Privacy at Carnegie Mellon University.

References

1. Cline RJ, Haynes KM. Consumer health information seeking on the internet: the state of the art. Health Educ Res. 2001 Dec; 16(6): 671-72.
2. Houston TK, et. al. Consumer health informatics: a consensus description and commentary from the American Medical Informatics Association Members. Proc AMIA. 2001; 269-273.
3. Eysenbach G, Jadad AR. Evidence-based patient choice and consumer informatics in the internet age. J Med Internet Res. 2001 Apr-Jun; 3(2): e19.
4. Gustafson DH, et. al. Impact of a patient-centered, computer-based health information/support system. Am J Prev Med. 1999; 16(1):1-9.
5. Fox S, et. al. The online health care revolution. Pew Internet & American Life Project. http://www.pewinternet.org.
6. Gagliardi A, Jadad AR. Examination of instruments used to rate quality health information on the internet. BMJ. 2002; 324: 569-573.
7. Pandolfini C and Bonati M. Follow up of quality of public oriented health information on the world wide web. BMJ. 2002; 324: 582-583.
8. Hernandez-Borges AA, et. al. Can examination of WWW usage statistics and other indirect quality indicators help to distinguish the relative quality of medical websites? J Med Internet Res. 1999; 1(1).
9. Federal Register, 45 CFR Sect. 160-4. HHS. Standards for privacy of individually identifiable health information, Final Rule. Feb 26, 2001.