Analysis & Visualization of EHR Patient Portal Clickstream Data

Full Paper

Farhan Mushtaq
Worcester Polytechnic Institute
farhanmushtaq@gmail.com

Bengisu Tulu
Worcester Polytechnic Institute
bengisu@wpi.edu

Diane Strong
Worcester Polytechnic Institute
dstrong@wpi.edu

Sharon Johnson
Worcester Polytechnic Institute
sharon@wpi.edu

John Trudel
Reliant Medical Group
John.Trudel@reliantmedicalgroup.org

Lawrence Garber
Reliant Medical Group
Lawrence.Garber@reliantmedicalgroup.org

Abstract

The purpose of this paper is the analysis of EHR clickstream data for patient portal to determine patient usage behavior by analyzing patterns found in the data. Clickstream data retains vital information trail of patient's usage. Utilizing directed and undirected data mining techniques for data exploration and then visualizing specific patterns provide valuable information about how a patient utilizes these services. The information can help service providers to understand the demographics and behavioral aspects of the patients to develop, enhance and improve their systems to make the best use for these portals.

Keywords

Electronic health records, patient portals, use statistics, e-health

Introduction

Electronic health record (EHR) systems are becoming increasingly popular after the HITECH Act of 2009 that provides incentive payments to eligible hospitals and physicians through the Medicare and Medicaid reimbursement that demonstrate meaningful use of certified EHR technology. To avoid penalties and achieve Meaningful Use, providers must follow a set of criteria that provide guidelines for effectively using an EHR. (Zhang 2013).

EHR systems allow providers to share health records with their patients and facilitate communication with them (Tang et al. 2005). They are a tool for improving patients' health understanding, which can transform them into better educated consumers (Kahn et al. 2009). Number of studies have demonstrated that EHRs have changed and improved communication between patients and providers (Tang et al. 2005). Less attention has been paid to how patients utilize these portals (EHR website/applications), which can provide insight into how to most effectively design portals as well as to incorporate them into clinic workflows. With the potential for improving health and healthcare delivery via EHRs, healthcare provider organizations want to understand the characteristics of patients who use these new EHR tools, as well as their usage patterns.

Internet users leave behind a trail of usage information, called clickstream data, when visiting a webpage. Clickstream data is inherently heterogeneous and naturally uncertain, which makes visualization and pattern recognition difficult (Cadez et al. 2000). Clickstream data analysis studies have mostly focused on retailer sites for consumer behavior centric to financial goals (Kim et al. 2005; Yu et al. 2004; Zeng et al. 2002). In these environments, various data mining techniques have been implemented to examine navigation patterns on websites and other web-based systems using very large data sets (Moe 2003; Moe et al. 2004; Montgomery 2004). For analyzing clickstream data from patient portals, our goals are different.

We are less concerned about financial aspects and more concerned about patient and usage characteristics that can help healthcare systems design, develop, manage and improve these portals for delivering better healthcare services.

This paper reports our findings from a study to detect and understand EHR usage patterns. We utilize both directed and undirected data mining approaches (Berry et al. 2000) to study this data. Directed data mining is a top-down approach used when it's known what needs to be find. The goal is to create a predictive model from the existing data whereas undirected approach is a bottom-up approach where the data itself determines the relationships, if patterns are found, user decides whether the patterns are useful or not (Berry et al. 2000).

For the ***directed approach***, we ask some basic, yet key questions, to help EHR providers understand the characteristics of users and their patterns of use, including:

Q1. What are the demographic characteristics of the patients who are the users of the system?

Q2. What is the frequency of use? Is there a change in frequency of use over time?

Q3. At what time of the day are users more likely to access these resources?

Q4. How long is a typical user session?

Q5. Is there a difference in patients' use of the portals based on their age and gender?

Q6. Do users of different ages/genders spend different amounts of time on each page?

For the ***undirected approach***, we explore clusters of users and how they differ using R.

Our results provide a more detailed analysis of patients' use of portals than is in the literature (Bengisu et al. 2012) where the focus was to analyze usage based on age and gender with the help of system logs and patient interviews, whereas approach of this paper is the multi-dimensional exploration of clickstream data of a patient portal for patterns recognition and then to demonstrate the value of using this data to better understand patient behaviors. One of the limitation in our analysis is that it is solely based on patient's portal use, we did not have access to the patient's clinical data and hence we have not been able to compare the behaviors with their health statuses.

Methodology

Our data source is EHR usage data from a multi-site, multi-specialty group medical practice located in the northeast United States, with approximately 250 physicians and 1,500 other employees. The practice serves 200,000 patients with over one million patient visits annually. The group practice has 25 clinic locations that utilize a well-respected EHR system, providing features to support a patient portal and ambulatory care clinics. Patient portal was made available to all clinic patients at the beginning of 2009.

Data collection

As part of a larger research study, 10,000 patient portal users among 40,000 total users were randomly selected and invited to participate in the study. Users were defined as those who had logged into the system at least once. Among those invited, 632 patient portal users signed a consent form and agreed to participate in the study (a 6.3% response rate). Clickstream data logs for these patients were extracted for the period from January 2009 until February 2013, i.e., four years of usage data.

Every user activity recorded as a click in the clickstream data is generated as a result of a user clicking on a menu item on the main interface, such as login, logout or lab results. No other clicks are recorded in the data except these menu clicks. That is, subsequent clicks within a menu item, e.g., within lab results, are not reflected in our data. Although this might be seen as a limitation, from the perspective of analyzing how users navigate the system, this clean set of data is more useful in understanding the core functionality utilized by patients.

Data Analysis

We analyzed the data with both directed and undirected data mining approaches (Berry et al. 2000). The directed approach allowed us to classify the data and provide visualizations to answer the six questions listed previously. To prepare the data for directed analysis, we first loaded the data in an SQL database, and then wrote queries for data extraction targeted to answer each of our questions above. The extracted data was then loaded in Excel for visual analysis and plotting. This approach allowed us to take advantage of the visualization capabilities of Excel, which is often used as a front end data presentation tool.

The undirected data mining approach allowed us to look further for data patterns and to compare them with existing results above. We utilized R, a free programming language and software environment for statistical computing and graphics that provides a variety of statistical analysis and graphical techniques and can be easily extended using a large library of packages (R Team 2012). In R, we performed clustering analysis used for "unsupervised" learning (Tibshirani et al. 2000) to create data groups that have similar patterns. We used K-means clustering algorithm since it is known as a "simple and elegant approach for partitioning a data set into K distinct non-overlapping clusters" (James, G. 2013)

Results

Our results are organized as follows. First, we present our results from our directed data mining analysis to answer the six questions. Second, we present our undirected data mining analysis of how patients cluster into common usage patterns using R.

Directed Data Mining Analysis Results

Age and Gender Patterns of Users

To answer our first question, we explored the gender and age characteristics of our participants, as shown in Figure 1. Of our 632 study participants, 61% were females and 39% were males. We divided the participants into seven age brackets. The first bracket was defined as patients 18-30 years old; a bracket was then defined for each subsequent 10 years. The final bracket was patients above 80 years old. As shown in Figure 1, over 80% of the study patients fell in the age bracket between 41-80 years old age.

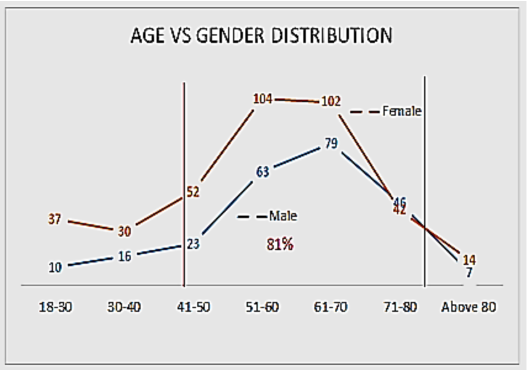

Figure 1. Age and gender distribution of patients

Active User pattern

When usage is voluntary, as it is with patient portal, organizations want to understand the number and type of users they are able to attract. Studies that investigated patient portal use typically reported increases in the number of patients who signed up over time to become users of the system (Bengisu et al. 2012). We considered the number of registered users (for this organization defined as those who had received an

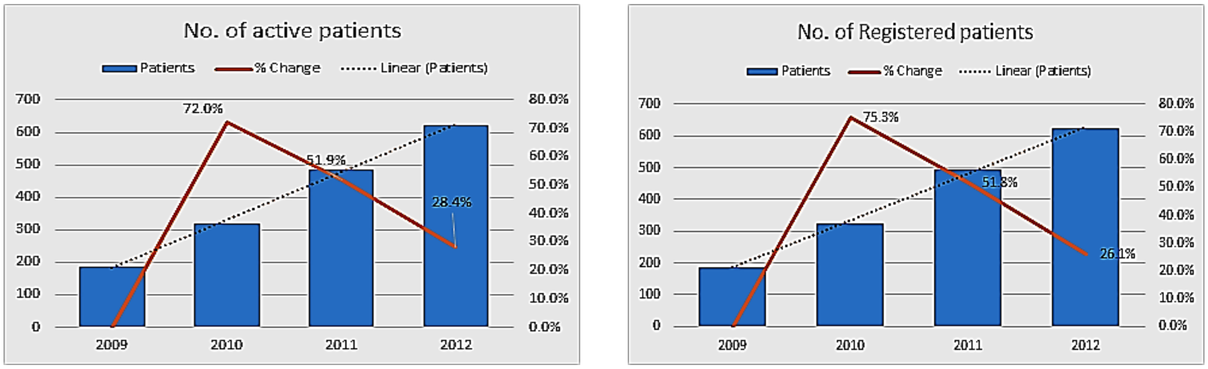

Figure 2. Yearly growth of participants who were active patients versus registered patients of the portal

activation code and logged into the system once), but we focused on the number of patients who were active within the system.

Active patients in a given year were defined as those who performed at least one session of activity in that year. For the 632 patients in our sample, the left half of Figure 2 presents the growth in the number

of active patients over the four years of analysis, in which 72% growth was observed in the first year (2009 to 2010), which then slowed down 51.9% and 28.4% in the years 2011 and 2012 respectively. On average, yearly growth in active patients was 67%. The right half of Figure 2 reflects the number of patients who were registered patients of the system, that is, they had signed up to use the patient portal by logging in at least once before or during the given year. The two graphs demonstrate that for our sample, most registered patients continued to be active portal users. In 2009 and 2012, 100% of the registered patients were active.

Amount of Activity per User Patterns

Due to yearly growth in the number of users (see Figure 2), total patient menu-click activity, as total clicks generated in a year or month, is likely to increase over time. We were interested in understanding if the growth in activity was also reflected at the patient level, i.e., was there more activity per user or only more activity because there were more users. To visualize how activity per user varied over time, the weighted average menu-click

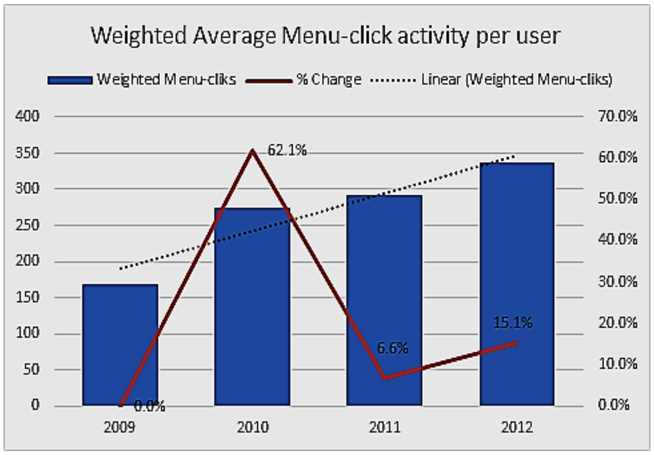

Figure 3. Weighted Average activity per user per year

Activity per user was calculated by:

Weighted average menu-click activity=Total menu-click activity per year/Total active patients per year

The weighted average menu-click activity is plotted in Figure 3. A significant increase was observed in 2010, when the menu-click activity increased by 62% over the initial year; it became steadier in the next two years. The average annual rate was 28%.

This increase indicates that individual patients were becoming more active. Although some of this increase in activity may be due to new features integrated into the system over the first three years of the implementation, the growth in the activity of individuals continued even after the system features stabilized. The non-weighted activity trend is plotted Figure 4, which presents a monthly analysis of patient menu-clicks, revealing that activity steadily increased at a rate of average 7% per month.

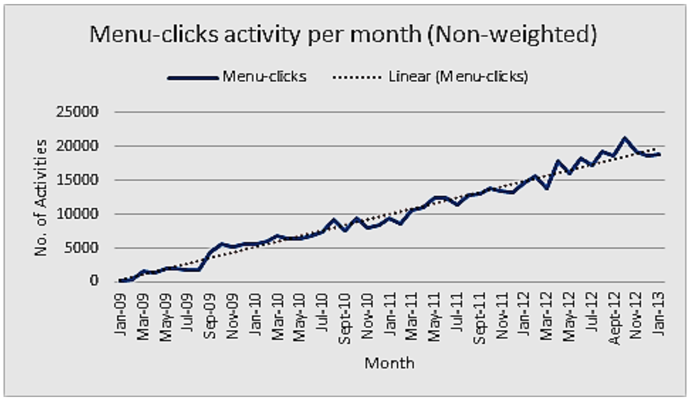

Figure 4. No. of Menu-clicks per month

Time-of-Day Usage

To answer the third question posed, we calculated the time of the day usage for different age groups.

Figure 5 shows the 24 hour patient activity usage pattern for the portal. The time of the day usage showed that 52% of the patient activity occurs between 8AM and 2PM and the trend clearly shows higher usage by the age group above 60 years old. The age group 18-40 is more active than the age group 40-60 between 9AM to 5PM. It is also interesting to note that the hourly trend is relatively consistent for all age groups.

The information can help providers and clinics as they plan for and educate patients about response time.

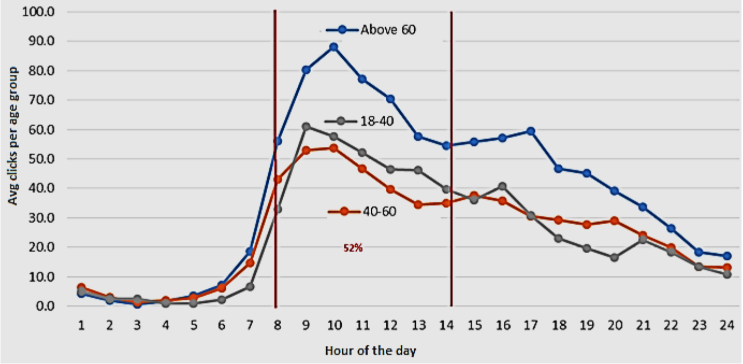

Figure 5. Time of the day usage

Clicks per Session Patterns

In this analysis, we turn our attention to patterns within a session of activity. A session of activity consists of a patient login, followed by various click activities, and terminates with either a logout or timeout after an idle time of 21 minutes (i.e., the session is terminated by the server). As shown in Figure 6 most patient sessions consisted of only a few clicks.

Specifically, 48% of the total sessions had fewer than six clicks and 80% of the sessions consisted of fourteen menu-clicks or fewer. One interpretation may be that patients are focused during their sessions, i.e., they logged in to get specific information or accomplish specific goals and then terminated their session

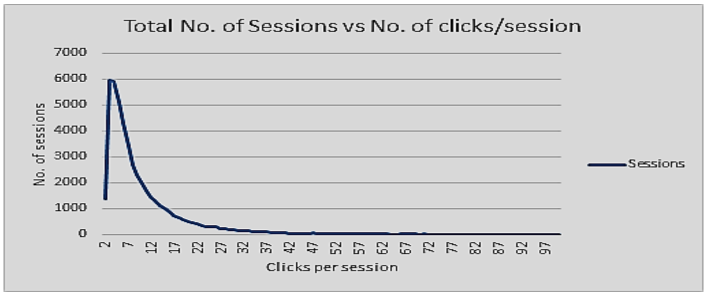

Figure 6. Clicks per session pattern

Menu-clicks vs Age and Gender Patterns

The fifth question we explored was whether patients' use of the portal differed depending on their age and gender. We observed a consistent difference in the behavior of the patients relative to age, with older patients demonstrating more activity than younger ones, as shown in Figure 7. This trend was steadier and more pronounced in the male patient group than for female patients. Specifically, the change in average

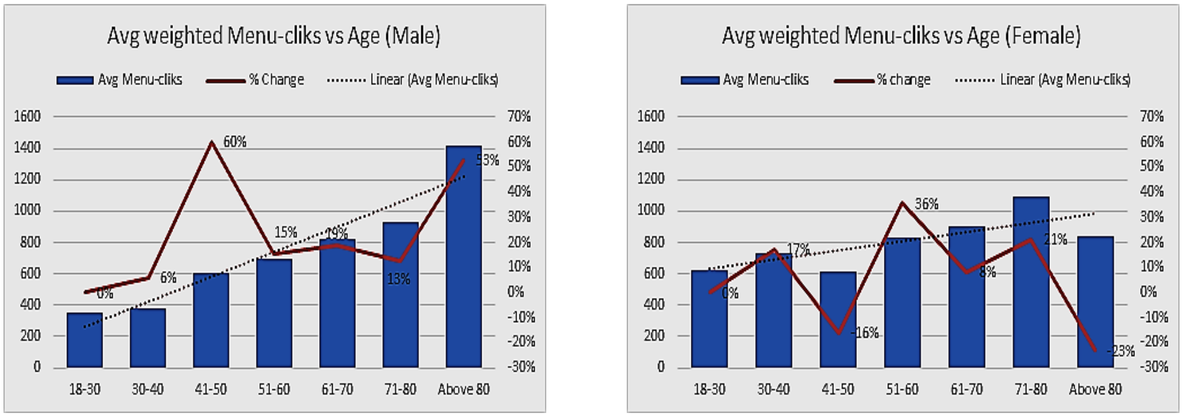

Figure 7. No. of menu-clicks by age and gender

menu-click activity was 28% per 10 years of age for male patients, whereas in female patients it was only 6%. It shows that usage is relatively low for young males that increases rapidly with age. It could be either because the male population below 40 might not be too concerned about their health and hence do not care much about using the patient portals or it might be that they are maintaining a good health that doesn't really require frequent clinic and subsequent portal visits. Unfortunately, we do not have access to the clinical data to prove these findings. In future, having this data would certainly help in understanding the patient's portal behaviors with respect to their clinical visits. For healthcare providers and researchers, this data might suggest that an opportunity exists to further study these behaviors and may be encourage the younger male population to become more active users of the portal.

The weighted menu-click was calculated as:

Weighted menu-clicks=Total menu-clicks in the group/ total active patients in the group.

Patterns of Average Time between Clicks

We also found that the time between the clicks within a patient session differed as a function of the age of the user. Specifically, the time between clicks generally increased with increasing age, as shown in Figure 8. (Recall that a click refers to a menu item, not clicks within a menu item.)

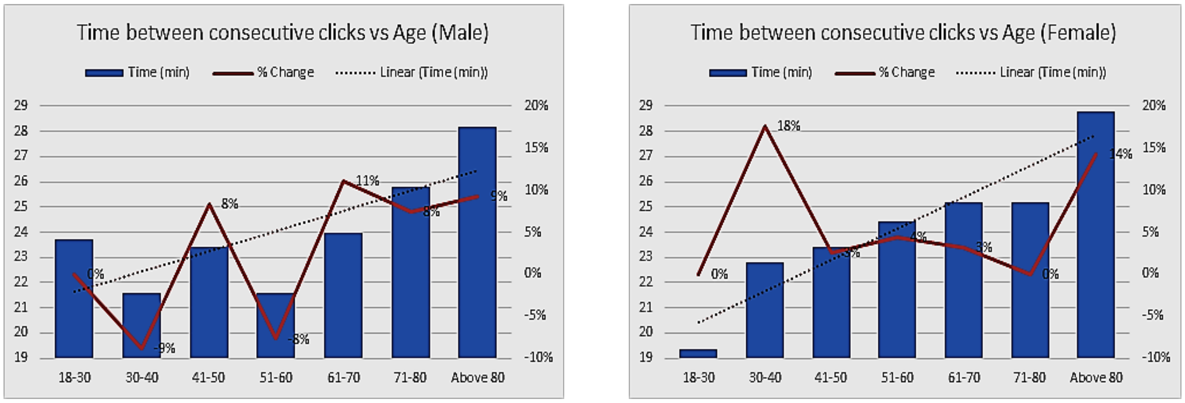

Figure 8. Time between clicks vs age and gender

Figure 8 shows that older patients tended to spend more time on a page than younger patients. For the female patient group in our sample, the increase in time occurred across age groups, whereas in male patients it was variable below 50 years but after that it increased at an average rate of 9%.

This information could help service providers to predict the time period their targeted patient segments will spend at the portal and the expected response time if required at a given page.

Undirected Data Mining Analysis Results

Our purpose with the undirected data mining analysis is to explore whether common clusters of patients exist, distinguished by their age and gender characteristics or their usage patterns, and how these clusters differ.

Once the data had been pre-processed, we used K-means clustering algorithm in R, which is the most popular partitioning method and is simple and flexible. The algorithm requires the desired number of clusters to be pre-defined, which makes it an ideal candidate for experimental and evaluation purposes that involve comparing multiple cluster results.

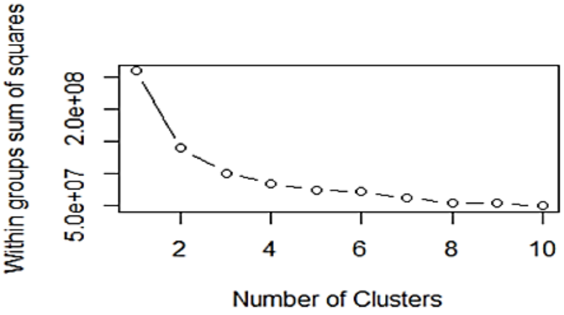

Figure 10. Number of Clusters

To select an appropriate number of clusters, we plotted the within groups sum of squares by number of clusters extracted (see Figure 10). The sharp decreases from 1 to 4 clusters with very little decrease from 5 to 10 suggest a number close to 4 would be the best solution. We then performed our analysis using 3, 4 and 5 clusters (see Table 1).

Some of the clusters are very small, for instance C_2 in 4-Cluster and C_5 in 5-Cluster have only three patients. Interesting, these are the same 3 patients that in both of these clusters. They can be considered as outliers because they exhibit behavior that is very different than the patients from other clusters. Specifically, these three patients demonstrate very little activity and have never used the Appointment schedule or cancel feature.

	Patients per cluster		
Cluster	3-Cluster	4-Cluster	5-Cluster
C_1	105	20	71
C_2	17	3	18
C_3	502	117	168
C_4		484	364
C_5			3

Table 1. Patients per Clusters

Because they also have the highest number of account re-enable requests, it is likely that they find it difficult to use the portal. So clustering provides very useful information to the service providers by associating each patient with a category of patients that demonstrate similar usage behavior. It is easier for the service provider to determine the specific needs and provide tailored solution for an identified group of patients.

	Average Age Distribution (Years)			**Gender Distribution (Female %)**		
Cluster	3-Cluster	4-Cluster	5-Cluster	3-Cluster	4-Cluster	5-Cluster
C_1	54.84	56.30	54.28	57%	45%	66%
C_2	53.59	49.00	56.00	47%	33%	44%
C_3	58.03	53.69	56.17	62%	59%	57%
C_4		58.36	58.66		62%	63%
C_5			49.00			33%

Table 2. Age and Gender Distribution within Clusters

As Table 2 shows, age and gender do not appear to be major discriminatory factors in the clustering distribution. The distribution by cluster is primarily due to the difference in patient's usage, as shown in Figure 11.

An interesting finding from Figure 11 is that the distribution of usage among the top 10 features of the EHR system was consistent across all cluster groups. That is, the main distinguishing pattern is the frequency of use of the five most frequent features, messaging, lab tests, lab results, appointment review, and appointment details. Note that the ordering of the features is the same for each category of users.

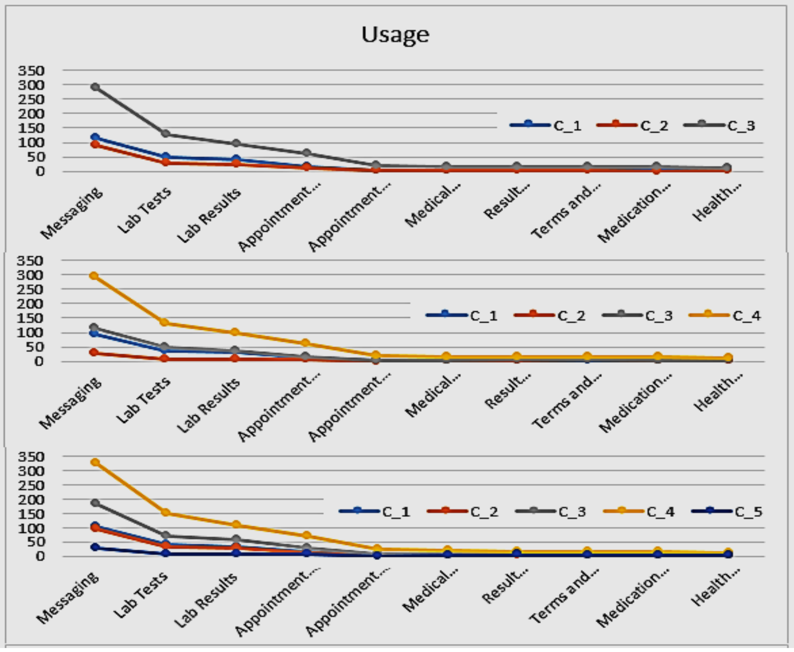

Figure 11. Cluster Usage distribution

Time of Day Usage

Time of the day usage was computed for a 4-cluter scheme. In Table 3, columns C1-C4 represents the average number of clicks recorded in the given time range for each cluster. 'Average for Time Block' represents the average number of clicks that occur in each time range, across all clusters.

The distribution of activities across each time range is almost identical across the clusters except C_2, which only consists of few patients who can be considered as outliers, so this group itself is not a significant cluster distributer because every cluster has the same proportion of time ranges.

	C_1	C_2	C_3	C_4	Average for Time Block
6AM-11:59AM	104.5	37.33	139.2	396.74	169.4
12PM-5:59PM	82.2	34.33	116.33	338.7	142.9
6PM-11:59PM	51.35	19	66.4	185.86	80.7
12AM-05:59AM	3.15	0	5.54	18.79	6.9

Table 3. Time of the Day Usage

These results are in a line with the Time of the day usage computed earlier in Figure 5 with a difference that Figure 5 plots the hourly data for three different age groups, whereas Table 3 provides the aggregate for each cluster group. But the overall trend is consistent where the most frequent time range is from 6AM to 11:59 AM, which decreasing as the day gets later.

Conclusion

In summary, in our study with 632 participants (61% female, 39% male) who were active users of a patient portal, over 80% fell in the 41-80 years old age bracket. Over the time period covered (2009-2013), there was a 67% yearly growth in the number of active portal patients. Activity per user (as measured by average weighted growth in menu-click portal activity) increased 28% annually. Patients' sessions were relatively short, with nearly half of the sessions consisting of fewer than six clicks. Most patient activity occurred between 8AM and 2PM, with a similar pattern for each age group. In our sample, the data also suggested that older patients tended to be more active than younger patients. The time required for patients to review information on the page also increased with age, with an average increase of 6% per 10 years in females and 3% in males.

Such information about patient portal activity is useful to physicians and practice leadership. For example, one goal of portal implementation is patient engagement. In our sample, which suggested that active users are engaging more with the system over time, the portal might be viewed as supporting this goal. When significant new functionality is added to the system, leaders could examine activity to see whether patient engagement (measured by activity) has increased. While the number of users (and any increase) is also important, data about the amount of activity per user provides one measure of the strength of the relationship between user and system. In addition, it supports resource planning for server capacity needed to host the patient portal. A consistent growth in this usage should be taken into account at the time of deployment for the planned lifespan of the associated resources. Exploring usage patterns to examine whether differences occur between groups can also yield valuable information. In our future work, we will to continue to explore different metrics and relationships to describe portal use.

Analysis of clickstream data of patient portals using simple data extraction and clustering techniques in R can reveal important findings that describe patients' portal usage behavior. Visualization suggests interesting patterns that can then be tested more formally and completely with statistical approaches to examine their significance. For clustering users and finding commonalities a tool like R can be used for data processing to discriminate patients by their portal usage patterns.

Ultimately, the information obtained from analyzing patient clickstream data can help service providers to build, manage, maintain and improve systems to make the most effective use of their portals.

REFERENCES

Berry, M., & Linoff, G. 2000. *Data Mining Methodology: The Virtuous Cycle Revisited*. Mastering data mining: The art and science of customer relationship management pp. 40-41. New York: Wiley Computer Pub.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S., 2000. *Visualization of navigation patterns on a web site using model-based clustering*. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 280-284, New York, NY, USA, ACM.

James, G. (2013). Clustering Methods. In *An introduction to statistical learning with applications in R* (1st ed.). Springer

Kahn, J. S., Aulakh, V., Bosworth, A. 2009. *What It Takes: Characteristics Of The Ideal Personal Health Record*, Health Affairs, pp. 369-376.

Kim, Y. S., Yum, B. J., Song, J. 2005. Development of a recommender system based on navigational and behavioral patterns of customers in ecommerce sites, Expert Systems with Applications, Vol. 28, No. 2, pp. 381-393.

Moe, W. W., 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream, Journal of Consumer Psychology, Vol. 13, No. 1, pp. 29-39.

Moe, W. W., Fader, P. S., 2004. *Capturing evolving visit behavior in clickstream data*, Journal of Interactive Marketing, Vol. 18, No. 1, pp. 5-19.

Montgomery, A. L., Li, S., Srinivasan, K. 2004. *Modeling online browsing and path analysis using clickstream data*, Marketing Science, Vol. 23, No. 4, 579-595.

R Development Core Team 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

Tang, P. C., and D. Lansky. 2005. *The Missing Link: Bridging the Patient-Provider Health Information Gap*. Health Affairs (Millwood) 24, no. 5 (2005): 1290–95.

Tibshirani, R., Walther, G., & Hastie, T. 2000. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 411-423.

Tulu, B., Strong, D., Johnson, S., Bar-On, I., Trudel, J., & Garber, L. 2012. *Personal Health Records: Identifying Utilization Patterns from System Use Logs and Patient Interviews*, HICSS, 45th Hawaii International Conference on System Sciences (HICSS) 2012, pp. 2716-2725, doi:10.1109/HICSS.2012.47

Yu, L., Liu, L., Li, X. F. 2004. *Research on personalized recommendation algorithm for user's multiple interests*, Computer Integrated Manufacturing Systems, Vol. 10, No. 12, pp. 1610-1615.

Zeng, C., Xing, C. X., Zhou, L. Z. 2002. *A survey of personalization technology*, Journal of Software, Vol. 13, No. 10, pp. 1952-1961.

Zeng, L., Xu, L., Shi, Z., Wang, M., Wu, W. 2007. *Techniques, process, and enterprise solutions of business intelligence*, 2006 IEEE Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan, Vol. 6, pp. 4722.

Zhang, J. H., 2013. *Impacts of US health care reform on IT firms' revenue: The case of EHR Meaningful Use criteria*. Engineering, Management Science and Innovation (ICEMSI), 2013 International Conference on, Issue Date: 28-30 June 2013