Published in final edited form as:
J Biomed Inform. 2011 April ; 44(2): 333–342. doi:10.1016/j.jbi.2011.01.007.

Learning Relational Policies from Electronic Health Record Access Logs

Bradley Malin[a], **Steve Nyemba**[a], and **John Paulett**[a]

[a]Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, Tennessee USA

Abstract

Modern healthcare organizations (HCOs) are composed of complex dynamic teams to ensure clinical operations are executed in a quick and competent manner. At the same time, the fluid nature of such environments hinders administrators' efforts to define access control policies that appropriately balance patient privacy and healthcare functions. Manual efforts to define these policies are labor-intensive and error-prone, often resulting in systems that endow certain care providers with overly broad access to patients' medical records while restricting other providers from legitimate and timely use. In this work, we propose an alternative method to generate these policies by automatically mining usage patterns from electronic health record (EHR) systems. EHR systems are increasingly being integrated into clinical environments and our approach is designed to be generalizable across HCOs, thus assisting in the design and evaluation of local access control policies. Our technique, which is grounded in data mining and social network analysis theory, extracts a statistical model of the organization from the access logs of its EHRs. In doing so, our approach enables the review of predefined policies, as well as the discovery of unknown behaviors. We evaluate our approach with five months of access logs from the Vanderbilt University Medical Center and confirm the existence of stable social structures and intuitive business operations. Additionally, we demonstrate that there is significant turnover in the interactions between users in the HCO and that policies learned at the department level afford greater stability over time.

Keywords

electronic health records; organizational behavior; knowledge discovery; access logs; auditing

1. Introduction

The healthcare community has made considerable strides in the development and adoption of information technologies [1]. Evidence indicates that the collection, storage, and processing of patient data in electronic form can decrease costs, strengthen staff productivity, and promote safety [2,3]. To realize these benefits on a broad scale, healthcare

To whom correspondence should be addressed: Bradley Malin, Ph.D. 2525 West End Avenue, Suite 600 Department of Biomedical Informatics Vanderbilt University Nashville, TN 37203 USA b.malin@vanderbilt.edu tel: +1 615 343 9096 fax: +1 615 322 0502.

organizations (HCOs) must deploy information technology in a manner that facilitates business processes without jeopardizing patients' privacy rights [4]. HCOs are beginning to utilize software tools to design clinical information systems with logically sound privacy and security controls (e.g., [5,6,7,8]), however, such tools are limited in practice. In part, this is because they depend on HCO administrators to supply clear policy definitions, which has technical as well as social complications.

From a technical perspective, the specification of policies is a nontrivial challenge because a healthcare system is an inherently dynamic environment where teams of clinicians and support staff interact [9,10]. The notion of dynamic teams, while effective in supporting healthcare operations, is more complex than the pairwise relationship typically expected by access control frameworks (e.g., provider-patient) [11]. The complexity of the problem is compounded by the reality that HCOs evolve to address technological and organizational pressures, such as the adoption of new patient management protocols, reaction to legislated changes, and assimilation of rotating medical students [12].

From a social perspective, traditional methods for defining security policies are problematic because they rely on the knowledge of domain experts (e.g., employees) or the observations of external specialists. These methods may be feasible within the context of small systems, but are unmanageable for modern HCOs where the number of policies can be large, defined in an *ad hoc* manner, and revised at a moment's notice. The traditional approaches are further limited because they are subjective and can suffer from informant accuracy [13], which will be magnified by the scale of modern HCOs.

As an alternative, the growing adoption of electronic health record (EHR) systems provides an opportunity to apply automated learning methods that are data-driven, so that policies can be docked to actual behavior. Modern EHR systems already generate access logs to enable the construction of audit trails regarding what information care providers observe and what actions they take. Furthermore, many countries have enacted regulations that mandate their inclusion, such as the Security Rule of the U.S. Health Insurance Portability and Accountability Act (HIPAA), which requires HCOs to retain access logs for a minimum of six years [14]. The automated extraction and representation of policies from such systems could complement traditional policy specification frameworks because they could assist in auditing existing policies, as well as discovering novel patterns of EHR system use. If such approaches can be designed in a scalable manner, they can endow administrators with a unique view of the stability, or volatility, of policies over time within their HCO.

We further anticipate that EHR user behavior can serve as the basis for automated surveillance tools that uncover policy violations. The problem could be framed as an anomaly discovery problem, such that accesses (or behaviors) that are sufficiently different than expected EHR system usage are flagged for administrative review. In this respect, we envision surveillance systems that mine policies from EHR access logs in a similar way to how intrusion detection systems mine rules from the audit logs of file systems [15,16]. However, a core assumption of such surveillance approaches is that behaviors manifest in regular patterns. As such, one of the primary goals of this paper is to assess if EHR system usage can be distilled into a representative set of patterns. While there are many different types of patterns that could be extracted, we limit our focus to patterns that summarize how EHR users and HCO departments collaborate. This is a fundamentally different problem than that studied in intrusion detection because EHR users are expected to function in a coordinated manner.

Specifically, we introduce a process to transform a database of EHR users' access of patient records into patterns of users' interactions, in the form of probabilistic rules. The process

consists of two primary components; first we infer a social network of EHR users based on co-accesses to patient records. We focus on networks, so that we can capture the collaborative model of HCOs and their employees and, as we show, this network can be decomposed into clusters that represent expected organization structure at a high-level (e.g., children's vs. main hospital). Second, we convert the pairwise interactions of users into probabilistic rules of association. These rules capture the frequency of interaction between two users, as well as the likelihood of a user accessing a patient record given that another user accessed the same record. We focus on association rules because (as we review in Section 2) they have been shown to be effective at decomposing a complex environment into simple statistical components that can be applied in intrusion detection frameworks, but can also be presented to humans for review if need be. To make our approach reusable, we have developed open-source software to support the replication and application of our investigations.[1]

While there has been a number of investigations into the extraction of business process rules from access logs, as well as computational approaches for organizational modeling, to the best of our knowledge; this is the first research to link the concepts. Beyond a theoretical treatment of the problem, we perform an empirical analysis with five months of access logs from a large HCO's EHR system that is well-integrated into core healthcare operations. Through this investigation, we confirm that the HCO is a highly dynamic environment, which is not necessarily appropriate for traditional access policies. At the same time, we demonstrate that automatically learned policies can be grouped into levels of stability with intuitive implications.

2. Related Work

Historically, automated learning approaches have proven to be capable of extracting relevant knowledge from access logs. Much of the work in log mining parallels the growth of the Internet, where it was shown that a domain's webpage access logs could be modeled as a database of transactions [17], from which association rules could be discovered to characterize users' tendencies [18,19]. Beyond log analysis methodology, it was demonstrated that knowledge, mined from access logs, can inform business decisions and can provide feedback to make systems more efficient. As a result, e-commerce sites routinely employ access log analytics to influence, attract, and retain customers [20,21,22,23,24,25]. Beyond marketing, there is some evidence in the healthcare domain which indicates that EHR access logs can reveal the clinical information use behavior for patients [26] and providers (e.g., in one study it was shown that care providers often viewed a patient's laboratory and radiology results in the same session) [27,28,29]. Further work has shown that logs are effective for understanding how medical information systems are used for educational purposes [30,31,32].

Nonetheless, research on access log analytics has, in general, focused on “what” users view, but HCO policies need to model “who” and is viewing the health records of whom and “why”. To address the latter, it helps to know the reason for a particular access; however, such information is not always adequately documented in an EHR. As such, in this research we focus on the extent that basic information in EHR access logs can provide intuition into the patterns of use that care providers generate when accessing patient records. In doing so, our goal is to uncover the relationships between individuals in the healthcare enterprise and the types of health record access workflows that exist. In this respect, methods from the social network analysis [33] and computational organizational science [34] domains are ideal candidates for policy extraction. In the biomedical realm, such techniques have been

[1]A freely available open source software tool, which can be downloaded from http: //code.google.com/p/hornet/source.

successfully applied to characterize the interactive and dynamic nature of multi-disciplinary research communities [35,36], to discover the dynamics of public health settings to facilitate change management [37], to explain physician adoption rates of EHR systems [38], and, most recently, investigate the relationships of healthcare providers in various clinical contexts [39,40,41].

In the context of health information security, we recognize that auditing through EHR access logs has become a practice both for HCOs and, more recently, for patients. One of the primary features that healthcare consumers desire in an EHR is transparency and, in particular, consumers want to be provided with the ability to access their medical records, as well as determine who has viewed them [42,43,44]. In support of this function, a system developed at New York Presbyterian Hospital was built [45] to enable the latter and found that such a tool can lessen the burden for EHR administrators by increasing the number of individuals that can monitor for violations. At the same time; however, it was observed that the task of analyzing the logs is complex due to the number of access that occurs for a typical patient's record, as well as the fact that patients are neither well-versed in an HCO's operations nor comprehend if a user has a legitimate reason to access their record.

To assist in EHR access logs monitoring, Asaro and Ries [46] prototyped a tool to automatically analyze EHR access logs by inspecting the distribution of records accessed by users. In their investigation, it was observed that some users accessed an excessive number of medical records; however, no methods were designed to determine if such behavior was in line with the behavior of the HCO. Thus, while the scientific literature exists for setup, collection, and summarization of EHR access logs, the study of the information within the access logs for health information security and policy evaluation has been neglected.

3. Relational Healthcare Policy Discovery

This section describes the framework and methods applied for extracting healthcare policies from EHR access logs and begins with a high-level overview. The framework consists of two core components: 1) HCO network construction and 2) association rule discovery. First, access transactions are transformed into a graph that summarizes how EHR users interact through patient records in the system. This graph constitutes an inferred social network of the EHR users. Second, the pairwise relationships of the users in the graph are converted into association rules that communicate the relative frequency of user interaction via a patient record, as well as the conditional likelihood of a user accessing a patient's record given that another user accessed the record. In addition, the graph and resulting rules are abstracted to represent relations between HCO departments to analyze the extent to which existing organizational models can dampen noise in the resulting rule sets.

Before detailing the methods, we take a moment to formalize the system. We consider an environment composed of a set of users and a set of patients, which we refer to as U and P, respectively. Each user and patient is characterized by a unique persistent identifier (i.e., system login or medical record number). Each access by a user to a patent's record is represented as a logged transaction in a database; we refer to the set of transactions as $T = \{t_1, t_2, \ldots, t_m\}$. The transaction is represented as a 3-tuple of the form $\langle u, p, stamp \rangle$, where $u \in 2U$, $p \in P$ and *stamp* is the timestamp associated with the access.[2] The upper section of Figure 1 provides an example of access transactions, where three EHR users access a total of four patients' records over a three day period.

[2]The transaction is likely to contain additional information relevant to the EHR system, such as the client's IP address and the action taken by the user. For the purposes of this study, we are interested in modeling organizational structure and not what actions are performed to medical record. This is a fruitful direction for study, but is beyond the scope of this study.

3.1. Network Construction

Though EHR systems rarely document which users explicitly collaborate, the patient record can serve as a proxy. In fact, at times the patient record may be a better indicator of collaboration because, in an HCO, care providers in a common workflow may not have face-to-face interaction. Thus, we derive a relational network of EHR users based on common patient record accesses within a certain time period. Without loss of generality, we assume that the set of transactions has been restricted to the time period of interest and henceforth consider 2-tuples of the form $\langle u, p \rangle$. The resulting network is modeled as a graph. Each vertex in the graph corresponds to a user $u_i \in U$ and is assigned a weight, which corresponds to the number of patients accessed by the user:

$$w_{u_i} = |\{p | \langle u_i, p \}|$$

Each edge is weighted as the number of patients that users u_i and u_j accessed in common:

$$w_{u_i,u_j} = \left| w_{u_i} \cap w_{u_j} \right|$$

Figure 1 depicts a user network for the example access logs. Notice, *Nurse S* accessed the records of *Bob*, *Charles*, and *Daniel*, thus $w_S = 3$. Also, *Nurse S* and *Dr. X* accessed two patients in common, so the weight of their corresponding edge is $w_{S,W} = 2$.

3.2. Network Abstraction

Models learned at the level of specific EHR users can obscure how modern HCOs function from a team-based perspective. For example, certain groups of nurses and physicians can perform the same actions, but with different sets of patients, such as when the patients reside in different wards of a medical center. User-level models based on such events might obscure the fact that similar policies, or workflows, are being followed. We hypothesize that an appropriate abstraction could mitigate such factors to detect more robust signals for policy discovery. An abstraction could be the mapping of EHR users to the department (e.g., Critical Care, Internal Medicine, or Oncology) or the role (e.g., physician, nurse, or biller) to which they are assigned in the HCO. In the abstracted network we construct in this work, each vertex corresponds to an abstracted value and each edge represents the combined edges of the constituent user-level vertices associated with the corresponding value. If a user is associated with more than one abstracted value (e.g., an attending physician with appointments two distinct departments of the HCO), we discount the user's contribution to the abstracted edges and vertices proportionally to the number of values to which they are mapped.

More formally, let $A = \{a_1, a_2, \ldots, a_x\}$ be the set of abstract values. Each value is composed of a set of users, such that $a_i \subseteq U$. Let x_i be the set of abstracted values to which u_i is assigned. Then, the weight for the abstracted vertex is:

$$w_{a_k} = \sum_{u_i \in a_k} \frac{w_{u_i}}{x_i}$$

And, the weight for each edge in the abstracted network is computed as

$$w_{a_k,a_l} = \sum_{u_i \in a_k, u_j \in a_l, u_i \neq u_j} \frac{w_{u_i,u_j}}{x_i x_j}$$

Figure 1 depicts a scenario in which *Dr. J* maps to *(E)mergency Medicine* and *(H)ematology*, *Nurse S* maps to *(G)astroenterology* and *(H)ematology*, and *Dr. X* maps to *(E)mergency Medicine*. Using the previous equations, we find the node weights:

$$w_E = \frac{w_J}{x_J} + \frac{w_X}{x_X} = \frac{1}{2} + \frac{2}{1} = 2.5$$
$$w_G = \frac{w_S}{x_S} = \frac{3}{2} = 1.5$$
$$w_H = \frac{w_J}{x_J} + \frac{w_S}{x_S} = \frac{1}{2} + \frac{3}{2} = 2$$

And, we compute the edge weights accordingly:

$$w_{E,G} = \frac{w_{S,J}}{x_S x_J} + \frac{w_{S,X}}{x_S x_X} = \frac{1}{2*2} + \frac{2}{2*1} = 1.25$$
$$w_{G,H} = \frac{w_{S,J}}{x_S x_J} = \frac{1}{2*2} = 0.25$$
$$w_{E,H} = \frac{w_{S,J}}{x_S x_J} + \frac{w_{S,X}}{x_S x_X} + \frac{w_{X,J}}{x_X x_J} = \frac{1}{2*2} + \frac{2}{2*1} + \frac{0}{1*2} = 1.25$$
$$w_{H,H} = \frac{w_{S,J}}{x_S x_J} = \frac{1}{2*2} = 0.25$$

3.3. Interactive Policies

From the relational network, we can determine the likelihood that users access the same patient record by mining associative patterns. There are numerous ways in which patterns can be extracted and represented, but for the purposes of this work, the patterns we focus on are association rules [47]. This choice was made for several reasons based on computability and interpretability. First, association rules can be calculated efficiently, requiring no more than a single scan of a database table (for rules with two users in the rule), which is desirable given the large number of users, patients and access log transactions. Second, association rules allow for probabilistic representations of a system, with intuitive statistical interpretation (i.e., joint and conditional probabilities, as explained below). Third, association rules are relatively simple for humans to comprehend. This criterion is a necessity when policies are presented to human administrative experts for review (though this is beyond the scope of the study). We note that the association rule mining model invoked in this work is not designed to address temporal relations or more complex, potentially informal interactions (e.g., users that are not directly connected). However, this is an artifact of the manner in which we model the social network and believe that alternative data structures and techniques based on sequential pattern mining and clustering can facilitate the learning of such information.

Specifically, we model the joint probability that users u_i and u_j access the same patient's record and the conditional probability that u_j accesses a patient record given it was accessed by u_i. To compute these probabilities, we treat each patient as a transaction of EHR users $\{u_i, \ldots, u_j\} \subseteq U$.

Support—In the context of association rules, the joint probability is referred to as the *support* of the rule and is calculated as:

$$sup\left(u_i, u_j\right) = \frac{w_{u_i,u_j}}{|P|} \quad (1)$$

where $|P|$ is the total number of transactions (i.e., the number of patients). Intuitively, the support for the association rule $u_i \Rightarrow u_j$ corresponds to the number of patients that u_i and u_j

accessed in common, normalized by the total number of patients. As a result, the support score of any rule is bounded by [0,1]. Additionally, the support score for a rule is symmetric, such that $sup(u_i \Rightarrow u_j) = sup(u_j \Rightarrow u_i)$. A high support score indicates that it the relationship between u_i ad u_j is more common in the HCO, such that a larger number of patients are accessed by both users. The greater the support, the more likely it is that the users work in high volume clinical care areas. However, given the typical delivery of care by all but the smallest HCOs, it is unlikely that any set of users will have relatively large support scores. For example, for an edge to achieve a support of 0.10, both users must access at least 1 in every 10 patients at the HCO.

Continuing the example from Figure 1, we can calculate the support of the *(E)mergency Medicine* and *(H)ematology* departments as:

$$sup(E,H) = \frac{w_{E,H}}{|P|} = \frac{1.25}{7} = 0.179 \quad (2)$$

Confidence—While the support provides the relative frequency of a relationship it lacks an implication and can obscure notable policies. For example, an EHR user can have multiple subordinate users (e.g., medical residents reporting to the same attending physician), each of which accesses a subset of the patients of their superior. If the subset of patients is small, the support for each superior and subordinate pairing will be small as well. Yet, there is a clear policy, which is that the subordinate accesses the same records as the superior. In this sense, a conditional probability can enable the discovery of strong relationships that have low support. The probability that a patient's record is accessed by a user u_j, given that the record was accessed by user u_i is:

$$conf\left(u_i \Rightarrow u_j\right) = \frac{w_{u_i,u_j}}{w_{u_j}} \quad (3)$$

This is referred to as the *confidence* of a rule, where user u_i is the antecedent and user u_j is the consequent. Unlike the support, the confidence score is asymmetric, such that $conf(u_i \Rightarrow u_j)$ may differ from $conf(u_j \Rightarrow u_i)$.

We can calculate the confidences between *(E)mergency Medicine* and *(H)ematology* departments from Figure 1:

$$conf(H \Rightarrow E) = \frac{w_{E,H}}{w_H} = \frac{1.25}{2} = 0.625 \quad (4)$$

$$conf(E \Rightarrow H) = \frac{w_{E,H}}{w_E} = \frac{1.25}{2.5} = 0.5 \quad (5)$$

Though the support and confidence scores are relatively simple to compute, they are powerful tools for reducing the huge amount of data that access logs provide into meaningful observations regarding events within the HCO.

3.4. Policy Decay

To characterize how the relational network evolves over time, we compute the number of weeks a rule manifests during the study period. A rule that appears for a large portion of the study period suggests that the users or departments often accessed records in common. Additionally, the greater the number of rules that exist across weeks, the greater the evidence we have in favor of a stable organizational structure. If, on the other hand, rules are detected for only a few of the weeks in the study period, we have more evidence that there are organizational changes or more sporadic events.

At one extreme, we can imagine an HCO that has a static relational network due to highly consistent behavior of all entities over time. For example, a practice with two doctors and three nurses will likely be stable over time because, on average, the nurses will likely interact with the doctors consistently, and the doctors will have relatively independent sets of patients they care for. At the other extreme, we can envision an HCO that has a dynamic relational network, which could be caused by a large number of care providers who perform roughly the same duties. Since multiple care providers function in an interchangeable manner, the possible combinations of care paths would increase significantly. We are not suggesting that care providers are commodities, but rather that as the number of qualified care providers in the HCO increases, the probability of a patient being assigned to a specific care provider decreases.

4. Results

We evaluated our approach with the access logs of StarPanel, a longitudinal electronic patient chart developed and managed by the Vanderbilt University Medical Center (VUMC) [48]. StarPanel is ideal for this study because it aggregates all patient data as fed into the system from any clinical domain at the medical center and is the primary point of clinical information management. StarPanel's user interfaces are Internet-accessible on the Vanderbilt intranet and remotely accessible via the World Wide Web. The system has been in operation for over a decade and receives data from over 50 diverse clinical sources. For this study, we obtained 21 weeks of StarPanel access logs, during the year 2006. In summary, there were 9,940 distinct EHR users, who accessed the records of 350,889 distinct VUMC patients. There were a total of 7,575,434 accesses of patient information.[3]

In addition to the StarPanel access logs, we obtained the VUMC department assignments for the EHR users, as well as their start and end dates within the departments, from human resource records at the VUMC enterprise data warehouse. These records contained a total of 321 departments for 3687 of the EHR users in the study, or 36.9%. Despite the lack of meta-information for all patients, as we demonstrate below, studying this subset of the user population has minimal impact on the generalizability of our results.

4.1. Verification of Social and Organizational Phenomena

To determine that the StarPanel access logs could be useful for policy mining, we performed a series of experiments to verify the system yields rudimentary business processes and known social phenomena. It is important to verify that this particular EHR system has the ability to capture the business operations of the healthcare domain.

First, in an HCO such as a large academic medical center it is expected that most accesses are generated during the work week, when outpatient and non-critical treatment dominates

[3]For the purposes of this study, all logs were de-identified, such that each user and medical record number was consistently replaced with a random pseudonym.

operations. In anticipation of this phenomenon, we investigated the aggregate number of accesses to medical records generated per day by the EHR user population. The result, which is summarized in Figure 2, confirms this expectation. Notice that the number of accesses on weekdays is significantly greater than on weekends. The difference is approximately a factor of 5 across the entire study period.

Second, in an HCO, we expect users and patients interact according to known social phenomena. In general, we observed that the number of users accessing patients' records is consistent throughout the study period with an average of 6,406 users per week (standard deviation of 126 or 2.4%). The average number of departments observed each week is 292 (standard deviation of 4.5 or 1.5%). Next, to determine how users interact with patient records, we modeled the number of patients records each user accessed. In summary, the number of records accessed by a user in one week was 28.8 on average, with a median of 11. We observed that the distribution of the number of patient records accessed per user follows a power-law distribution (log-log linear), a property often seen in social networks [49]. Figure 3 depicts this phenomenon for an arbitrary week in the study, where 861 of the 6,389 users accessed a single patient's medical record only, while one user accessed 1097 records. This finding suggests that a large portion of the EHR user population interact with only a few patients' medical records each week, but there exists a relatively small number of "power users" who access hundreds of patients' records. We confirmed the notion of power users by examining the job titles of those users who were on the upper end of this distribution. Both users have jobs that entail accessing the medical records of numerous patients. Specifically, the role of the user who accessed over 1000 patients' records during the week in consideration is that of a medical records analyst in the *Emergency Medicine* department. Additionally, the user who accessed the second most records, over 600, is assigned the role of a medical coding specialist. Both of these occupations are often expected to work with higher volumes of patient records to facilitate insurance processing and quality assurance within a medical center.

Third, we expect the interactions of users in an HCO to reflect high-level organizational boundaries. For instance, it may be difficult to discern if two particular users or departments should interact, but it should be the case that general groups exist. Consider, the VUMC consists of several major hospitals, including a primary facility (i.e., the main hospital) and specialized facilities such as a children's hospital. It is anticipated that the majority of interactions between care providers will infer such structures. To verify if this was the case, we considered if the relational network between the users implies such groupings.

Figure 4A depicts the user network from an arbitrary weekday in the study period. On this day, approximately 900 users accessed 1,900 patients' records. In this network, we modeled users as nodes and added an edge if two users accessed at least one patient in common on the day in question. This is a simplification of the weighted network studied for association rule mining (as described in Section 3), but note that the goal of this portion of the investigation was to determine if there is a general organizational clustering of users.

We applied a principal components analysis (PCA) [50] to the network and projected the network on the first two components.[4] The result, depicted in Figure 4B revealed two major groups, with several users spanning the groups. Upon further inspection, we found the

[4]This was achieved by representing the network as an adjacency matrix of Boolean values. Each "dimension" of the matrix corresponds to the adjacency vector of a single user; i.e., a column in the matrix. PCA is a dimensionality reduction technique that reduces the dimensions of the matrix along orthogonal vectors which explain the variance in the system. As a consequence, each component is a vector of weights that characterize how each user contributes to the component. The principal components are sorted by the amount of variance they explain. In the social networking community, such methods are applied to derive the *eigencentrality* of entities in the system. [51]

majority of users in these groups correspond to the main university hospital and children's hospital, respectively, which suggests access logs can be analyzed to reveal high-level HCO structure. We note that it may be possible to uncover additional organizational relationships that are smaller in nature, such as particular wards or sections of a hospital, by modeling beyond the first two principal components. However, this aspect of an organizational analysis is beyond the scope of this study.

Similar results were observed for other randomly selected time points in the study period. Thus, our model appears to capture expected properties of HCOs and known phenomena in social networks.

4.2. Stability of HCO Interactions

We applied the relational model to extract association rules for EHR users and HCO departments. Figure 5 illustrates that the number of rules detected each week is consistent and significantly smaller than the maximum number of possible relationships. At the user-level, we detect on average 886,784 user-user interactions per week (standard deviation of 48,972, or 5.5%). This suggests that the user-level network is relatively sparse because the upper limit of user-user interactions is characterized by $\frac{|U|^2 - |U|}{2}$; since the average number of users per week was 6,406, the average upper limit of interactions is 20,515,215. Thus, only 4.32% of all possible user relationships are realized, which suggests non-random structure and affiliations within the organization.

For departmental rules, we ignored the users lacking departmental assignments, and detected 27,261 rules on average per week (standard deviation of 862, or 3.2%). In this case, the upper limit to the number of department-department relationships is $\frac{n^2+n}{2}$, where n is the number of departments. In comparison to user interactions the additional n accounts for the fact that departments can have relationships to themselves (i.e., two users from the *Emergency Department* accessing the same record. Using the average number of departments per week, 292, the upper limit is 42,778. Thus, our observation corresponds to 63.7% of the possible department relationships, indicating a greater amount of cross-community relationships at the department-level than at the user-level. Though we neglected a portion of the users for departmental analysis, we believe this has minimal influence on the finding because the two populations (i.e., those with vs. those without departments) have statistically indistinguishable sets of rules learned from the access logs. This claim is based on a Mann-Whitney U test we performed in which we rejected the null hypothesis, with a $p < 0.001$, that the two groups are different in terms of how many records each user in the two groups accessed. Additionally, when we compare a complete version of the user-level network to a version with all users who lack department information filtered out, we find that the ranks of the remaining rules do not change in relational to each other. Thus removing the users retains a meaningful sub-network.

When we examine the distribution of relationships per user for an arbitrary week in the study, as shown in Figure 6, we discover that it roughly follows a power-law distribution. This highly skewed distribution has an average of 139 rules per user and a median of 77. The distribution shows there are certain StarPanel users who are the lone person accessing a patient's record, whereas other users are connected to a very large number of other users. During the week, 118 users (1.8%) were the only users to access a specific patient's record and 95 users (1.5%) accessed records that only one other user accessed. At the other end of the spectrum, one highly connected person generated rules with 2584 other users. At the

department level, the number of rules per department in a week averages 98 with a median of 88.

4.3. Policy Decay

We visualized the stability of policies as a decay curve with respect to the user and department level rules. The results are summarized in Figure 7. For rules at the user level, we found that 56% exist for only a single week. Less than 1% of the rules exist for at least 14 weeks and only 0.07% of the rules exist for the entire study period. The steep decay regarding the longevity of rules stems from several possible causes, including the possibility that the organization is changing or that departments do not interact every week. Moreover, the VUMC is a large academic medical center, in which residents rotate through different clinical areas and in which care is delivered by teams. Therefore, we looked at department level abstractions of the users to determine if the delivery is by similar clinicians.

For departmental level rules, we see a much slower rate of decay in Figure 7. After one week 83% of the departmental rules remain, after 7 weeks 50% of the rules are retained, and over 16% of rules exist over the entire 21-week study period, which is much greater than the user rules. Notably, this represents a consistent and highly stable set of rules that could form the core of an access control system. A subset of such rules are depicted in Table 1.

The dynamic nature of the results indicates that even at the abstracted level, any statistical model of the HCO should be regenerated on a regular basis with recent access transactions.

4.4. Departmental Interaction Rules

Since rules at the user level lack HCO semantics, we examine rules at the departmental level. The departmental rules are more stable and therefore more likely to be incorporated into an HCO policy model. For the study period, a total of 58,415 department rules were discovered on a weekly basis. For space, we report on a small subset of the rules.

The rules with the highest confidence occurred for only a small part of the 21-week period. It is not until the 184th rank, according to confidence, that we discover a rule (*Office Of Research* ⇒ *Child & Adolescent Psychiatry*) that exists in all weeks of the study. To dampen noise in the system, we applied a filter to discard rules with a reoccurrence of less than three weeks.[5] The first portion of Table 1 reports the five highest post-filter confidence rules. In the filtered set, we obtain rules that are clinically intuitive. For instance, there is a 0.378 probability of an EHR user in the *Nephrology Clinic* accessing a patient who was also accessed by an EHR user in the *Hypertension Clinic*. This result is intuitive because nephrology, the branch of internal medicine that focuses on the kidneys, often deals with kidney diseases that include hypertension. As such, we expect coordination between these two departments.

In contrast, the rules with the highest support occur in every week of the study period. The top-5 rules are reported in the second portion of Table 1. Since the support represents the relative frequency of the pair of departments in a week, we expect to find departments that traditionally treat many patients to appear at the top of the list. This expectation is validated by the departments in Table 1, where we find *Emergency Medicine*. Additionally, the support of the top-20 rules capture 38% of the total support across all rules. For example, the rules with *Emergency Medicine* in the antecedent and consequent (i.e., *Emergency Medicine* ⇒ *Emergency Medicine*) account for nearly 6.5% of all rules detected in the study period. Moreover, 9 of the top-20 rules in this class have the same antecedent and consequent. This

[5]We chose three, as opposed to two, weeks to discard rules that spanned the end of the first and the beginning of the second week.

finding suggests there is a large amount intra-departmental interaction; i.e., people in the same department tend to work frequently with each other. This, however, is not always the case and some have different departments. For instance, the support of (*Allergy, Pulmonary, & Critical Care*, *Emergency Medicine*) is 0.013, that is it accounts for 1.3% of all the pairs of departmental co-accesses in the system. Yet, these relationships are also clinically intuitive.

We also observe HCO structure in the rules with the lowest support and confidence. Using the aforementioned filter, the rules with the lowest confidences are depicted in the third section of Table 1. Notice, again we find the *Emergency Medicine* department is a common antecedent for 17 of the 20 rules with the lowest confidences, which is not surprising because it collaborates with many other departments. From a clinical perspective, this implies that the emergency department functions as a triage point of the HCO. Furthermore, we find that the rules with the lowest support, depicted at the bottom of Table 1 are potentially important for HCO policies as well. This is because low frequency events may still lead to highly confident predictions of associations.

5. Discussion

In summary, our findings suggest that 1) HCOs are dynamic environments in which associations fluctuate over time, 2) departmental interactions are more stable than those of EHR users, and 3) intra-departmental relations tend to be more likely than inter-departmental. We have also shown that there are exceptions to these rules and that each policy should be monitored on a case-by-case basis. It is important to recognize that our results are not necessarily indicative of the policies at HCOs beyond the VUMC. Rather, each HCO should generate a customized statistical model of itself that is monitored over time.

Moreover, there are several limitations of our study that we wish to point out, some of which are due to the reality of data reuse and others based on the methodology itself. First, there are various possible extensions and refinements of our modeling technique that can be developed. For instance, the time windows upon which our networks are built can be designed to incorporate a patient's length of stay in the hospital (e.g., time of admission and discharge). Of course, it should be noted that a patient's record is often accessed after discharge for follow-up, billing, and processing. Moreover, we could impose more structure on user and departmental relationships by accounting for the ordering of events within the system. If an Emergency Department, for example, truly acts as a triage service, then it is anticipated that accesses from this department will precede other groups in the HCO.

Second, we suspect that our model can be improved upon through the incorporation of certain semantics associated with HCO information and by restructuring the access transactions. For instance, we only considered if a user accessed a patient or not, but it might be useful to consider a weighting schema that accounted for the number of times the user accessed a patient's record as well as how many user's accessed a particular record. From a semantic perspective, we could document which accesses led to care provider referrals, which could confirm existing work patterns and enable finer tuning of our statistical model. Similarly, additional improvements could result from incorporating a temporal view of the data. Consider, we might expect the emergency medicine department to access records first, then some other department such as internal medicine or surgery. From a restructuring perspective, we note that our models consider pairwise relationships between users, such that group detection via clustering techniques could provide rules that are more team-oriented.

Third, the study revealed that meta-information, such as departmental affiliations, is not a necessary condition to manage a serviceable EHR. Rather, it is often more appropriate for various services and programs that are outside the realm of EHR adminstration to extract this information from other data sources. For instance, in addition to the human resources database applied in this study, we discovered at least four distinct sources of partial meta-information at the medical center. The first source was a database designed to keep track of the medical residents. Residents often rotate through different care areas in the HCO, but this database was not designed to capture what department or care area each resident was assigned to and the dates over which that assignment was relevant. The second source was designed to allow customization of experience within the EHR based on a self-selected role; e.g. attending physician, staff nurse, etc. We believe that the combination of role and department would make for a compelling statistical model. For example, we could inquire if attending physicians in one department always work with charge nurses in a different department. Unfortunately, a majority of users never customized their experience and thus this system lacked role-related information on a significant majority of the user population. A third source contained departmental affiliations, but it lacked a primary key through which we could link the users in this source to the unique logins of the users in the access logs.

Finally, this work focused on automated policy extraction, but not on a rigorous human review of the resulting association rules. The number of rules generated by our approach will certainly be greater than capability of humans to process and assess. Nonetheless, though the resulting rules reflect how the EHR system was utilized by HCO employees, the rules themselves may be indicative of endemic problems within the HCO. Thus, we believe that the proposed approach could be evaluated by HCO administrators to assess the extent to which known policies are being followed. This will require an in-depth survey with experts within the HCO.

6. Conclusions

In this paper, we introduced a data-driven methodology to automatically extract HCO policies from on EHR access logs. The policies are probabilistic in nature and represent a wide range of relationships that transpire within the HCO. The methodology is generalizable and can be applied across HCOs, such that it can incorporate specific HCO knowledge about the EHR user population (e.g., user-department assignments). Beyond a theory, we empirically evaluated our approach with the access log of a large healthcare provider. Our investigations confirmed that policies at the HCO department-level afford greater stability than policies at the user-level. Moreover, our analysis verified the existence of certain expected business processes, such as the fact that intra-departmental interactions tend to occur more often that inter-departmental interactions and that intuitive collaborations (e.g., nephrology and hypertension clinics) manifest with relatively high likelihood. In future research, we intend on extending our methods to account for temporal components of access patterns and verify the methodology in additional clinical domains.

Acknowledgments

This research was funded in part by grants CCF-0424422 and CNS-0964063 from the National Science Foundation and R01-LM010207 from the National Library of Medicine, National Institutes of Health. We thank D. Giuse, D. Staggs, and E. Shultz for the data in this study. Additionally, we thank D. Aronsky, E. Boczko, J. Denny, C. Gadd, N. Lorenzi, R. Miller, J. Peterson, and G. Smith for their valuable comments regarding to this research.

References

1. Hersh W. Health care information technology: progress and barriers. JAMA. 2004; 292:2273–4. [PubMed: 15536117]
2. Wang S, Middleton B, Prosser L, Bardon C, Spurr C, Carchidi P, et al. A cost-benefit analysis of electronic medical records in primary care. Am J Med. 2003; 114(5):397–403. [PubMed: 12714130]
3. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med. 2006; 144(10):742–52. [PubMed: 16702590]
4. Stead, W.; Lin, H., editors. Computational technology for effective health care: immediate steps and strategic directions. National Research Council of the National Academies; Washington, DC: 2009.
5. Alam M, Hafner M, Breu R. Constraint based role based access control in the sectet-framework: A model-driven approach. Journal of Computer Security. 2008; 16(2):223–60.
6. Chu, S. From component-based to service oriented architecture for healthcare; Proceedings of the HEALTHCOM Workshop; 2005. p. 96-100.
7. Mathe J, Werner J, Lee Y, Malin B, Ledeczi A. Model-based design of clinical information systems. Methods Inf Med. 2008; 47(5):399–408. [PubMed: 18852913]
8. Mykknena J, Riekkinena A, Sormunena M, Karhunenb H, Laitinenc P. Designing web services in health information systems: From process to application level. Int J Med Inform. 2006; 76(2-3):89–95. [PubMed: 17118700]
9. Kohn, L.; Corrigan, J.; Donaldson, M. To Err is Human: Building a Safer Health System. National Academy Press; Washington, DC: 2000.
10. Corrigan, J. Crossing the Quality Chasm: A New Health System for the 21st Century.. National Academy Press; Washington, DC: 2001.
11. Blobel B, Nordberg R, Davis J, Pharow P. Modelling privilege management and access control. Int J Med Inform. 2006; 75:597–623. [PubMed: 16199198]
12. Lorenzi N, Riley R. Managing change: an overview. J Am Med Inform Assoc. 2000; 7(2):116–24. [PubMed: 10730594]
13. Bernard H, Killworth P, Kronenfeld D, Sailer L. The problem of informant accuracy: the validity of retrospective data. Annual Reviews in Anthropology. 1985; 13:495–517.
14. U.S. Dept. of Health and Human Services. Office for Civil Rights Standards for protetion of electronic health information; final rule. Federal Register. 2003; 164 45 cfr.
15. Lee W, Stolfo S. A framework for constructing features and models for intrusion detection systems. ACM Transactions on Information and System Security. 2000; 3(4):227–61.
16. Luo J, Bridges S. Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. Journal of Intelligent Systems. 2000; 15:687–703.
17. Cooley, R.; Mobasher, B.; Srivastava, J. Web mining: information and pattern discovery on the world wide web; Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence; 1997. p. 558
18. Srivastava J, Cooley R, Deshpande M, Tan P. Web usage mining: discovery and applications of usage patterns from web data. ACM SIGKDD Explorations Newsletter. 2000; 1:12–23.
19. Mobasher, B. Lecture Notes In Computer Science: The Adaptive Web: Methods and Strategies of Web Personalization. Vol. 4321. Springer-Verlag; 2007. Data mining for web personalization; p. 90-135.
20. Bucklin R, Sismeiro C. A model of web site browsing behavior estimated on clickstream data. Journal of Marketing Research. 2003; 40:249–67.
21. Johnson E, Moe W, Fader P, Bellman S, Lohse G. On the depth and dynamics of online search behavior. Management Science. 2004; 50:326–35.
22. Huang P, Lurie N, Mitra S. Searching for experience on the web: An empirical examination of consumer behavior for search and experience goods. Journal of Marketing. 2009; 73:55–69.
23. Hui S, Fader P, Bradlow E. Path data in marketing: An integrative framework and prospectus for model building. Marketing Science. 2009; 28:320–35.

24. Moe W, Fader P. Capturing evolving visit behavior in clickstream data. Journal of Interactive Marketing. 2004; 18:5–19.
25. Montgomery A, Li S, Srinivasan K, Liechty J. Modeling online browsing and path analysis using clickstream data. Marketing Science. 2004; 23:579–95.
26. Malin B. Correlating web usage of health information with patient medical data. Proc AMIA Symp. 2002:484–8. [PubMed: 12463871]
27. Chen E, Cimino J. Automated discovery of patient-specific clinician information needs using clinical information system log files. Proc AMIA Symp. 2003:145–9. [PubMed: 14728151]
28. Chen E, Bakken S, Currie L, Patel V, Cimino J. An automated approach to studying health resource and infobutton use. Stud Health Technol Inform. 2006; 122:273–8. [PubMed: 17102263]
29. Hripcsak G, Sengupta S, Wilcox A, Green R. Emergency department access to a longitudinal medical record. J Am Med Inform Assoc. 2007; 14:235–8. [PubMed: 17213496]
30. D'Alessandro M, D'Alessandro D, Galvin J, Erkonen W. Evaluating overall usage of a digital health sciences library. Bull Med Libr Assoc. 1998; 86:602–9. [PubMed: 9803306]
31. Dev P, Rindfleisch T, Kush S, Stringer J. An analysis of technology usage for streaming digital video in support of a preclinical curriculum. Proc AMIA Symp. 2000:180–4. [PubMed: 11079869]
32. Nieder G, Nagy F. Analysis of medical students' use of web-based resources for a gross anatomy and embryology course. Clin Anat. 2002; 15:409–18. [PubMed: 12373731]
33. Wasserman, S.; Faust, K. Social network analysis: methods and applications. Cambridge University Press; New York, NY: 1994.
34. Carley K. Computational organization science: a new frontier. Proc Natl Acad Sci U S A. 2002; 99(Suppl 3):7257–62. [PubMed: 12011404]
35. Malin B, Carley K. A longitudinal social network analysis of the editorial boards of medical informatics and bioinformatics journals. J Am Med Inform Assoc. 2007; 14(3):340–8. [PubMed: 17329730]
36. Merrill J, Hripcsak G. Using social network analysis within a department of biomedical informatics to induce a discussion of academic communities of practice. J Am Med Inform Assoc. 2008; 15(6):780–2. [PubMed: 18756000]
37. Merrill J, Bakken S, Rockoff M, Gebbie K, Carley K. Description of a method to support public health information management: organizational network analysis. J Biomed Inform. 2007; 40(4): 422–8. [PubMed: 17098480]
38. Zheng K, Padman R, Krackhardt D, Johnson M, Diamond H. Social networks and physician adoption of electronic health records: insights from an empirical study. J Am Med Inform Assoc. 2010; 17:328–36. [PubMed: 20442152]
39. Boyer L, Belzeaux R, Maurel O, Baumstarck-Barrau K, Samuelian J. A social network analysis of healthcare professional relationships in a french hospital. Int J Health Care Qual Assur. 2010; 23:460–9. [PubMed: 20845677]
40. Gray J, Davis D, Pursley D, Smallcomb J, Geva A, Chawla N. Network analysis of team structure in the neonatal intensive care unit. Pediatrics. 2010; 125:e1460–7. [PubMed: 20457681]
41. Lurie S, Fogg T, Dozier A. Social network analysis as a method of assessing medical-center culture; three case studies. Acad Med. 2010; 84:1029–35. [PubMed: 19638768]
42. Hassol A, Walker J, Kidder D, Rokita K, Young D, Pierdon S, et al. Patient experiences and attitudes about access to a patient electronic health care record and linked web messaging. J Am Med Inform Assoc. 2004; 11:505–13. [PubMed: 15299001]
43. Weitzmann E, Kaci L, Mandl K. Sharing medical data for health research: The early personal health record experience. J Med Internet Res. 2010; 12:e14. [PubMed: 20501431]
44. Wiljer D, Urowitz S, Apatu E, DeLenardo C, Eysenbach G, Harth T, et al. Patient accessible electronic health records: exploring recommendations for successful implementation strategies. J Med Interet Res. 2008; 10:e34.
45. Gallagher R, Sengupta S, Hripcsak G, Barrows R, Clayton P. An audit server for monitoring usage of clinical information systems. Proc AMIA Symp. 1998:1002.
46. Asaro P, Ries J. Data mining in medical record access logs. Proc AMIA Symp. 2001:855.

47. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record. 2003; 22:207–16.

48. Giuse D. Supporting communication in an integrated patient record system. AMIA Annu Symp Proc. 2003:1065. [PubMed: 14728568]

49. Newman M. The structure and function of complex networks. SIAM Review. 2003; 45:167–256.

50. Duda, R.; Hart, P.; Stork, D. Pattern Classification. 2nd edition. Wiley-Interscience; Maiden, MA: 2000.

51. Borgatti S. Centrality and network flow. Social Networks. 2005; 27:55–71.

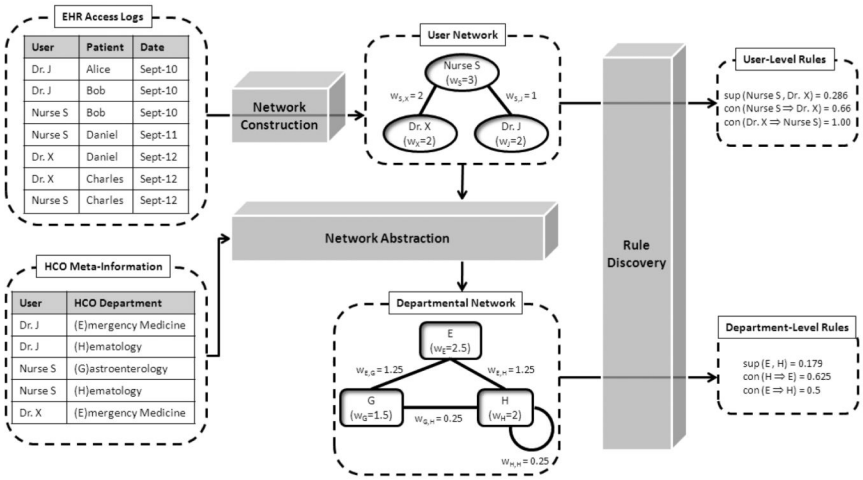

Figure 1.
Example of network construction from EHR access logs.

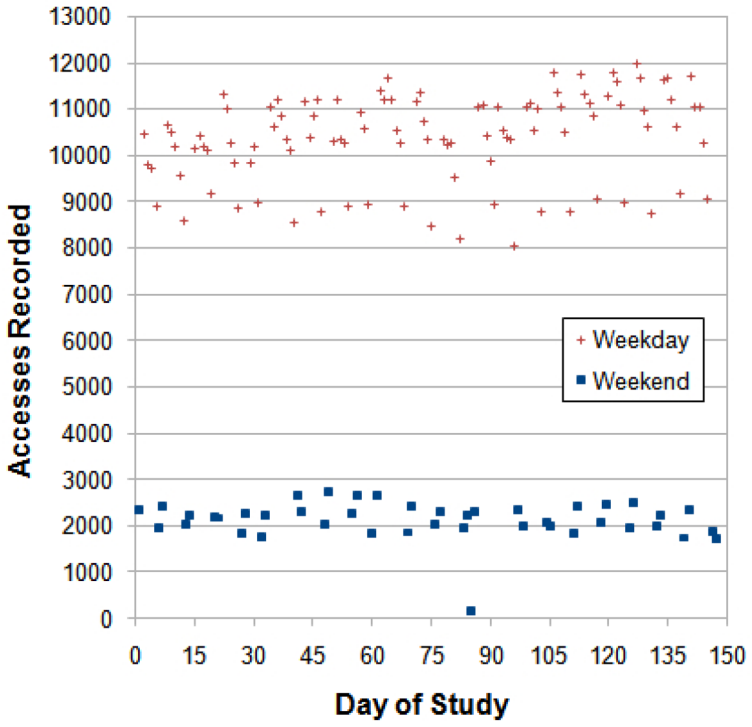

Figure 2.
Number of distinct patient records accessed per day during the study period.

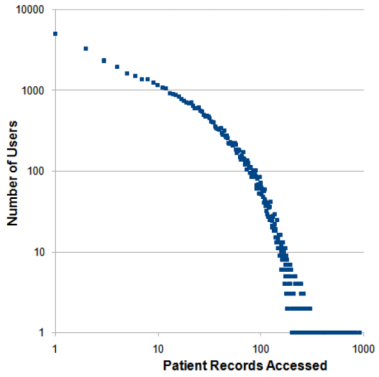

Figure 3.
Distribution of patients accessed by each user (log scale) during an arbitrary week in the study.

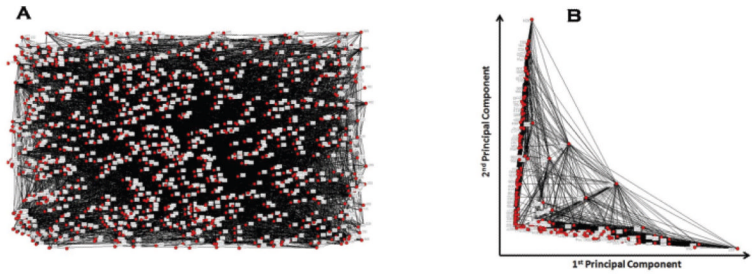

Figure 4.

A) Social network of EHR users for an arbitrary weekday in 2006. B) Projection of network against first two principal components. In these graphs, each node correspond to a distinct EHR user and each edge represents that two users accessed at least one patient in common.

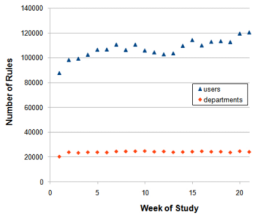

Figure 5.
Number of rules mined each week at the user and department levels.

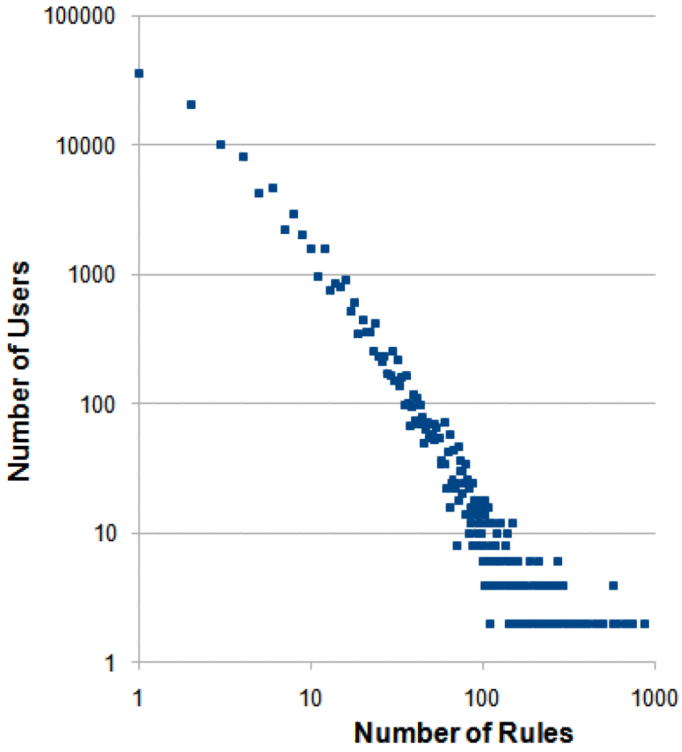

Figure 6.
Distribution of rules for each user (log scale) during an arbitrary week in April 2006.

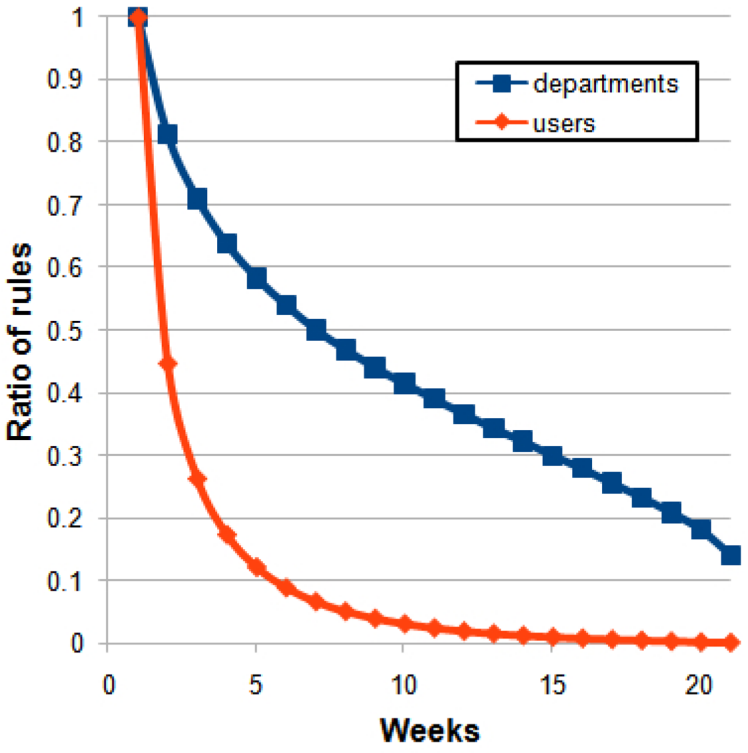

Figure 7.
Presence of rules throughout the lifetime of the study (interpolated).

Table 1

HCO departmental association rules. The scores correspond to the mean over the weeks when the rule was detected

Department *A*	Department *B*	Support (*A* and *B*)	Confidence (*A* ⇒ *B*)	Weeks
High Confidence Rules				
Organizational Development	School of Nursing	3.60×10^{-5}	0.4	7
Hypertension Clinic	Nephrology Clinic	4.16×10^{-5}	0.378	3
Business Office: Correspondence	Learning Center	5.71×10^{-6}	0.33	3
Business Office: Fees Group	Office of Compliance	7.94×10^{-6}	0.25	4
Main OR	Vanderbilt Medical Group - [Anonymized City]	5.94×10^{5}	0.2	6
High Support Rules				
Emergency Medicine	Emergency Medicine	6.48×10^{-2}	6.47×10^{-4}	21
Ophthalmology	Ophthalmology	2.85×10^{-2}	3.36×10^{-3}	21
Obstetrics & Gynecology	Obstetrics & Gynecology	2.69×10^{-2}	1.55×10^{-3}	21
Orthopedics & Rehabilitation	Orthopedics & Rehabilitation	2.06×10^{-2}	1.47×10^{-3}	21
Pediatric Hematology	Pediatric Hematology	2.06×10^{-2}	4.28×10^{-3}	21
Low Confidence Rules				
Emergency Medicine	Post Anesthesia Care Unit	2.31×10^{-6}	2.25×10^{-8}	5
Emergency Medicine	Developmental Disability Center	2.00×10^{-6}	2.29×10^{-8}	3
Emergency Medicine	Psychology & Human Development	2.00×10^{-6}	2.29×10^{-8}	3
Emergency Medicine	Cardiology	2.40×10^{-6}	2.35×10^{-8}	9
Emergency Medicine	Pulmonary Clinic	2.73×10^{-6}	2.93×10^{-8}	4
Low Support Rules				
Nursing Support Services	Medical Intensive Care Unit	5.11×10^{-7}	1.18×10^{-5}	3
Medical Intensive Care Unit	Nursing Support Services	5.11×10^{-7}	1.37×10^{-6}	3
General Surgery	Special Procedures	5.81×10^{-7}	4.45×10^{-6}	3
General Surgery	Hematology	5.81×10^{-7}	4.45×10^{-6}	3
Special Procedures	General Surgery	5.81×10^{-7}	4.98×10^{-4}	3