The Curious Case of the Medical Student - EMR Use and Anomaly Detection From Access Logs

Mirza S. Khan
Vanderbilt University
mirza.s.khan@vanderbilt.edu

ABSTRACT

Medical students represent a unique group of users who engage with the electronic medical record given the frequent change in clinical setting expected during their training. Failure to consider and manage this distinct group of users may lead to a high proportion of false positive anomalous cases. Using unsupervised learning methods, the pattern of EMR use among medical students can be discerned from other clinical colleagues. Existing anomaly and changepoint detection algorithms appear to be effective tools to identify deviations from typical medical student EMR engagement.

1. INTRODUCTION

Health care systems are required to maintain access logs for routine auditing and monitoring. Often, administrators responsible for monitoring access logs to detect improper and unwanted health record access have limited resources and time. With health record accesses in the millions per day for many health care systems, it is necessary to filter the signal from the noise to ensure a manageable workload for manual review [3]. There exists the trade-off between specificity and sensitivity of filtering; an excess may fail to capture misbehaving users and too little filtering may return an unmanageable number of records for manual review.

During their clinical training, medical students are exposed to a series of different clinical practice settings. This variation in hospital system setting can easily appear anomalous. Several medical students also engage in research activities that may include chart review and other electronic medical record (EMR) use. Given this unique EMR usage, medical students may appear anomalous relative to other EMR users and have their use flagged despite exhibiting appropriate behavior. Moreover, many cases of improper access may be related to curiosity, rather than malicious intent [2]. Such may also be the case for medical students, and assessing anomalous use within this sub-group may better inform improper EMR usage. All medical students face limitations with respect to work hours and what they may and may not be able to do in the EMR that can be useful features to identifying a distinct pattern among medical students compared to others. Given this, I hypothesized that EMR use among most medical students inferred from access logs will be comparable and that this could be used to construct a medical student EMR access fingerprint. In this work, I compared different unsupervised learning methods to determine if medical student EMR usage was distinct from others in a clinical practice setting. Then, I used established anomaly detection techniques to identify outliers based on access log information. One student identified as being an outlier was then examined further to determine why his/her accesses were considered a deviation. Last, I explored temporal changes in the number of accesses by two users to study how this may also be useful in identifying deviations in access behavior using changepoint detection.

2. BACKGROUND

Insider threats and inappropriate health record accesses are a significant possibility for health care systems. Medical students are no less susceptible to inappropriate access than other groups. Several earlier studies have used data-driven approaches for detection of anomalous EMR access behavior. The community based anomaly detection (CADS) method proposed by Chen and Malin uses an unsupervised learning framework to detect insider threats from access logs. CADS accounts for the team-based communities reflecting the practical structure of the health care system by inferring communities from relationships between EMR users and patient records [1]. Yet, this method identifies those users with low affinity to well-established communities as being anomalous. By design, medical students are transient members of communities, and may falsely be flagged by the CADS system. Other anomaly detection methods may overcome this limitation, but the specific case of medical students does not appear to have been explored. In particular the patient flow-based anomaly detection (PFAD) method uses past patterns for temporal modeling and assess deviations in workflows across different medical services [6]. Although not explcitly described in their work, applying PFAD to medical students as a distinct group may be a useful and robust strategy.

Supervised learning frameworks may be of limited utility in this domain. For instance, this would require labels, which are difficult and costly to acquire. Researchers and administrators may be unaware of all of the possibile expected anomalous patterns. Moreover, nefarious actors could then bypass detection by avoiding these anomalous approaches.

Unsupervised learning methods overcome this challenge and may be less susceptible to changes in EMR usage patterns, e.g. as may occur after an upgrade or other modifications are made to an organization's EMR system.

Another challenge involves the review of those accesses and users that are deemed to be anomalous. To assist with this process, Fabbri and Lefevre proposed the explanation-based auditing system [2]. Providing the rationale and contextual information for flagged records may assist with more rapid review and minimize the friction that otherwise exists in the review process.

3. METHODS

I retrieved a subset of access log data from Vanderbilt University Medical Center over a four month interval from August to November, 2019 using PostgreSQL. This interval was chosen to avoid many major holidays and a the beginning of the academic calendar when medical students may first be getting acclimated to the EMR system.

Medical students were identified as those employees with the medical student identifier from the available employee information. This approach was also used to identify attending physicians (PCPs), registered nurses (RNs) and medical assistants (MAs). Only those PCPs, RNs and MAs affiliated with the Adult Primary Care Clinic were included for this study. I assumed that those in the Adult Primary Care Clinic were engaged in typical behavior such that a user-patient access involves interaction with a given patient's chart on the date of the clinical encounter. I examined the subset of user-patient record accesses where a medical student and Adult Primary Care Clinic PCP, RN or MA accessed the same patient on the same day.

EMR access log records were dummy encoded to a sparse numeric representation containing count values for each action during each user-encounter interaction. Those instances where an action was missing were excluded. These null actions also tended to have a timestamp of '00:00:00.' Using the date and time information for the EMR accesses, I created additional features, including after hours (7PM to 6AM) and weekend accesses. To avoid collinearity that may affect some analytic procedures, certain datetime features were excluded during the analyses, e.g. business hours and weekday accesses. Given the size of the dataset, I used Dask for parallel and batch loading and processing of the data.

3.1 Clustering analysis

To perform group segmentation, the following clustering algorithms were applied to a dataset comprising the encoded access log records and another that also includes the date and time features described above. The following algorithms were applied: k-means clustering, agglomerative hierarchical clustering and hierarchical density-based spatial clustering of applications with noise (HDBSCAN). Using the medical student or non-medical student label for each employee, I computed the accuracy of each cluster and overall clustering accuracy for each algorithm. Hyperparameter tuning was performed to help optimize algorithm performance.

3.2 Anomaly detection

Using the access records and datetime features for medical student EMR accesses, I used the Local Outlier Factor (LOF) algorithm and k-nearest neighbors (KNN) anomaly detection from the from `scikit-learn` and `PyOD` libraries, respectively. For each algorithm, the number of neighbors parameter was set to 5. I compared the outliers from each method, and selected a shared outlier for further examination to determine why this user may have been considered to be an outlier.

3.3 Changepoint analysis

I selected two medical students for changepoint detection: the student with the greatest total number of EMR access actions and the student with the most days with more than one access over the 4 month study period. Mean changepoint analysis using the number of weekly accesses over the 4 month study period and the year 2019. I used the binary segmentation method for changepoint detection provided by the `changepoint` package [5].

4. RESULTS

This study included 748 distinct EMR users; 281 of whom were medical students. These users accounted for 2.72 million accessess across the shared 35,395 patients over the 4 month study period. Compared to PCPs, RNs and MAs, students accounted for 38.0% of accesses. Comparing accesses between students and PCPs, students make up 58.9% of accesses.

4.1 Clustering analysis

For k-means clustering comparing student and PCP accesses, maximum accuracy was approximately 92.0% with $k = 6$. This is shown in Figure 1 where I used PCA with 2 components to visualize the distinction in a two-dimensional space. Using agglomerative hierarchical clustering, the overall accuracy was 92.5% using the Ward method and Euclidean distance and a distance threshold of 25, which yields 6 distinct clusters. HDBSCAN resulted in an overall accuracy of 76.9% using a minimum cluster size of 2.

I compared students to PCPs, RNs and MAs using access log information alone and with the datetime features. Using access records information only, k-means clustering was maximized at $k = 9$ with an overall accuracy of 94.2%. Overall accuracy using agglomerative clustering was 94.7% using a distance threshold of 15 to 35 resulting in 19 to 5 distinct clusters, respectively. HDBSCAN was maximized at a minimum cluster size of 2 to 81.8% leading to 170 distinct clusters. Including the datetime features, overall accuracy from k-means clustering was greatest with $k = 10$ at an overall accuracy of 94.7%. Overall accuracy using agglomerative clustering was 95.1% using a distance threshold of 15 to 25 resulting in 19 to 9 distinct clusters, respectively. HDBSCAN was maximized at a minimum cluster size of 2, which yielded 180 distinct clusters.

4.2 Anomaly detection

For the anomaly detection task, the LOF algorithm identified one user who appeared anomalous. The KNN anomaly detection algorithm returned 27 distinct students as outliers. The sole user identified by LOF was also detected

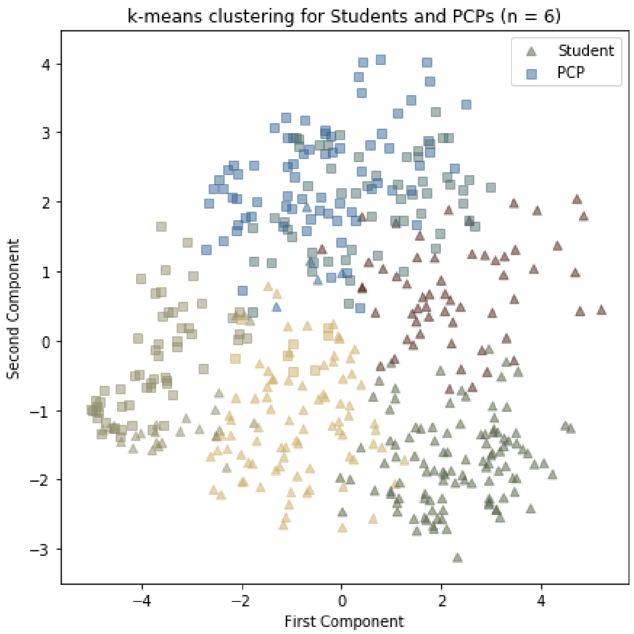

Figure 1: k-means clustering (k = 6) shown using the first and second components of a principal component analysis for a two-dimensional representation. Each of the 6 clusters is shown using different colors. True medical student labels are shown using a triangle marker; primary care physicians with a square marker.

using KNN. This user accessed patient records on 7 different days, ranging from 1 access to more than 1500 accesses (Figure 2). On further examination, his/hermost common access type was viewing a form; the fifth most common was 'research selection.'

4.3 Changepoint Analysis

For the student with the greatest number of accesses over the study period, there is a dramatic increase in the number of accesses (Figure 3A-B). On further examination, this individual had greater than 9,000 accesses in a single day. Figure 3C-D show the changepoints for the student with accesses on the greatest number of days over the 4 month study period.

5. DISCUSSION

We demonstrate that medical student EMR usage is discernible based on access records when compared to other clinical users, namely PCPs, RNs and MAs. Our findings show a marginal improvement by including datetime features. This may be related to some students interacting with the EMR due to participation in the student-run free clinic. Inspection of the clusters may reveal distinct phenotypes for medical students, such as those involved primarily in clinical activities, those who also rely on the EMR to conduct research, etc. This finding may be a useful instrument for role prediction for medical students based on access logs. For instance, it is unclear if the medical student designation expires at the time of a students expected graduation. An alternative approach could involve updating a users' role once it becomes evident that they are no longer behaving like a

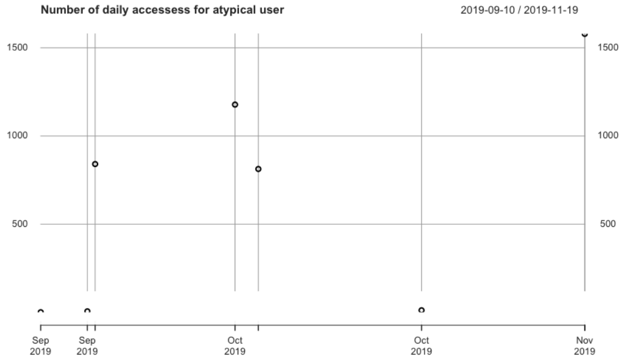

Figure 2: Number of accesses (y-axis) by date (x-axis) for the medical student identified as anomalous by both the Local Outlier Factor and k-nearest neighbors anomaly detection algorithms.

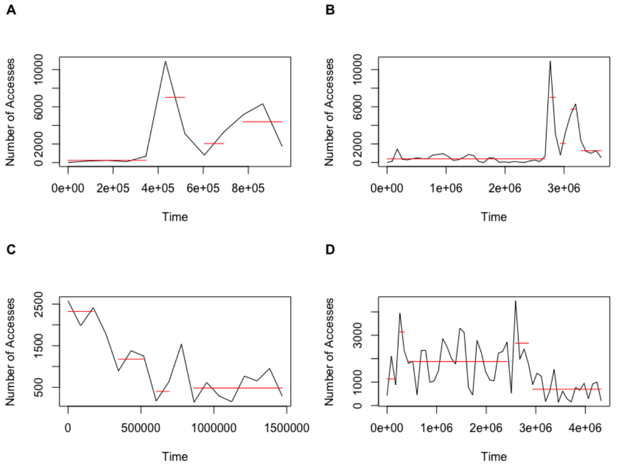

Figure 3: Mean changepoint detection using binary segmentation for the student with the greatest number of weekly accessess and the greatest number of days with at least 1 access from August to November, 2019 (A and C, respectively) and from the entire year 2019 (B and D, respectively).

student and engaging in a more senior role.

Treating medical students as a distinct group, anomaly detection methods successfully identified a medical student who was not engaging with the EMR like his/her peers. I validated that this individual was designated as a medical student in the system. On review of the most common access types this user was engaged in, s/he appeared to be conducting research-related activities and was not performing any frequent clinical duties. This may be appropriate use for first and second year medical students or those in the physician-scientist training program. One would expect to see some ebb and flow of EMR accesses by medical students based on the clinical rotation with a decrease around holidays or the examination period for each rotation. This expected pattern is highlighted in Figure 3C-D. By contrast, we find an atypical pattern for the user depicted in Figure 3A-B. For an administrator or compliance officer, this may be useful to exclude users with changepoints that occur at expected intervals, and flag those that with unexpected changepoints for further review. Presenting a figure, such as Figure 3A-B also provides context regarding an individual users typical behavior while emphasizing the where the change has occurred. Changepoint detection may be performed in real-time using tools that provide online learning. When using changepoint analysis or anomaly detection, one must weigh the cost of false positives and false negatives. This can then guide which method(s) and parameters may be most pragmatic to identify anomalous usage without being burdensome to those responsible for conducting manual review of access logs.

One approach to determine network structures and roles using the medical record may involve the use of metadata, such as department. Yet, classifying users based on department has some limitations. For example, some users, e.g. medical students and float nurses, may frequently change departments. Additionally, this method fails to capture some of the social complexities within health care institutions. For example, one PCP may work more closely with some specialists compared to others and oncologists who focus on lung cancer may work closely with select pulmonologists. As such, a network structure should be able to account for these relationships as described in Chen and Malin's SCAD framework [1]. I wished to use the label propagation algorithm (LPA) for community detection to illustrate this network structure and to examine how transient community members, such as medical students may be seen. A simple approach may be to flag all users who are outside of the dense community network(s), but I suspect that this may lead to many false positives. My effort to apply this algorithm was unsuccessful. This may be due to this particular sample of users and patients may reflect a weak community structure. LPA is known to produce an extremely large community that dominates the network and encompasses small communities [4]. Future studies should examine if properties within the network may be useful to identify deviations from expected usage.

There may be some limitations to this study. It is unclear to me what happens to the medical student designation upon graduation. There remains the possibility that the medical student attribute may remain for historical reasons, and, if so, this could have resulted in inappropriate labeling of medical students based on my query of the access logs. For example, residents at Vanderbilt University Medical Center who graduated from the medical school could have been mislabeled as medical students. Given the results of the clustering analysis, it is possible that these residents may represent one of the clusters designated as 'student' (Figure 1). Additionally, it is possible that the 4 month interval is too small of a window of study given that most patients are seen by their PCP every 6 months to 1 year unless they are acutely ill. Moreover, there are many other features available in the access log data that may be useful for role prediction and outlier detection. Introduction of these features may help improve performance.

6. CONCLUSION

We show that medical student EMR access log patterns are distinct from other clinical care providers. Using medical student access logs, existing changepoint and anomaly algorithms show promise in identifying and contextualizing deviations from typical medical student EMR usage. These findings suggest that detection of atypical EMR use can be effective when applied in isolation to this distinct group, and changepoint analysis may provide a useful tool to assist with review. This may augment existing outlier detection methods. Application of anomaly detection for medical students may benefit from analysis in isolation as they may not adhere to traditional community structures inferred from network analyses.

References

[1] Chen, Y. and Malin, B. 2011. Detection of anomalous insiders in collaborative environments via relational analysis of access logs. *Proceedings of the first ACM conference on Data and application security and privacy - CODASPY '11* (San Antonio, TX, USA, 2011), 63.

[2] Fabbri, D. and LeFevre, K. 2013. Explaining accesses to electronic medical records using diagnosis information. *J Am Med Inform Assoc.* 20, 1 (Jan. 2013), 52–60.

[3] Hedda, M. et al. 2017. Evaluating the Effectiveness of Auditing Rules for Electronic Health Record Systems. *AMIA Annu Symp Proc.* 2017, (2017), 866–875.

[4] Hu, X. et al. 2016. Role-based Label Propagation Algorithm for Community Detection. *arXiv:1601.06307 [physics].* (Jan. 2016).

[5] Killick, R. and Eckley, I.A. 2014. Changepoint: An R Package for Changepoint Analysis. *J. Stat. Soft.* 58, 3 (2014).

[6] Zhang, H. et al. 2013. Mining Deviations from Patient Care Pathways via Electronic Medical Record System Audits. *ACM Trans. Manage. Inf. Syst.* 4, 4 (Dec. 2013), 17:1–17:20.