CAESAR-ALE: An Active Learning Enhancement for Conditions Severity Classification

Nir Nissim[1], Mary Regina Boland[2], Robert Moskovitch[2], Nicholas Tatonetti[2], Yuval Elovici[1], Yuval Shahar[1], George Hripcsak[2]
[1]Information Systems Engineering, Ben Gurion University, Beer Sheva, Israel
[2]Biomedical Informatics, Columbia University, New York, New York, USA
nirni,elovici,yshahar@bgu.ac.il; mb3402,robert.moskovitch,nick.tatonetti,hripcsak@columbia.edu

ABSTRACT

Electronic Health Records (EHRs) are a treasure trove of health-related data. Prioritizing conditions extracted from EHRs is important for minimizing the burden on medical experts, who often need to manually review patient conditions for accuracy. Severity is useful for prioritizing and discriminating among conditions. Recently, a framework called CAESAR (Classification Approach for Extracting Severity Automatically from Electronic Health Records), for classifying condition-level severity, was proposed. However, it used passive learning that requires extensive manual labeling efforts by medical experts of each condition severity. We present CAESAR-ALE, an Active Learning (AL) based framework that uses several Active Learning (AL) to decrease the manual labeling efforts required. At each step in the algorithm, only the most informative conditions are labeled to train the algorithm. Our results show that our first method, which we refer to as Exploitation, reduced labeling efforts by 64% while achieving a true positive rate equivalent to that achieved by passive methods. Additionally, our second method that we refer to as Combination_XA reduced labeling efforts by 48% while achieving accuracy equivalent to that achieved by passive learning. Our proposed methods (Exploitation and Combination_XA) were superior in identifying a larger number of severe conditions, compared to SVM-Margin and the Random methods with a reduction of 46% in labeling efforts. As for the PPV (precision) measure, CAESAR-ALE achieved a 71% relative improvement in the predictive capabilities of the framework when classifying conditions as severe. These results demonstrate the potential of AL methods to decrease the labeling efforts of medical experts, while increasing accuracy given the same or even a smaller number of acquired conditions.

Keywords

Active Learning, Electronic Health Records, Phenotyping.

1. INTRODUCTION

Connected health is increasingly becoming a common framework to improve health service. An important component is diagnosing patients and labeling the severity of their diagnoses, which we focus in this study using active learning approach to decrease labeling efforts. Many national and international organizations study conditions and their clinical outcomes. The Observational Medical Outcomes Partnership (OMOP) standardized condition/phenotype identification and extraction from electronic data sources including Electronic Health Records (EHRs) [1]. The Electronic Medical Records and Genomics Network [2] successfully extracted some 20 phenotypes from EHRs for Phenome-Wide Association Studies [3]. Defining phenotypes from EHRs is a complex process because of discrepancies in definitions [4], data sparseness, data quality [5], bias [6], and healthcare process effects [7]. Currently, around 100 conditions/phenotypes have been successfully defined and extracted from EHRs. However, a short list of conditions ranked by priority (or severity) remains lacking.

To generate a prioritized list of conditions, we sought to rank them by their severity status by classifying conditions as either severe or mild at the condition-level. Classifying severity at the condition-level distinguishes acne as a mild condition from myocardial infarction (MI) as a severe condition. In contrast, patient-level severity assesses whether a given patient has a mild or severe form of a condition (e.g., acne). The bulk of the literature focuses on patient-level condition severity with many indices being unique to a given condition [8-11]. However, none of these indices capture severity at the condition-level.

Others developed methods to study patient-specific condition severity at the whole-body level, e.g., the Severity of Illness Index [12], for a wide range of conditions. This is useful for characterizing patients as severe or mild manifestations of a given disease condition. However, it does not measure severity at the condition-level (e.g., acne vs. MI), which is required to prioritize conditions by severity and thereby reduce the selection space to only the most severe conditions.

In this paper, we describe the development and validation of an Active Learning (AL) approach to classifying severity from EHRs. This builds on a previous passive learning approach called CAESAR (**C**lassification **A**pproach for **E**xtracting **S**everity **A**utomatically from Electronic Health **R**ecords). We call our Active Learning Enhancement of CAESAR, CAESAR-ALE, and we demonstrate that it reduces the burden on medical experts by minimizing the number of conditions requiring severity assignment. CAESAR-ALE works well in the biomedical domain by utilizing EHR-derived variables to assess severity of EHR-derived conditions.

2. BACKGROUND

In prior work, a classification method was developed called CAESAR (**C**lassification **A**pproach for **E**xtracting **S**everity **A**utomatically from Electronic Health **R**ecords) using a passive learning approach to capture condition severity from EHRs [13]. This method required medical experts to manually review conditions and assign severity status to each (severe or mild). They assigned severity to a set of 516 conditions included in the reference standard. These severity assignments were then used to evaluate the quality of the classifier. The review of conditions was limited to 516 conditions out of the 4,683 conditions included in the reference standard, because medical expert review is time-consuming and costly.

2.1 SNOMED-CT

SNOMED-CT (Systemized Nomenclature of Medicine-Clinical Terms) is a specialized ontology developed to capture conditions from EHRs obtained during the clinical encounter [14, 15]. SNOMED-CT is the terminology of choice of the World Health Organization and the International Health Terminology Standards Development Organization (IHTSDO). It also satisfies Meaningful Use requirements of the Health Information Technology component of the American Recovery and Reinvestment Act of 2009 [16], and often clinical ontologies are used for the retrieval of clinical guidelines [39]. Therefore, we used SNOMED-CT to extract patient conditions from EHRs treating each coded clinical event as a "condition" or "phenotype," knowing that this is a broad definition [4].

2.2 Classification of Conditions

Classification of conditions in the biomedical domain typically is based on two main methods: 1) manual approach where experts assign labels to conditions; and 2) passive classification approaches (typically supervised) where a dataset is labeled based on a subset of labeled data.

The Chronic Condition Indicator (CCI) was developed, as part of the Healthcare Cost and Utilization Project, using a totally manual approach [17] to assign chronicity categories (acute vs. chronic) to ICD-9 codes. Medical experts were asked whether or not a particular ICD-9 code was chronic. Disagreements were handled by consultation with one of the physician panel members [17]. The CCI built on original work by Hwang et al. [18], and it has been used successfully in multiple studies [18, 19] demonstrating the value of manual expert labeling.

Others employed passive learning approaches, including Perotte et al. who classified International Classification of Disease version 9 (ICD-9) codes and showed that leveraging the ICD-9 hierarchy outperformed treating ICD-9 codes as a flat list [20]. Another work by Perotte et al. classified conditions into chronicity categories [21]. Other machine learning approaches have been used in biomedicine typically in the subfield of text classification. Torii et al. showed that the performance of automatic taggers improved when trained on a dataset comprised of multiple data sources [22]. They also mention the need to have more documents available for training to improve performance [22], a common issue in passive learning techniques. Nguyen et al. built an algorithm for classifying lung cancer stages (tumor, node, and metastasis) using pathology reports and SNOMED-CT [23].

2.3 Active Learning

Labeling examples, which is crucial for the learning process, is often an expensive task since it involves human experts. Active learning (AL) was designed to reduce labeling efforts by actively selecting the examples with the highest potential contribution to the learning process of the classification model. AL is roughly divided into two major approaches: membership queries [24], in which examples are artificially generated from the problem space and selective-sampling [25], in which examples are selected from a pool. Selective-sampling is used in this paper. Studies in several domains have successfully applied active learning in order to reduce the time and money required for labeling examples. Unlike random learning, in which a classifier randomly selects examples from which to learn, in active learning the classifier actively indicates the specific examples that should be labeled and which are commonly the most informative examples for the training task.

Active learning approaches can be useful for selecting the most discriminative conditions from the entire dataset in order to minimize the number of conditions that experts need to manually review. Doing this focuses experts' efforts on a smaller subset of conditions, thereby saving time and money.

2.4 Active Learning in Biomedicine

Although applications of active learning algorithms have been widely demonstrated in other domains, their applications in the biomedical domain have been limited. Liu described a method similar to relevance feedback [26]. Warmuth et al. used a similar approach to separate active (positive side) and inactive (negative side) compounds [27]. More recently, active learning was reported to be useful in biomedicine for classification of text [28] and radiology reports [29]. In all these cases active learning methods were found to perform better than passive learning.

3. METHODS

3.1 Dataset Development

The dataset used in this study was developed as a result of prior work [13]. It contains 516 conditions (SNOMED-CT codes) labeled as mild and severe. The labeling was performed using a set of expert heuristics, described in detail elsewhere [13] and validated with five independent evaluators. The dataset also contains six severity measures for each condition: number of comorbidities, number of procedures, number of medications, cost, treatment time, and a proportion term. These six measures were used to classify conditions previously in the original passive learning method (CAESAR). Therefore, when constructing and testing CAESAR-ALE, we used exactly the same dataset used in developing and evaluating CAESAR.

3.2 The CAESAR-ALE Framework

The purpose of the CAESAR-ALE framework is to decrease the labeling efforts of an expert. Using active learning, the framework actively asks the expert to label a specific condition as severe or mild, rather than label randomly selected conditions. The workflow of the CAESAR-ALE framework is described in figure 1.

Figure 1 illustrates the framework and the process of labeling and acquiring new conditions by maintaining the updatability of the classification model. If the AL method finds the condition to be more informative than others, the conditions will be acquired for labeling. Conditions are collected and scrutinized within our framework. Then, they are transformed into a vector form (as with the CEASER method) for the advanced check. The conditions are then introduced to the classification model based on a Support Vector Machine (SVM) and Active Learning (AL) method. The classification model scrutinizes the conditions and provides two values for each condition: a classification decision using the SVM classification algorithm and a distance calculation from the separating hyper-plane using Equation 1. A condition that the AL method recognizes as informative and which it has indicated should be acquired is sent to an expert who labels it. The labeled conditions are then added to the training set for further use. By labeling the most informative conditions, we aim to frequently update and improve the classification model. Note that informative conditions are defined as those that when added to the training set improve the classification model's predictive capabilities.

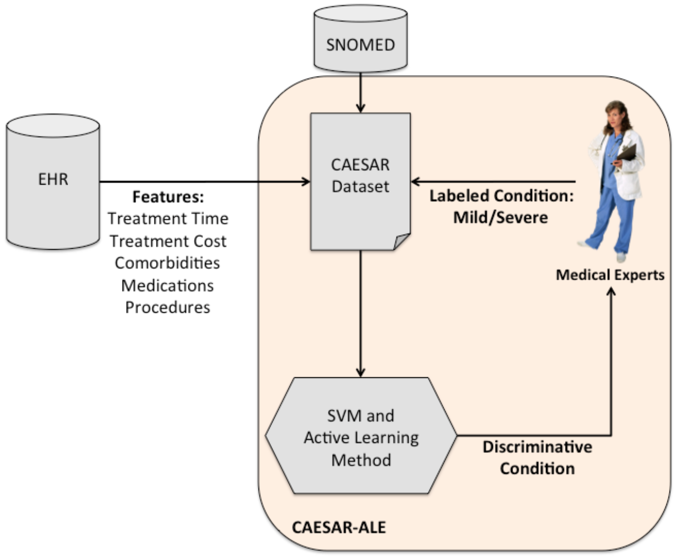

Figure 1: The process of using AL methods to detect discriminative conditions requiring medical expert annotation.

The AL methods are based on the predictive capabilities of the classification model, thus an updated classification model directly affects the AL method's ability to select the most informative conditions and by doing so decreases the labeling efforts; with few and well-selected labeled conditions we can maintain an accurate model and decrease the labeling efforts, in contrast to a situation in which the expert is required to labeled a large number of less informative conditions. Accordingly, in our context, there are two types of conditions that may be considered informative. The first type includes conditions in which the classifier has limited confidence as to their classification (the probability that they are mild is very close to the probability that they may be Severe). Labeling them would improve the model’s classification capabilities. In practical terms, these conditions will have new combinations of features (e.g., low in cost and requiring a long treatment time) or special combinations of existing features that represent their particular permutations. Therefore, these conditions will probably lie inside the SVM margin and consequently will be acquired by the SVM-Margin strategy that selects informative conditions, both severe and mild, that are a short distance from the separating hyper-plane. The second type of informative conditions includes those that lie deep inside the severe side of the SVM margin and are a maximal distance from the separating hyper-plane according to Equation 1. These conditions will be acquired by the Exploitation method (which will be further explained below) and are also a maximal distance from the labeled conditions. This distance is measured by the KFF calculation that will be further explained below as well.

The motivation underlying the selection of the conditions that are most likely to be classified as severe is based on two reasons: first, severe conditions have a higher medical and practical value, since they provide information about high priority conditions, those that should be treated more urgently. Second, by attempting to select conditions from deep within the "severe" instances sub-space of the SVM's separating hyper-plane, we may encounter a mild condition that was erroneously considered as being severe; the addition of this mild condition to the classifier provides highly valuable information, which greatly improves the classification model. Finally, the informative conditions are then added to the training set for updating and retraining the classification model (Figure 1). The framework integrates two main phases: training and classification/updating.

Training: A classification model is trained over an initial training set that includes both severe and mild conditions. After the model is tested over a test set that consists only of unknown conditions that were not presented to it during training, the initial performance of the classification model is evaluated.

Classification and updating: For every condition in the pool of unknown conditions the classification model provides a classification, while the AL method provides a rank representing how informative the condition is. The framework will then consider acquiring the conditions based on this. After being selected and receiving their true labels from the expert, the informative conditions are acquired by the training set (and removed from the pool). The classification model is retrained over the updated and extended training set that now also includes the acquired conditions that are regarded as being very informative. At the end of the update, the updated model again receives the pool of unknown conditions from which the updated framework and model again actively select informative conditions.

We employed the SVM classification algorithm using the radial basis function (RBF) kernel in a supervised learning approach. We used the SVM algorithm, because it has proven to be very efficient when combined with AL methods [26], [27]. In our experiments we used Lib-SVM implementation [30], because it also supports multiclass classification.

3.3 Active Learning Methods

Since our framework aims to provide solutions to real problems it must be based on a sampling method. We compared our proposed AL methods to other strategies, and all the methods considered are described below.

3.3.1 Random Selection (Random)

Random selection (also referred to as random learning or passive learning) is the default case in machine learning, in which the classifier is given by a set of labeled training examples. Thus, this is used as a baseline method.

While random selection is obviously not an active learning method, it is at the "lower bound" of the selection methods discussed. Consequently, we expect that all AL methods will perform better than a selection process based on the random selection of examples.

3.3.2 The SVM-Simple-Margin AL Method (SVM-Margin)

The SVM-Simple-Margin method [31] (referred to as SVM-Margin) is directly related to the SVM classifier. Using a kernel function, the SVM implicitly projects the training examples into a different (usually a higher dimensional) feature space denoted by *F*. In this space there is a set of hypotheses that are consistent with the training set, and these hypotheses create a linear separation of the training set. From among the consistent hypotheses (referred to as the version-space (*VS*)), the SVM identifies the best hypothesis with the maximum margin. To achieve a situation where the VS contains the most accurate and consistent hypothesis, the SVM-Margin AL method selects examples from the pool of unlabeled examples reducing the number of hypotheses. This method is based on simple heuristics that depend on the relationship between the VS and the SVM's maximum margin. The heuristics are used since calculating the VS is complex and impractical where large datasets are concerned. Examples that lie closest to the separating hyper-plane (inside the margin) are more likely to be informative and new to the classifier, and these examples are selected for labeling and acquisition.

This method selects examples according to their distance from the separating hyper-plane only to explore and acquire the informative conditions without relation to their classified labels, i.e., not specifically focusing on severe or mild conditions. The SVM-Margin AL method is very fast and can be applied to real problems, yet as its authors indicate [N-18], this agility is achieved because it is based on a rough approximation and relies on assumptions that the VS is fairly symmetric and that the hyper-plane's Normal (*W*) is centrally placed, assumptions that have been shown to fail significantly[32]. The method may query instances whose hyper-plane does not intersect the *VS* and therefore may not be informative.

3.3.3 Exploitation

We have developed a method, called "Exploitation", for efficient detection of malicious contents [34, 35], such as for malicious files [36, 38] or documents [37]. Exploitation is based on the SVM classifier principles and is oriented towards selecting examples most likely to be severe that lie furthest from the separating hyper-plane. Thus, this method supports the goal of boosting the classification capabilities of the classification model by acquiring as many new severe conditions as possible. For every condition *X* that is suspected of being severe, Exploitation rates its distance from the separating hyper-plane using Equation 1 based on the Normal of the separating hyper-plane of the SVM classifier that serves as the classification model. As explained above, the separating hyper-plane of the SVM is represented by *W*, which is the Normal of the separating hyper-plane and actually a linear combination of the most important examples (supporting vectors), multiplied by LaGrange multipliers (alphas) and by the kernel function *K* that assists in achieving linear separation in higher dimensions. Accordingly, the distance in Equation 1 is simply calculated between example *X* and the Normal (*W*) presented in Equation 2.

$$Dist(X) = \left(\sum_{1}^{n} \alpha_i y_i K(x_i\, x) \right)$$

(1)

$$w = \sum_{1}^{n} \alpha_i y_i \Phi(x_i)$$

(2)

In **Figure 2** the conditions that were acquired (marked with a red circle) are those conditions classified as severe and have maximum distance from the separating hyper-plane. Acquiring several new severe conditions that are very similar and whose values share nearly the same features is considered a waste of manual analysis resources. Thus, acquiring one representative condition for this set of new severe conditions will serve the goal of efficiently updating the classification model. In order to enhance the training set as much as possible, we also check the similarity between the selected conditions using the kernel farthest-first (KFF) method suggested by Baram et al. [33] which enables us to avoid acquiring conditions that are quite similar. Consequently, only the representative conditions that are most likely severe are selected. In Figure 2 it can be observed that there are sets of relatively similar conditions (based on their distance in the kernel space), however, only the representative conditions that are most likely to be severe are acquired. The SVM classifier defines the class margins using a small set of supporting vectors (i.e., conditions). While the usual goal is to improve classification by uncovering (labeling) conditions from the margin area, Exploitation's goal is to acquire conditions in order to enhance the detection of severe conditions. Contrary to SVM-Margin which explores examples that lie inside the SVM margin, Exploitation explores the "severe side" to discover new and unknown severe conditions that are essential for the detection of severe conditions which might cost more money and requires different treatment in hospitals (conditions which will likely become support vectors and update the classifier).

Figure 2 presents an example of a condition lying far inside the severe side that was found to be mild. The distance calculation required for each instance in this method is quick and equal to the time it takes to classify an instance in a SVM classifier, thus it is applicable for products working in real-time.

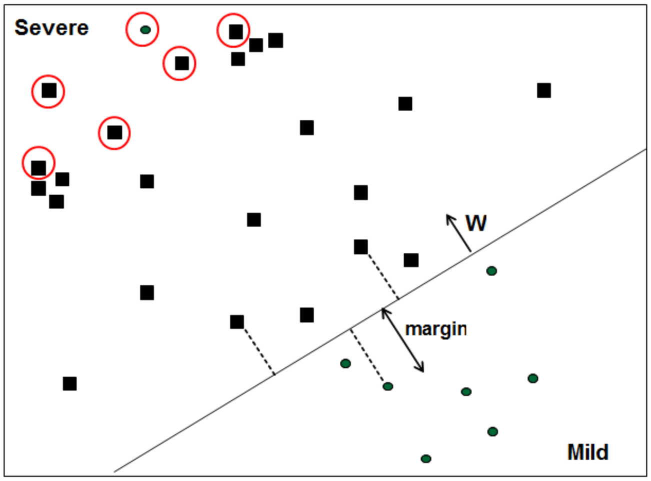

Figure 2: The criteria by which Exploitation acquires new unknown severe conditions. These conditions lie the farthest from the hyper-plane and are regarded as representative conditions.

3.3.4 Combination_XA: A Combined Active Learning Method

The "Combination_XA" method lies between SVM-Margin and Exploitation. It conducts a cross acquisition of informative conditions, which means the first trial it selects conditions according to SVM-Margin criteria and next trial it selects according to Exploitation criteria and so on with cross selection. Thus, on the odd trials the Combination_XA method selects examples based on SVM-Margin criteria in order to acquire the most informative conditions, acquiring both severe and mild conditions; this exploration phase is important in order to enable the classification model to discriminate between severe and mild conditions. While on even trials, the Combination_XA method then tries to maximally update the detection capabilities of severe conditions using the exploitation phase, drawing on the Exploitation method. On the one hand, this strategy is aimed at selecting the most informative conditions, both mild and severe, and on the other hand, it tries to boost the classification model with severe conditions or very informative mild conditions which are confusing and lie deep inside the severe side of the SVM's hyper-plane.

4. EVALUATION

The objective in our primary experiment was to evaluate and compare the performance of our new AL methods to the existing selection method, SVM-Margin, on the tasks of:

- Updating the predictive capabilities (Accuracy) of the classification model that serves as the knowledge store of AL methods and improving its ability to efficiently identify the most informative new conditions.
- Identifying which of the AL methods better improve the capabilities of the classification model to correctly classify the severe conditions (TPR) with minimal errors (FPR), a task which is of particular importance given the need to identify severe conditions from the outset.

During a variable number of acquisition trials that ended with acquiring every condition in the pool of unlabeled condition, we compared the acquisition of conditions based on AL methods to random selection based on the performance of the classification model. In our acquisition experiments we used 516 conditions (372 mild, 142 mild) in our repository and created 10 randomly selected datasets with each dataset containing three elements: an initial set of six conditions that were used to induce the initial classification model, a test set of 200 conditions on which the classification was tested and evaluated after every trial in which it was updated, and a pool of the remaining 310 unlabeled conditions, from which the framework and the selective sampling method selected the most informative conditions according to that method's criteria. The informative conditions were sent to a medical expert who labeled them. The conditions were later acquired by the training set that was enriched with an additional X new informative conditions. The process was repeated over the next trials until the entire pool was acquired. The performance of the classification model was averaged for 10 runs over the 10 different datasets that were created. Each selective sampling method was checked separately on four different acts of condition acquisition (each consisting of a different number of conditions). This means that for each act of acquisition, the methods were restricted to acquiring a number of conditions equal to the amount that followed, denoted as X: five conditions, 10 conditions, 20 conditions and 30 conditions.

The experiment's steps are as follows:

1. Inducing the initial classification model from the initial available training set (the initial training set includes six conditions).
2. Evaluating the classification model on the test set of 200 conditions to measure its initial performance.
3. Introduction of the pool of unknown and unlabeled conditions to the selective sampling method, which chooses the X most informative conditions according to its criteria and sends them to the medical expert for labeling.
4. Acquiring the informative conditions, removing them from the pool and adding them to the training set.
5. Inducing an updated classification model from the updated training set and applying the updated model on the pool (which now contains fewer conditions).

This process repeats itself on our dataset from first trial until the entire pool is acquired.

5. RESULTS

We evaluated the efficiency and effectiveness of our framework by comparing four selective sampling methods: 1) a well-known existing AL method, termed SVM-Simple-Margin (SVM-Margin) based on [27]; our proposed methods 2) Exploitation, 3) Combination_XA, and 4) random-selection (Random) as a "lower bound." Each method was checked against all four acquisition amounts, in which the results were the mean of 10 different folds. Due to space limitations we present the results of the most representative acquisition amount of five conditions in each trial.

We now present the results of the core measures in this study the accuracy, TPR, and the improvement in the classification model regarding these measures after each acquisition and retraining trial. In addition, we also measured the number of new severe conditions that were discovered and finally acquired into the training set. As explained above, five conditions (while the acquisition amount varied, we present only the most pertinent results from our experiments using an acquisition amount of five conditions because of page constraints) were selected from a pool of new unlabeled conditions during each trial of CAESAR-ALE. It is well known that selecting more conditions per trial will improve accuracy. However, we wanted to reduce the medical experts' efforts in labeling conditions, and therefore, we used AL methods to maximally improve the classification model's accuracy while minimizing the number of conditions acquired. More specifically, we used two of our methods (Exploitation and Combination_XA) to reduce the number of conditions selected by SVM-Margin.

Figure 3 presents the Accuracy levels and their trends in the 62 trials at acquisition level of five conditions per trial (62*5=310 conditions in pool). As can be seen, in most of the trials all of the AL methods outperformed Random. This shows that the use of AL methods can reduce the number of conditions required to achieve similar accuracy to the passive learning methods (i.e., Random). The classification model had an initial accuracy of 0.72, and all methods converged at an accuracy of 0.975 after the pool was fully acquired into the training set.

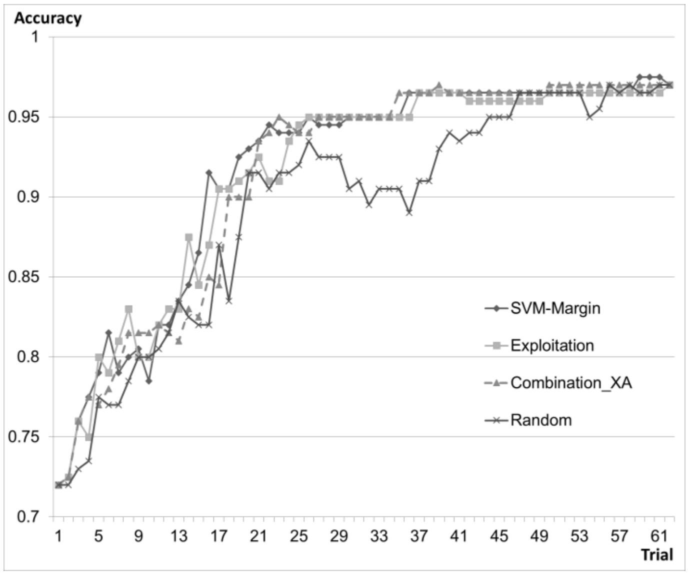

Figure 3: The accuracy of the framework over 62 trials for different methods (Five conditions acquired during each trial).

The first selection method that arrived at a 0.95 rate of accuracy was our Combination_XA method, which required 23 acquisition trials (acquisition of 115 conditions out of 310 conditions) while other AL methods required 26 trials. When compared to random selection, the Combination_XA method performed almost twice as well (23 vs. 44 trials) while achieving the same accuracy (i.e., 0.95).

Figure 4 presents TPR levels and their trends over 62 trials. Five new conditions are acquired during each trial (62*5=310 conditions in pool). Using TPR, Exploitation outperformed the other selection methods. It achieved a 0.85 TPR rate after only 17 trials (85 conditions out of 310), while random selection achieved a TPR of 0.85 after 47 trials.

In addition, the performance of the classification model improved as more conditions were acquired. After 36 trials, all AL methods converged to TPR rates around 0.92. Our results show that using AL methods to select discriminative conditions for classification can reduce the number of trials required for training the classifier. In turn, this will reduce the total number of conditions requiring medical expert review and thereby reduce costs.

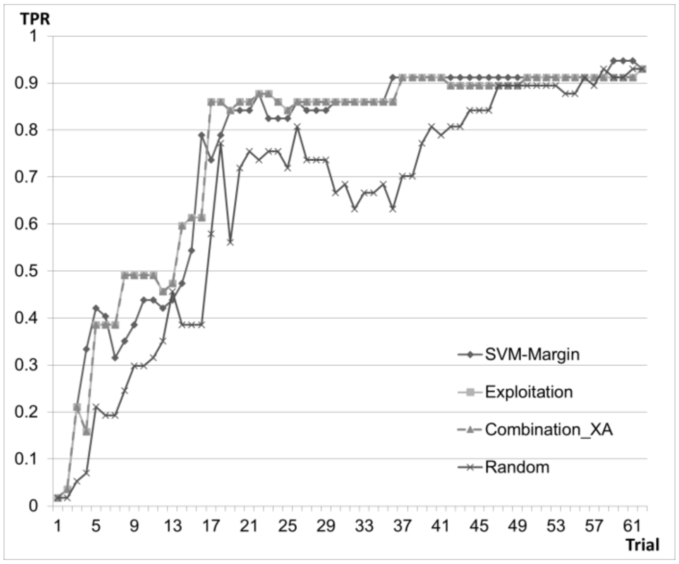

Figure 4: The TPR of the framework over 61 trials for different methods through the acquisition of five conditions in each trial.

We now present the results of another important measure in this study, the number of new severe conditions that were discovered and finally acquired into the training set. As explained above, during each trial the framework deals with pool of conditions beginning with 310 conditions, consisting of about 82 new severe conditions. Statistically, the more conditions selected during each trial, the more severe conditions will be acquired. Yet, using AL methods, we tried to improve the number of severe conditions acquired by means of existing solutions. More specifically, using our methods (Exploitation and Combination_XA) we also sought to improve the number of files acquired by SVM-Margin.

Figure 5 presents a cumulative number of severe conditions obtained by acquiring the five conditions during each trial by each of the four methods until the pool was fully acquired. From the fifth trial, Exploitation and Combination_XA outperformed the other selection methods (their graph intersects in Figure 5). It can be observed that after 23 trials (115 conditions) both of our AL methods acquired 73 severe condition out of the 82 severe conditions in the pool, whereas SVM-Margin and Random achieved it after 42 trials (210 conditions) and 60 trials (300 conditions), respectfully. This represents a reduction of 46% compared to SVM-margin and 62% compared to Random. The greatest difference between our AL methods and SVM-Margin was 15 severe conditions after 23 trials, while during the same trial we also observed the greatest difference between our AL methods and Random, a difference of 43 severe conditions after the 23 trials. The difference between our AL methods and Simple-SVM can be explained by the way this method acts: The SVM-Margin acquires examples about which the classification model is less confident. Consequently, they are considered to be more informative but not necessarily severe. As was explained previously, SVM-Margin selects informative conditions inside the margin of the SVM. Over time and with the improvement of the detection model towards more severe conditions, it seems that the severe conditions are less informative. Since these severe conditions might not lie inside the margin anymore, SVM-Margin may actually be acquiring informative mild conditions, rather than severe conditions.

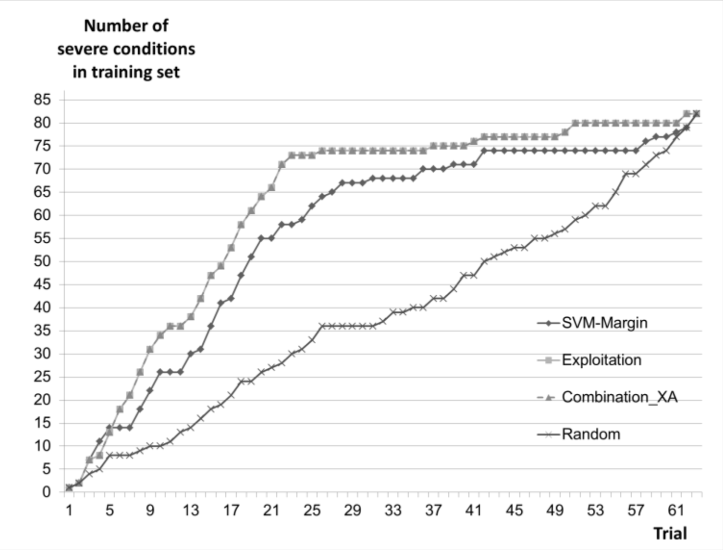

Figure 5: The number of severe conditions in the training set acquired by the framework for different methods with the acquisition of five conditions in each trial.

However, our methods, Combination_XA and Exploitation, are more oriented toward acquiring the most informative severe conditions by obtaining conditions from the severe side of the SVM margin. As a result, an increasing number of new severe conditions are acquired in the earliest trials, thus improving the classification model's performance and subsequently reducing labeling efforts; In addition, if an acquired mild condition lies deep within the severe side, it is still informative and can be used for learning purposes and to improve the upcoming trial's classification capabilities.

6. DISCUSSION AND CONCLUSION

We present a framework, CAESAR-ALE, used to detect informative conditions that, if labeled by experts, improve the classifier. We presented results from the lowest acquisition amount, because our primary goal was to minimize the number of conditions sent to medical experts for manual labeling.

Two different measures were mainly used to evaluate our algorithm: accuracy and TPR. TPR is important in severity classification, because of the great importance of detecting severe conditions. Therefore, because of the consequences inherent in misclassifying severity, it is better to classify a condition as severe when it is actually mild instead of classifying a condition as mild when in reality it is severe.

Bearing this in mind, traditional passive learning approaches require large amounts of training data to achieve sufficient performance. However, our Exploitation method achieved a TPR of 0.85 after only 17 trials. Only 85 conditions would require manual expert labeling in this scenario. In contrast, random selection required 47 trials or 235 conditions to achieve the same TPR (representing a 64% reduction). This would cut costs by almost two-thirds and allow medical experts to focus their energy elsewhere.

In terms of accuracy, the Combination_XA AL method performed the best with a reduction in the number of trials from 44 to 23 (when compared to random). If we translate this to the number of conditions, we find that the Combination_XA method required 115 vs. 220 conditions (representing a 48% reduction). Therefore, because for our purposes, FPR is less important (we don't mind calling some mild conditions severe as long as we accurately capture all severe conditions), we can reduce efforts and cost by 64% without compromising the classification performance. However, in some instances we may desire maximal accuracy, and in those cases we would still achieve a reduction in the number of trials required of 48% when using AL methods vs. passive learning.

Considering the number of severe conditions acquired across the trials, we observed that our methods (Exploitation and Combination_XA) were more successful at the identification of severe conditions, acquiring many more severe conditions in the early stages of the enhancement process, compared to the SVM-Margin and Random methods. The results showed a reduction of 46% in the number of acquired conditions needed for identification of most of the severe conditions.

The stronger acquisition performance of the Exploitation and Combination_XA methods can be explained by the way they function. Both methods have an exploitation phase during which they attempt to acquire conditions that are most likely severe. In fact, these two methods also acquire mild conditions that are thought to be severe. Although these mild conditions are indeed initially confusing, they are actually very informative to the classification model, since they lead to a major modification of the SVM margin and its separating hyper-plane. As a consequence, their acquisition improves the performance of the classification model better than the SVM-Margin method, which focuses on acquiring conditions that are known to be confusing, lead to only small changes in the SVM margin and its separating hyper-plane, and thus contribute less to improving the classification model. We understand from this phenomenon that there are often noisy "mild" conditions lying deep within what seems to be the sub-space of the "severe" conditions, as was explained in recent study that focused on the detection of PC worms [34]. As noted, these "surprising" cases are very informative and valuable to the improvement of the classification model (these conditions will probably become support vectors after acquiring them and retraining the model). In addition, they are helpful in the acquisition of severe conditions that eventually update and enrich the knowledge store. It should be noted that these conditions seem to be more informative than severe conditions, because they provide relevant information that was previously not considered (they were initially classified tentatively as severe by the classifier). (That is, the classifier initially considered them as being severe, but they were eventually discovered as being mild). It seems that our Exploitation and Combination_XA methods for acquiring conditions that are most likely severe induce a better classification model, a model that will eventually also acquire confusing but valuable and informative mild conditions.

When calculating the *Positive Predictive Value* (PPV) (i.e., Precision) of our enhanced framework CAESAR-ALE and comparing it to the basic approach CAESER, we observed that by acquiring 200 conditions (40 trials), CAESAR-ALE achieved a 96.6% PPV, compared to the 56.2% PPV that was achieved by CAESAR. This represents more than a 40% absolute, and a 71% relative improvement in the predictive capabilities of the framework when classifying conditions as severe, an improvement that was achieved along with the simultaneous significant reduction of labeling efforts.

In our future work, we hope to extend our efforts and provide a tool that prompts medical experts to label only pertinent and discriminative conditions. This should significantly reduce the workload of already busy clinicians.

In conclusion, we presented a framework called CAESAR-ALE that reduces the manual efforts of medical experts by identifying the most important phenotypes for labeling. Our AL framework reduced labeling efforts significantly, with a reduction by 64% for the same TPR, and 48% for the same accuracy level. We also demonstrated the strength of using AL methods on EHR data in the biomedical domain.

7. ACKNOWLEDGMENTS

This research was partly supported by the National Cyber Bureau of the Israeli Ministry of Science, Technology and Space. Support for portions of this research provided by **R01 LM006910** (GH).

8. REFERENCES

[1] Stang, P. E., Ryan, P. B., Racoosin, J. A., Overhage, J. M., Hartzema, A. G., Reich, C., Welebob, E., Scarnecchia, T. and Woodcock, J. 2010. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med.*, 153, 9 (Nov 2), 600-606.

[2] Kho, A. N., Pacheco, J. A., Peissig, P. L., Rasmussen, L., Newton, K. M., Weston, N., Crane, P. K., Pathak, J., Chute, C. G., Bielinski, S. J., Kullo, I. J., Li, R., Manolio, T. A., Chisholm, R. L. and Denny, J. C. 2011. Electronic medical records for genetic research: results of the eMERGE consortium. *Science translational medicine*, 3, 79 (Apr 20), 79re71.

[3] Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M. and Crawford, D. C. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26, 9 (May 1), 1205-1210.

[4] Boland, M. R., Hripcsak, G., Shen, Y., Chung, W. K. and Weng, C. 2013. Defining a comprehensive verotype using electronic health records for personalized medicine. *J Am Med Inform Assoc.*, 20, e2 (December 1), e232-e238.

[5] Weiskopf, N. G. and Weng, C. 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc.*, 20, 1, 144-151.

[6] Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A. and Melton, G. B. 2011. Bias associated with mining electronic health records. *Journal of biomedical discovery and collaboration*, 6, 48.

[7] Hripcsak, G. and Albers, D. J. 2013. Correlating electronic health record concepts with healthcare process events. *J Am Med Inform Assoc.*, 20, e2 (December 1), e311-e318.

[8] Rich, P. and Scher, R. K. 2003. Nail psoriasis severity index: a useful tool for evaluation of nail psoriasis. *Journal of the American Academy of Dermatology*, 49, 2, 206-212.

[9] Bastien, C. H., Vallières, A. and Morin, C. M. 2001. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Medicine*, 2, 4, 297-307.

[10] McLellan, A. T., Kushner, H., Metzger, D., Peters, R., Smith, I., Grissom, G., Pettinati, H. and Argeriou, M. 1992. The fifth edition of the Addiction Severity Index. *Journal of substance abuse treatment*, 9, 3, 199-213.

[11] Rockwood, T. H., Church, J. M., Fleshman, J. W., Kane, R. L., Mavrantonis, C., Thorson, A. G., Wexner, S. D., Bliss, D. and Lowry, A. C. 1999. Patient and surgeon ranking of the severity of symptoms associated with fecal incontinence. *Diseases of the colon & rectum*, 42, 12, 1525-1531.

[12] Horn, S. D. and Horn, R. A. 1986. Reliability and validity of the severity of illness index. *Medical care*, 24, 2, 159-178.

[13] Boland, M. R., Tatonetti, N. and Hripcsak, G. 2014. CAESAR: a Classification Approach for Extracting Severity Automatically from Electronic Health Records. *Intelligent Systems for Molecular Biology Phenotype Day*, Boston, MA, In Press, 1-8.

[14] Elkin, P. L., Brown, S. H., Husser, C. S., Bauer, B. A., Wahner-Roedler, D., Rosenbloom, S. T. and Speroff, T. *Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists*. Elsevier, City, 2006.

[15] Stearns, M. Q., Price, C., Spackman, K. A. and Wang, A. Y. *SNOMED clinical terms: overview of the development process and project status*. American Medical Informatics Association, City, 2001.

[16] Elhanan, G., Perl, Y. and Geller, J. 2011. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. *Journal of the American Medical Informatics Association*, 18, Suppl 1 (December 1, 2011), i36-i44.

[17] 2011. HCUP Chronic Condition Indicator for ICD-9-CM. *Healthcare Cost and Utilization Project (HCUP)*, https://http://www.hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp - download, Accessed on February 25, 2014.

[18] Hwang, W., Weller, W., Ireys, H. and Anderson, G. 2001. Out-Of-Pocket Medical Spending For Care Of Chronic Conditions. *Health Affairs*, 20, 6 (November 1, 2001), 267-278.

[19] Chi, M.-j., Lee, C.-y. and Wu, S.-c. 2011. The prevalence of chronic conditions and medical expenditures of the elderly by chronic condition indicator (CCI). *Archives of Gerontology and Geriatrics*, 52, 3, 284-289.

[20] Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F. and Elhadad, N. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21, 2 (March 1, 2014), 231-237.

[21] Perotte, A. and Hripcsak, G. 2013. Temporal properties of diagnosis code time series in aggregate. *IEEE journal of biomedical and health informatics*, 17, 2 (Mar), 477-483.

[22] Torii, M., Wagholikar, K. and Liu, H. 2011. Using machine learning for concept extraction on clinical documents from multiple data sources. *Journal of the American Medical Informatics Association*(June 27, 2011).

[23] Nguyen, A. N., Lawley, M. J., Hansen, D. P., Bowman, R. V., Clarke, B. E., Duhig, E. E. and Colquist, S. 2010. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association*, 17, 4 (July 1, 2010), 440-445.

[24] Angluin, D. 1988. Queries and concept learning. *Machine Learning*, 2, 319-342.

[25] Lewis, D. and Gale, W. 1994. A sequential algorithm for training text classifiers. *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag*, , 3-12.

[26] Liu, Y. 2004. Active learning with support vector machine applied to gene expression data for cancer classification. *Journal of chemical information and computer sciences*, 44, 6, 1936-1941.

[27] Warmuth, M. K., Liao, J., Rätsch, G., Mathieson, M., Putta, S. and Lemmen, C. 2003. Active learning with support vector machines in the drug discovery process. *Journal of chemical information and computer sciences*, 43, 2, 667-673.

[28] Figueroa, R. L., Zeng-Treitler, Q., Ngo, L. H., Goryachev, S. and Wiechmann, E. P. 2012. Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association*, amiajnl-2011-000648.

[29] Nguyen, D. H. and Patrick, J. D. 2014. Supervised machine learning and active learning in classification of radiology reports. *Journal of the American Medical Informatics Association*, amiajnl-2013-002516.

[30] Chang, C. C. and Lin, C. J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, 3, 27.

[31] Tong, S. and Koller, D. 2000-2001. Support vector machine active learning with applications to text classification. *Journal of Machine* Learning Research, 2, 45-66.

[32] Ralf, H., Graepel, T. and Campbell, C. 2001. Bayes point machines. The Journal of Machine Learning Research 1, 245-279.

[33] Baram, Y., El-Yaniv, R. and Luz, K. 2004. Online choice of active learning algorithms. . Journal of Machine Learning Research, 5, 255-291.

[34] Nissim, N., Moskovitch, R., Rokach, L., & Elovici, Y. (2012). Detecting unknown computer worm activity via support vector machines and active learning. Pattern Analysis and Applications, 15(4), 459-475.

[35] Nissim, N., Moskovitch, R., Rokach, L., and Elovici, Y., Novel Active Learning Methods for Enhanced PC Malware Detection in Windows OS, Expert Systems With Applications, 41(13), 2014.

[36] Moskovitch, R., Nissim, N., and Elovici, Y., Malicious code detection using active learning, ACM SIGKDD Workshop In Privacy, Security and Trust in KDD, Las Vegas, 2008.

[37] Nissim, N., Cohen, A., Moskovitch, R., Barad, O., Edry, M., Shabatai A., and Elovici, Y., ALPD: Active Learning framework for Enhancing the Detection of Malicious PDF Files aimed at Organizations, Proceedings of JISIC, 2014.

[38] Moskovitch, R., Stopel, D., Feher, C., Nissim, N., Japkowicz, N., Elovici, Y., Unknown Malcode Detection and The Imbalance Problem, Journal in COmputer Viorology, 5 (4), 2009.

[39] Moskovitch, R., and Shahar, R., Vaidurya: A Multiple Ontology, Concept Based, Context Sensitive Clinical Guideline Search Engine, Journal of Biomedical Informatics, 42 (1), 2009.