Computer-Assisted Cohort Identification in Practice

BESAT KASSAIE, ELIZABETH L. IRVING, and FRANK WM. TOMPA, University of Waterloo, Canada

The standard approach to expert-in-the-loop machine learning is active learning, where, repeatedly, an expert is asked to annotate one or more records and the machine finds a classifier that respects all annotations made until that point. We propose an alternative approach, *IQRef*, in which the expert iteratively designs a classifier and the machine helps him or her to determine how well it is performing and, importantly, when to stop, by reporting statistics on a fixed, hold-out sample of annotated records. We justify our approach based on prior work giving a theoretical model of how to re-use hold-out data. We compare the two approaches in the context of identifying a cohort of EHRs and examine their strengths and weaknesses through a case study arising from an optometric research problem. We conclude that both approaches are complementary, and we recommend that they both be employed in conjunction to address the problem of cohort identification in health research.

CCS Concepts: • **Social and professional topics** → **Personal health records**; • **Applied computing** → **Document searching**; • **Information systems** → **Retrieval effectiveness**;

Additional Key Words and Phrases: Electronic health records, subpopulation discovery, record annotation, inclusion and exclusion protocol, vision research, patient eye care, routine eye examinations, asymptomatic patients

ACM Reference format:

Besat Kassaie, Elizabeth L. Irving, and Frank Wm. Tompa. 2021. Computer-Assisted Cohort Identification in Practice. *ACM Trans. Comput. Healthcare* 3, 2, Article 17 (December 2021), 28 pages.
<https://doi.org/10.1145/3483411>

1 INTRODUCTION

Electronic Health Record (EHR) systems have been developed to provide individuals with high-quality and continuing health care. As well as supporting the needs of clinicians and health care providers, the data included in health records also has the potential to be included in broad-based studies to improve the lives of others. Researchers from various domains, such as public health, social science, and economics, could extract invaluable insights from EHR data.

The identification of patients having specific properties, *cohort identification*, is an important step in clinical research studies. To prepare data for each study, researchers determine some essential attributes about their population of interest. They then sift through medical files of patients for whom they have obtained appropriate consent in order to find patients with those specified attributes. In some cases, follow-up interactions are

We gratefully acknowledge financial assistance received from the University of Waterloo and NSERC, the Natural Sciences and Engineering Research Council of Canada.

Authors' addresses: B. Kassaie and F. Wm. Tompa, David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1; emails: {bkassaie, fwtpa}@uwaterloo.ca; E. L. Irving, School of Optometry and Vision Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1; email: elizabeth.irving@uwaterloo.ca.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

2637-8051/2021/12-ART17 \$15.00

<https://doi.org/10.1145/3483411>

conducted with the chosen patients, but often the clinical histories already contain all the data needed for the research study. Finally, the researchers run statistical analyses over the data for the selected subset of the cohort. The first step of this process, identifying a cohort, is time consuming and highly prone to errors, and the research is not easily reproducible due to the intensive manual work.

Because the patient population is dynamic and changing over time, there are always demands for repeating previous studies as well as conducting new studies in the medical research community. We propose a system to reduce the amount of manual work required for cohort identification. Providing researchers with a reliable framework to work with EHRs entails addressing many problems.

Generally, EHR systems comprise two types of data elements: structured data (such as age and date of examination) and unstructured or semi-structured data (such as family history and diagnoses). Unstructured data is inherently difficult to process by machine. Unfortunately in most EHR systems, critical medical information is recorded as text, with its breadth of expressiveness as well as many spelling errors and non-standard abbreviations. Thus, cohort identification often relies on some understanding (or computational processing) of data fields stored as text [3, 13, 14, 35, 36, 39, 43, 49, 56, 57].

The University of Waterloo's School of Optometry and Vision Science hosts an Optometry Clinic, which is one of the largest vision care centers in Canada. As well, researchers in the School of Optometry and Vision Science conduct a variety of studies based on patients' medical data [45]. Thus, the problem of cohort identification is an important and recurring problem for researchers in the school.

A study performed a few years ago at Waterloo's Optometry Clinic [30] used the **Waterloo Eye Study (WatES)** dataset, extracted from (paper) patient records and covering 6,397 patient visits in 2007 [44], to determine the percentage of asymptomatic patients for which a routine eye examination uncovers critical changes in their eye-related status (stratified by age). The main objective of that study was to determine whether there is any evidence to propose new guidelines for the frequency of routine eye examinations for *asymptomatic* patients. Such data-driven guidelines can (and should) be re-examined periodically and potentially updated or refined as the population's characteristics change.

Since the time that the earlier study was conducted, the clinic adopted an EHR system to maintain its patient records. To repeat the study, we were tasked to identify the set of patient visits (examinations) deemed to be "asymptomatic routine eye exams," that is, those for which the patient had no eye-related problems (e.g., headaches or blurry vision) or diseases (e.g., glaucoma) and was not specifically instructed to have an examination. (The subsequent analysis to be performed on this "asymptomatic REE" cohort includes the time since the last visit, the patient's age, and whether or not the examination revealed the presence of new disease, a significant change of prescription, or a change in patient management, but these are used to define sub-populations within the cohort and are not part of the cohort identification criteria.) In the optometry EHR system, most of the information related to the cohort inclusion and exclusion criteria occurs in free-text fields, such as "*Chief Complaint*" or "*Reason for Visit*." For each examination record, either before or after selecting patients based on the free-text fields, there may be additional structured fields that can be used straightforwardly to apply inclusion/exclusion criteria.

In this article, we present our solution for this particular encounter-level cohort identification problem, which we believe to be applicable to many medical cohort identification problems in practice. For example, health care researchers repeatedly encounter the need to find similar (or dissimilar) characteristics among patients; they often wish to identify clusters of cases (based on symptoms, chronology, and/or outcomes), and they often wish to track cohort outcomes and natural progression retroactively. Of course, this need is not limited to searching optometric records but extends to searching for any disease within any health record system to determine the prevalence of that disease, condition, or outcome within the database.

Adelman noted that "Case studies have been routinely criticized on external validity grounds for, the argument goes, how can one generalize from a sample size of one? Yet, the same criticism can be made for attempting to generalize from a single experiment; it is just as precarious" [1]. We suggest that because many practitioners in

the health and medical research industry will face challenges similar to the ones we addressed, they will be able to adapt our solutions to fit their situations.

In the following sections, we explain the architecture of our system and decisions that have been made along the way.

- We explain why conventional experimental approaches to addressing the problem are of limited assistance in practice (Section 2).
- We present **Iterative Query Refinement (IQRef)**, an alternative to active learning for expert-in-the-loop cohort identification (Section 3).
- Our approach relies on a hold-out sample to assess the accuracy of classifiers, and we justify its use based on others' prior work (Section 4).
- We present the details of our case study for a difficult, exemplary cohort identification problem, in which both active learning and IQRef were applied, and we highlight the strengths and weaknesses of both approaches in addressing the problem. We conclude that both approaches are valuable and complementary (Sections 5 and 6).
- We recommend that both approaches to expert-in-the-loop be used together to address complicated cohort identification problems in health research (Section 7).

Our major research contributions to the state of the art in computing for healthcare are the design of the IQRef system, the strong recommendation to use a hold-out sample for evaluating cohort selection criteria, the observation that adaptive data analysis techniques must be incorporated so as not to overfit the hold-out data, and the final recommendation to use IQRef in conjunction with active learning to achieve the best results.

2 EXPERIMENTAL SYSTEMS FOR MEDICAL COHORT IDENTIFICATION

The goal of automating cohort identification is to reduce manual examination of medical records [62]. The complete elimination of such examination is impossible, however, because to verify any solution, we need to have a ground truth against which to compare the computer-produced results. The task of preparing a ground truth, regardless of techniques that are deployed to solve the problem of cohort identification, must get reduced to labeling a set of records, i.e., annotation. There are considerations that make it challenging to annotate medical records. *Crowdsourcing*,¹ a common way to annotate records, is ruled out because of the confidentiality of personal medical records, training required to understand medical records, and requirements for very high-quality annotations, as the research outcomes will be used in critical health-related tasks. However, as the level of expertise and authority of an annotator increases, the cost of annotation increases significantly. Therefore, for any possible solution, the number of samples to annotate manually by domain experts must remain within a limited budget.

To assist in addressing this need, the **Text Retrieval Conference (TREC)** sponsored a Medical Records Track in 2011 and 2012. Systems participating in the track were required to identify a cohort of records that match a short description. Detailed explanations can be found in Voorhees's summary [68] and in the TREC proceedings including that track [70, 71].

The data for the track includes a set of 17,250 independent hospital visits, each with a free-text field for "chief complaint," a discharge diagnosis code, and a set of 1 to 415 medical test reports (e.g., radiology, cardiology, and discharge reports). The median number of reports per visit is three.

Cohort criteria for a single retrieval task (a "topic" in the track) are expressed as a single brief phrase such as "Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression," and track participants are required to return a ranked list of 100 visits in decreasing order of likelihood of meeting the criteria.

¹Employing non-professionals with minimum cost to obtain the labels.

In 2012 (the second, and last, year for the track), 24 participants submitted a total of 88 runs (6 of which required some manual effort) against 47 topics. The best eight runs (two of which required manual intervention) had an “inferred nDCG score” (*infNDCG*, an estimate of how well the ranked result compares to the best possible ranking) between .49 and .68 and a precision at 10 (*P@10*, the fraction of the top 10 ranked responses that are judged to be correct) between .52 and .75. Importantly, the best *infNDCG* scores per topic varied between 0.2 and 1.0, and these best results were achieved by different systems. Voorhees reports: “Also typical for retrieval performance is that the difficulty of a topic, as measured by the evaluation scores obtained for it, is independent of the number of relevant visits it has” [68].

In her summary, Voorhees concludes: “As anticipated, the search results demonstrate that language use within electronic health records is sufficiently different from general use to warrant domain-specific processing. Top-performing systems each used some sort of vocabulary normalization device specific to the medical domain to accommodate the array of abbreviations, acronyms, and other informal terminology used to designate medical procedures and findings in the records. The use of negative language is also much more prevalent in health records (e.g., patient denies pain, no fever) and thus requires appropriate handling for good search results” [68].

As practitioners who must identify a cohort of records, we can learn much from the experimental performance reported for these systems, but, unfortunately, there are several reasons we cannot adopt a solution directly.

- (1) The emphasis for TREC’s Medical Records Track was on precision (returning records that mostly satisfy the cohort criteria, especially toward the start of the ranked list) rather than recall (returning most of the records that satisfy the criteria). However, for our purpose, we need not only a set of visits that meet the criteria but also as large a set of such visits as possible—ideally, we want to distinguish all patient visits that were for asymptomatic routine eye examinations from all other patient visits. With system performance in the track based on ranked results and performance reported in terms of *infNDCG* and *P@10* only, there is no indication whatsoever where to cut off an arbitrarily large ranked list to determine how well a system is able to identify the complete cohort.
- (2) The experimental results do not identify which system or approach might be most suited for our specific problem.
 - System performance depends critically on the topic statement. We could use a phrase as short as “asymptomatic routine eye examination” or a paragraph or two such as our explanation in Section 1. How should our cohort’s inclusion and exclusion criteria be stated so as to maximize the likelihood of a system’s success?
 - The relative success for each system in identifying a cohort varies widely across the topics. It is therefore unclear which system will work best for our particular problem. In fact, there is little information to help us predict how well each system would perform on our specific task.
 - There is no indication of how each system’s performance will transfer to a new corpus. Document size in the TREC track is significantly larger than in our optometry EHR setting. Many of the participating systems emphasized how to manage and weight multiple reports (which are not present in our corpus) and how to use outside sources to expand medical vocabulary (but MeSH and similar medical resources are weak in optometry-specific vocabulary and acronyms). Changes in these characteristics are bound to affect the systems’ performance.
- (3) Finally, neither the track organizers nor the participants have described a procedure or mechanism to evaluate the trustworthiness of the results. After a system returns a set of records, how is the health researcher to determine the validity of that set vis-à-vis the required cohort?

The first of these problems has been addressed by TREC’s Total Recall Track in the context of several corpora (one of which is health related) [25, 52]. Unlike other tracks, this one uses an iterative approach to assessing document relevance. Following this so-called *expert-in-the-loop* architecture (see Section 3), as participants explore the corpus, they periodically submit a document to an expert to provide an assessment on the fly. The “cost” is

the number of assessments requested and the goal is to achieve high recall with low cost. A major difficulty is to determine when to stop the iteration, and it is suggested that high recall vs. cost is achieved when stopping after $2R + 1,000$ documents have been assessed, where R is the number of results judged up to that point to be relevant. Such an approach might be feasible for cohort identification in situations where there are very few members in the cohort being sought. However, if the cohort is relatively large, the cost would be prohibitive.

We see no way to overcome the second problem: when trying to find a particular cohort, we do not have the luxury of trying many different systems and many different statements of our problem for each system to determine which combination seems to perform well. Our budget (in terms of the time commitment required of the expert) is limited. Furthermore, without addressing the third problem, we have no criteria to compare multiple systems' experimental performance to determine which one outperformed the others on our problem.

The answer to the third problem is to follow the well-accepted protocol of measuring performance against a hold-out sample of data [61] labeled as being in the desired cohort or being outside that cohort. Unfortunately, when there are few relevant documents, assessing members of a random sample of the collection is likely to include very few (and quite possibly no) positive relevance judgments. It is for this reason that when comparing the performance of various information retrieval systems, the practice used for TREC and other evaluation laboratories is to pool the top results for a variety of systems and to have experts assign relevance judgments to members of the pooled set [69]: if sufficiently many systems compete with diverse search strategies, nearly all relevant documents are likely to be ranked sufficiently highly by at least one system.

Based on the earlier study [30], we expect our cohort to comprise approximately one-third of the patient visits. This means that in any random sample of visits we can expect that approximately one-third will be in our cohort, so we can adopt the protocol of measuring performance against a hold-out sample. Furthermore, we cannot adopt the suggested stopping criteria from the Total Recall Track: assuming high recall and a collection size of fewer than 10,000 documents, once we have assessed $2R + 1,000$ documents, we might as well assess the remaining quarter or so of the collection. On the other hand, because we can rely on a meaningful hold-out sample, we have a way to compare systems' performance and a mechanism to monitor the development of suitable values for input parameters or to manage expert-in-the-loop iterative solutions. This is elaborated in Section 4.

3 EXPERT-IN-THE-LOOP ARCHITECTURES

Many researchers have explored how to incorporate human expertise in machine learning problems [28]. For example, Guo et al. have examined how experts could support semi-automated medical image grouping [26], and Giraldi et al. have investigated how experts could assist in semi-automated knowledge discovery [22]. In this section we describe two approaches to using expert participation when classifying EHRs.

3.1 Active Learning

Pool-based active learning has been devised to address machine learning in contexts where labeling data is expensive [38, 58]. Active learning has been successfully adopted in the legal domain to support e-discovery, where annotation is extremely expensive and high recall is the primary retrieval objective [12]. It has also been proposed for classifying health records using combinations of structured fields and terms extracted from text fields [11, 72].

With this approach, there is assumed to be a large set of unlabeled data D . After some pre-processing, the system chooses a small subset of records S_0 as seeds to show to the (human) annotator and initiates the learning process. The classifier (machine) uses S_0 to find a function f_0 that discriminates sample records in the class from the other records in S_0 . Then we apply f_0 to all the records in $D \setminus S_0$ and select a sample S_1 to show to the annotator. If the discriminator is deemed inadequate, the annotator will decide to continue learning by annotating a few more records from the returned sample and then use all the records annotated so far $S_0 \cup S_1$ to find a new classifying function f_1 . These last three steps (classifying the records, annotating an additional sample, training a new classifier) are then repeated until some stopping condition is met. When the discriminator

is deemed to have satisfactory performance (or we have exceeded an annotation “budget”), we end training and use the learned discriminator f_n to extract the final results from the sub-population that is the subject of further analysis.

Clearly the success of any machine learning approach, including active learning, depends on the success of finding a classifier that distinguishes records in the target class of interest from those not in that class. **Support Vector Machines (SVMs)** have been successfully employed in active learning contexts to find good text classifiers [66]. When using SVMs, the success of finding a good classifier depends on selecting an appropriate set of features to represent each record (including appropriate weights to reflect the relative importance of those features) and selecting an appropriate kernel for (effectively) transforming the chosen feature space into a feature space in which a linear separator can be found. Luckily, for text classification, feature and kernel selection is usually straightforward [17, 31].

As well as choosing an appropriate classification mechanism to use, a second critical aspect of active learning is choosing informative samples to show to the human annotator, the so-called *query strategy*. Several strategies have been proposed to formulate sample selection, such as *uncertainty sampling* [38], *query-by-committee* [60], and *expected model change* [59], among others. The intention is to find a sequence of example records that help to choose a good classification without the need for annotating many records in all.

A third major aspect of active learning is to determine when to stop the annotation cycle. For this purpose it would be helpful to have an accurate estimate of errors imposed by the process. Webber describes how to estimate the confidence of the recall metric in an active learning setting and presents experimental evidence indicating how large a sample is needed and whether to bias the sample toward documents matching a predicate or documents not matching the predicate in order to increase confidence in the recall reported for the sample [73]. In this article, because the cohort being sought is often fairly large, we suggest instead to use an accuracy estimate provided by a hold-out sample of records.

3.2 An Alternative: Iterative Query Refinement

Health practitioners describe patients’ characteristics and attributes with short phrases (“lost her glasses,” “occasional dryness”), sometimes expressing positive characteristics, and sometimes their absence (“no migraines,” “– pain”). Thus, it is natural for them to identify a cohort using terms (words or short phrases) chosen from an application-dependent *termset* $\mathcal{E}$. In fact, looking back at the 2012 TREC Medical Records Track, the top-performing system used a manual approach in which phrase-based queries “were interactively modified ... until either the top ten retrieved documents appeared mostly relevant or no relevant documents could be found.” [16] However, there is no explanation of how this exploration was conducted nor how many iterations were used (i.e., how much expert time was required), nor how the performance of the top 10 results can be translated into a measure of overall accuracy.

In a recent study examining approaches to patient-level cohort discovery from compound electronic health records, it was found that “structured Boolean queries, searching over unstructured and [structured] data, outperformed word-based automated methods over the same data” [8]. Again, looking back at the 2012 TREC Medical Records Track, the system with the fourth-best performance “took a manual, interactive approach to the task, and focused on the construction of a search interface that would allow its users to rapidly formulate queries, review their results, and iterate. Using our system, users could search through the corpus by any of the various fields (chief complaint, report text, etc.) using a robust search syntax, and could also include ICD-9 codes in their queries. This allowed for the easy construction of queries representing complex Boolean criteria” [5]. Furthermore, the experts who created the topics for the track searched the corpus “using a Boolean retrieval system to develop an estimate of the number of relevant visits in the document set” [68] (i.e., the approximate size of the cohort). Thus, it would seem that a Boolean query language over words and short phrases is likely to be an effective approach that is also appealing to experts in the medical and health domains.

Fig. 1. Cohort identification with IQRef.

As a result, in consideration of the challenges and constraints imposed by our problem, we propose IQRef, an architecture supporting interactive, adaptive data analysis [18, 27] with an expert in the loop (Figure 1). Similar to the standard setup for active learning, we start with a database of documents $\mathcal{D}$ (the *population*) from which we select a subset (the *training set* $\mathcal{T}$) to be used for developing the cohort identification function f —a Boolean expression over terms in $\mathcal{E}$ comprising words and short phrases (Figure 2)—and a fairly small annotated sample $S \subset \mathcal{D}$, where $S \cap \mathcal{T} = \emptyset$. However, unlike active learning, S is closed; i.e., neither deletion nor insertion occurs. Rather than the expert in the loop annotating additional records and the computer learning which Boolean function is a good separator, the expert learns how to distinguish cohort records and the computer serves as an oracle reporting the performance of a candidate function Q on S through calls to one or more pre-determined, black-box verification functions ϕ_j .

It is important to note that IQRef requires domain experts to serve in two distinct roles: an *annotator* who tags a sample in advance and a *learner* who builds the Boolean discriminator function, occasionally testing potential functions against the annotated sample. We require the annotator and the learner to have similar knowledge about the records and the information need; i.e., both interpret the same record features to be significant. Nevertheless, it is critically important that the learner not have any knowledge of the individual records in S beyond what is revealed by the calls to any of the verification functions ϕ .

Fig. 2. Boolean retrieval model.

Using IQRef, the learner devises Boolean predicates² over $\mathcal{E}$ that can be used to retrieve and examine a few EHRs $D_j^+ = \{d_1^+ \dots d_{m^+}^+\} \subseteq \mathcal{T}$ that satisfy the predicate and a few other EHRs $D_j^- = \{d_1^- \dots d_{m^-}^-\} \subseteq \mathcal{T}$ that do not satisfy the predicate. Each predicate is oblivious to any characteristic of an EHR except for the absence or presence of terms. Thus, either all or none of the records with identical phrases are among the retrieved set (Figure 2), and results are not ranked. In practice, only phrases and terms with bounded length b appearing in relevant text fields in the EHRs are interesting, which means that the class of potential discriminator functions is finite.

The ultimate goal of IQRef is to develop a predicate with minimum error to isolate a cohort of patients from the population. This is accomplished in two phases. In the first *exploratory* phase, after examining the records in D_j^+ and D_j^- , the expert chooses additional terms to include in the predicate to distinguish EHRs in the cohort from the rest. When satisfied after several iterations that Q_i serves as a suitable explanation for capturing a subset of the cohort records [32], it can be saved and Q_{i+1} can be begun to capture another subset. In the second, *aggregating* phase, the chosen sub-queries can then be combined with disjunction so that the aggregated expression $Q = \bigvee Q_i$ represents a classifying function with both high precision and high recall. The exploratory and aggregating phases can be arbitrarily interleaved, depending on whether the expert is attempting to build a sub-query or combine sub-queries into more complete queries.

During the query building process, the user may ask the system to evaluate some verification function(s) $\phi_j(Q)$ over S , the pre-selected hold-out sample of documents. Because these have previously been judged by an expert to be in or not in the cohort, the system can give an estimate of how well the predicate matches the desired EHRs. No document in S is retrieved and examinable by the user, but based on the outcome of $\phi_j(Q)$, she may or may not decide to terminate building Q_i or Q . Upon termination, she saves Q to be used as the cohort identification function f , applies f to the target subpopulation $D \subseteq \mathcal{D}$, and uses those extracted records to complete her intended research.

²Because the predicates are used to retrieve a subset of patient records, we use the terms *predicate*, *classifying function*, and *query* interchangeably to refer to the user-formulated Boolean expressions.

Fig. 3. Exploration mode.

We now describe the IQRef Learner's Interface in more detail.

3.2.1 Exploration. The first phase of operation is exploratory, in which the learner attempts to create a predicate that has high precision in identifying the target cohort, without worrying about recall. In this phase, the expert learner can create, run, validate, edit, and save Boolean expressions over phrases, where a phrase comprises one or more terms separated by blanks.

There are three text boxes used to specify a query (Figure 3). Every record that is retrieved meets the criteria provided in the nonempty text boxes, i.e., contains all the phrases appearing in “*All words/phrases*,” any of the phrases in “*Any words/phrases*,” and none of the phrases in “*None of words/phrases*.” Multiple phrases in a box are separated by commas. To match the indexed terms, each phrase is similarly normalized: stop words, such as was and on, are removed and all tokens are converted to lowercase. For example, if $p_1, \dots, p_i$ are in the topmost box, $q_1, \dots, q_j$ are in the middle box, and $r_1, \dots, r_k$ are in the bottom box, the corresponding query is $(\hat{p}_1 \wedge \dots \wedge \hat{p}_i) \wedge (\hat{q}_1 \vee \dots \vee \hat{q}_j) \wedge (\neg \hat{r}_1 \wedge \dots \wedge \neg \hat{r}_k)$, where $\hat{x}$ is the normalized form of x .

The learner may view the set of records in $\mathcal{T}$ that match the query by pressing the “Search” button. To avoid any bias, the set of records that match the Boolean query is presented in random order. Similarly, by pressing the “Browse Rejects” button, the learner may view the set of records in $\mathcal{T}$ that do *not* match the query. The number of examination records and the number of distinct patients matching the query are also displayed. The expert learner can browse through those records to determine how she might improve the query by adding or removing terms from any of the boxes in order to better discriminate among the records she sees.

As well as examining some qualifying (or disqualified) records, the learner can also explore *similar* terms to add to either the middle or bottom boxes (Figure 4). This is particularly useful, because there are many misspellings, synonyms, and non-standard abbreviation used in the EHRs [67]. For each of the terms in a specified box, 10 suggested terms are chosen based on the application of word2vec [46] trained on the records in $\mathcal{T}$. Thus, two terms are considered similar if they have appeared sufficiently often in similar contexts in those records. With

Fig. 4. Similar terms.

Fig. 5. Aggregation mode.

this technique, many standard and non-standard abbreviations (such as fbs in the first two rows and ee and cee in the sixth row of the figure) and alternatively spelled and misspelled terms (such as itchy in the second row and haloes in the third row) will appear in the lists of suggested synonyms. In fact, it is because of this simple ability to choose variant words that we have omitted stemming when creating our terms: for example, the last row shows that review and reviews can be included as alternative terms independently.

Using these screens, the expert learner iteratively develops a query to retrieve records that belong to the patient cohort. She can evaluate the current query over the previously annotated sample S , without seeing any of those records. The metric that is reported is *implication*, which represents the fraction of records from S that had been annotated as relevant, i.e., belongs to the cohort of patients, or that do not satisfy the learner's current predicate. This statistic helps the learner to decide whether to refine the query further. Once the learner is satisfied by the query's performance with respect to implication, she saves the query as a potential disjunct for the final classifier function.

3.2.2 Aggregation. Having created one or more predicates that imply membership in the cohort, the learner uses the second phase of operation to decide whether the records in the union of records from those predicates

are, in fact, sufficiently accurate to identify members of the cohort. In the aggregating phase, she builds, edits, and evaluates disjunctions over queries saved from the exploratory phase (Figure 5). The goal of this phase is to build an aggregate query with high accuracy. As in the exploratory phase, the learner can evaluate any potential aggregate query over S , again without seeing any of those records. In this phase, however, the performance of each aggregate query is reported using *accuracy* over the annotated sample. If this statistic is deemed to be satisfactory, the query is exported so that it can be applied to the desired subset of $\mathcal{D}$ to retrieve the patient cohort for the remainder of the research study. Otherwise, the learner can try to combine and test alternative sets of saved queries or return to the exploratory phase to continue to develop further queries to be used as additional or replacement disjuncts.

4 USING HOLD-OUT DATA TO ACHIEVE PERFORMANCE GUARANTEES

Because health management is a sensitive domain, researchers must be able to trust the cohort identification system. In another context, it has been observed that “Whilst every effort is made to ensure that the training and validation data captures the features present in the clinical setting, evidence is required to verify that the model will continue to perform as expected when deployed for real world diagnosis. To provide such assurances requires the test data to be both representative of the clinical setting and independent of the training data and learning process” [50]. Translating this requirement to cohort identification, the performance of a system (or for Boolean systems, a predicate) to identify a cohort during the development of its input and tuning parameters must be indicative of its performance when it is finally deployed, that is, when it is applied to the dataset as a whole. More specifically, those using the system need to know an estimate of the error that is included by solutions, possibly expressed as a bound on the size of the error. Approaches that are oblivious to or ignore this requirement fail to give confidence to researchers.

The proposed process (Figure 1) is similar to a learning mechanism. It is assumed that the set of EHRs to be used in the research study (the population $\mathcal{D}$) is generated by an unknown mechanism with unknown probability distributions and that the given set of EHRs (the training data $\mathcal{T}$ and the hold-out test data S) are representative of that population. The goal of learning is to find a Boolean expression that minimizes the error in identifying the target cohort over the training data $\mathcal{T}$ while generalizing to the population $\mathcal{D}$. That is, the cohort identification expression should isolate the cohort represented by $\mathcal{T}$ and S well, but not over-fit them, as we would like to re-apply the same expression to identify similar target cohorts from unseen (future) collections of EHRs, as long as the text in those records is similar to what is present in $\mathcal{T}$ and S .

To create S , a predetermined number of records n is picked through simple random sampling [61]; i.e., they are **independently and identically distributed (i.i.d.)** according to the underlying distribution. The annotator tags all the records in the sample S as being “in the cohort,” “outside the cohort,” or (possibly in a few cases) “undecided.” We remove the undecided records from the sample and define the Boolean predicate $C(r)$ to be 1 on record $r \in S$ if and only if the tag on r is “in the cohort.” The resulting annotated sample is then *held out*, so that the only way to acquire information about those records is through statistical queries.

For the purpose of providing feedback about S , statistical queries [34] are defined in terms of a Boolean predicate p , such that when p is evaluated on an individual record r , it returns 1 if r satisfies p and 0 if it does not. When such a query is evaluated on any set D drawn randomly i.i.d. from $\mathcal{D}$, the statistical value returned is the mean of the predicate value for each of the records in D , i.e., $\frac{1}{|D|} \sum_{r \in D} p(r)$.

In classical machine learning, the learner finds the best discriminator based on $\mathcal{T}$ and then evaluates its performance using S . The purpose of separating the training set $\mathcal{T}$ from the test set S is to avoid over-fitting the discriminator f to the training data, and therefore developing a discriminator that does not generalize to the whole population $\mathcal{D}$. To be able to describe the performance of f on $\mathcal{D}$ with confidence, it is vital that its development does not overly depend on the observed performance of f on S . In classical machine learning, this is

accomplished by ensuring that all queries posed against the hold-out data S must be chosen non-adaptively (i.e., they may not depend in any way on any answers to previous queries run against S).³

Our major hindrance is that we have very limited data with annotations: the purpose of using our approach at all is to avoid the need to annotate more than a few EHRs. Thus, we cannot afford to have records in $\mathcal{T}$ be annotated. This rules out, for example, the use of cross-validation to avoid over-fitting. Because the records in $\mathcal{T}$ are not annotated, we need an adaptive approach on which to build training.

In our architecture, the learner is human rather than an algorithm. Nevertheless, her objective falls within the *Empirical Risk Minimization* paradigm [61]. We trust the learner to estimate the error of various hypotheses over the unlabeled training data $\mathcal{T}$ based on her expertise. The limited set of labeled records is used to give an unbiased estimate of performance of hypothesized discriminators on the population to avoid over-fitting on $\mathcal{T}$. Thus, the learner can decide to validate her hypothesized expression using the hold-out records S . (Then, if she is not satisfied with the reported statistics, she will continue to refine the expression.) However, after this point, despite being a blind call to S , we cannot pretend that the hold-out set is completely unseen and fresh. Because the process is iterative and adaptive, the learner can over-fit the hold-out set, which may impact generalization to the whole population. Because of the expense of annotating records, we cannot afford a fresh sample for every validation call. That is, the main challenge for us is to measure and control the effects of over-using the hold-out set while still giving informative guidance to develop an effective cohort discriminator.

Dwork et al. have described how to apply the principles of differential privacy to adaptive data analysis in order to leak as little information as possible from the sample S when answering statistical queries [18–20]. In essence, given a statistical query to be evaluated on the hold-out sample S , the evaluating oracle should return an *approximate* answer rather than an exact answer to the query, by adding some noise. Their analysis shows that, even when the learner is trying to maximize the information leaked from the hold-out data, the generalization error can then be bounded when no more than $o(n^2)$ queries to S , where $n = |S|$, return an answer that differs substantially from applying the same statistic to the training set $\mathcal{T}$. Unfortunately, in our setting, we have no way to evaluate any statistic on $\mathcal{T}$ computationally because none of the records in $\mathcal{T}$ have annotations. Therefore, we cannot evaluate how far the statistic on the training data differs from that statistic evaluated on the hold-out data, which is a key component of this approach. On the other hand, our learner is also not trying to game the system. Instead, we wish to protect her from unintentionally mis-using the released information and therefore being misled about the quality of her queries.

In follow-up work, Russo and Zou and then Xu and Raginsky analyzed the effect of information leakage from adaptive data analysis by applying information-theoretic reasoning, thus removing the dependence on differential privacy mechanisms [55, 74]. More recently, Rogers et al. have developed tighter bounds on the generalization error rate for several data-hiding mechanisms that add noise to answers to queries on S [53]. These tighter bounds allow guarantees to be made with even fewer hold-out records in S .

THEOREM 4.1 (ROGERS ET AL. [53] THEOREM 2.1). *Given a sample of size n drawn for a population $\mathcal{D}$, confidence parameter β , k statistical queries posed against the sample, and parameter ρ , adding Gaussian noise $N(0, \frac{1}{2n^2\rho})$ to each answer a_i to the k queries, guarantees that the answers to those queries posed against $\mathcal{D}$ fall within the interval $a_i \pm \tau$ with confidence $1 - \beta$, where*

$$\tau = \sqrt{\frac{1}{2n\beta} \min_{\lambda \in [0,1]} \left(\frac{2\rho kn - \ln(1-\lambda)}{\lambda} \right)} + \frac{1}{2n} \sqrt{\frac{1}{\rho} \ln \left(\frac{4k}{\beta} \right)}. \quad (1)$$

IQRef relies on these bounds, which depends on a procedure where Gaussian noise is added to *statistical queries* against the hold-out sample. Given a Boolean query q and record r annotated by C , we can evaluate the

³Some learning architectures divide S into a development test set S_D that can be re-used if the learning is found to be unsatisfactory in finding f and a *final* test set S_F , where S_F is truly a hold-out sample, i.e., used only once at the very end of learning.

predicate $q(r) \wedge C(r) \vee \neg q(r) \wedge \neg C(r)$ for each $r \in S$ and compute the mean over all records in S to estimate the accuracy of the query in identifying relevant records. This information is informative when an aggregated query is tested to see how well it characterizes the cohort of interest.

However, *accuracy* is not the goal when the learner is in the exploratory phase of developing that predicate. The goal of the exploratory phase is to establish a precise, but not necessarily complete, sub-query Q_i . Unfortunately, the *precision* of a request, the fraction of records returned by a retrieval request that are relevant, is not a statistical query, as is required by the theory: precision cannot be expressed as a mean of some function evaluated on individual data items. To address our need, and motivated by Juba’s investigation into machine learning for abductive reasoning [33], we observe that we can use *implication* as a proxy for precision. In particular, for query q and cohort annotation C , $\neg q(r) \vee C(r)$ is a statistical query whose mean value is high whenever either precision is high or the selectivity of q is low. As a result, if the learner tries a query that matches at least several records in the training set $\mathcal{T}$ and it returns a high value for implication, she can be fairly certain that the precision is also high.⁴

Thus, using implication as our statistical query ϕ_E in the exploratory phase and accuracy as our statistical query ϕ_A in the aggregating phase, we can rely on the proofs by Rogers et al. that the performance of our final query Q on the hold-out sample accurately predicts its performance on any representative subset of the population $\mathcal{D}$. Because this approach adds noise when reporting performance measures on S , we have the additional benefit that we are not reliant on the assessor making no mistakes when tagging S .

5 CASE STUDY WITH IQREF

In this section we present our experience with using IQRef to identify visits to Waterloo’s Optometry Clinic that can be classified as asymptomatic routine eye exams.

5.1 Data Preparation

In our EHRs, there are at least 10 free-text fields for every patient visit, among which only 4 might contain relevant data for selecting our cohort: CHIEF_COMPLAINT, REASON, PRIMARY, and REASON_FOR_VISIT. For this work, we obtained Optometry records for 1 year (2015), among which 9,046 exams contain one or more of the text fields of interest. For each of these visits, a file was created by concatenating all relevant fields present in the record; each such file serves as a proxy EHR for IQRef.

We then selected 459 of the records, using simple random selection, for tagging by the expert annotator.⁵ The annotation interface chooses records from the sample, in random order; shows the contents of the four text fields of interest for that record; and offers three options: “in the cohort,” “outside the cohort,” and “undecided.” The annotator may choose any of these options, using “undecided” when there is insufficient information in the text fields to make a relevance judgment. The interface allows the annotator to see all records that have already been annotated and to change annotations at any time.

For our case study, five of the records were labeled “undecided,” and the remaining 454 records were set aside to serve as our hold-out subset S , with 186 tagged as “in the cohort” and 268 as “outside the cohort.” This left 8,587 records for our training set $\mathcal{T}$. We indexed the records in $\mathcal{T}$ using *Lucene*,⁶ an open-source search engine for indexing and searching. For tokenization, we adopted *WhitespaceTokenizer* to separate the text at characters that are considered as white space according to *JAVA*. We also applied *LowerCaseFilter* to normalize token text to lowercase.

⁴This echoes the reasoning behind asking for high support (selectivity) as well as high confidence (precision) when mining for association rules between sets of items in market basket analyses [2].

⁵Unfortunately, when we asked our expert to annotate a subset of records in 2018, we had not yet determined the sample complexity required for our approach.

⁶<https://lucene.apache.org>.

Fig. 6. Evolution of Boolean expressions.

5.2 Familiarizing the Learner in Using the Interface

In preparation for the case study, we created a practice environment in which the learner could become familiar with the system. We selected 927 records from the WatES database (using simple random sampling) used for the previous study and already annotated as being part of the cohort or outside the cohort in the earlier study. Of these, we selected 96 records, again by simple random sampling, to be the practice hold-out sample S_p and the remaining 831 records to be the practice training set $\mathcal{T}_p$. The expert who would eventually serve as the learner for the case study could then use this environment as if the data were drawn from more recent EHRs, creating Boolean expressions using $\mathcal{T}_p$ and checking for implication or accuracy against S_p .⁷ By using this separate environment for practice, no information leakage from the actual EHRs in $\mathcal{T}$ or S was possible.

As well as using the practice set to learn how to use the IQRef system, the expert also used the opportunity to refresh herself on what sorts of vocabulary might appear in the EHRs (broke, broken, overdue, wants, forms, leaving, first eye exam, OVA) and how such terms might indicate whether an exam should be deemed to be in the cohort.

5.3 Observations on the Case Study

5.3.1 Observations on Annotation. The expert annotator spent 90 minutes to annotate 459 records and finished the task in one session. She reported that she used the text content of records exclusively to determine labels; i.e., no structural features such as a term's position or text length were used. She considered phrases up to three words long as *interesting*, which is why we decided to index records using unigrams, bigrams, and trigrams.

⁷Because the list of suggested synonyms for each term was drawn from the actual training set $\mathcal{T}$, not from $\mathcal{T}_p$, the learner might be led to include vocabulary that did not appear in $\mathcal{T}_p$; this, however, was not problematic.

5.3.2 Observations on IQRef Learning. The expert learner⁸ spent 16 hours and 40 minutes (after completing the practice sessions) spread out over a week to examine 626 Boolean expressions, either in the exploration or in the aggregation mode. This is about half the time it would have taken to annotate all 9,046 records from 2015, based on the time taken for annotating S , where the effort would provide no help whatsoever in identifying cohorts from other years.⁹ For each expression, she started with a small number of phrases to include (disjuncts) or exclude (negations) and gradually expanded to form more complex expression (Figure 6). Notably, the topmost box (for conjuncts that *must* all be present, Figure 3) was never used. After building a query that captured many alternatives for characterizing a set of potentially routine eye exams, she added more and more negated terms to rule out terms that indicated pre-existing conditions (glaucoma), symptoms (headaches), or other problems (lenses scratched) that indicate that the exam was triggered by some reason other than merely elapsed time. She explored the vocabulary by examining records that did and did not match the Boolean expressions formed, and only once she requested the system to show her some similar terms for some of the phrases in her query (adding 65 synonyms, including a few abbreviations and one misspelling, from the suggested lists). The constructed expressions, when they seemed to be promising, were used to build groups in the aggregating phase.

The expert queried the hold-out sample 5 times during exploration (implication queries) and 12 times during aggregation (accuracy queries) (Figure 7). The only data reported back to the expert were the implication values or accuracy values, always with Gaussian noise added. However, one can see from the figure that recall was significantly lower than precision throughout the experiment, ending at 0.75, whereas precision was 0.80 and accuracy was 0.82.

The final Boolean classifier Q is a group that specifies the union of six sub-queries, four of which correspond to queries that evolved during the exploration phase (and denoted Q_1 through Q_4 in Figure 6) and two of which were created without evolution and without evaluation. Among those sub-queries, there are 83 disjuncts and 417 negations, some phrases appearing multiple times and as disjuncts in some sub-queries but negations in others. The performance of Q against S and against WatES (excluding the 927 records used for practice) is summarized in Figure 9. Interestingly, the recall (sensitivity) was significantly higher on the WatES dataset than on the hold-out sample; we hypothesize that this is because the text in the WatES records was summarized by a single person from paper-based health records during a fairly short period of time, and thus that text is far less varying than what is found in actual EHRs, which are entered by numerous clinicians and students over a period of at least several months or years. For the IQRef learning case study we set the parameters in Theorem 4.1 to $\beta = 0.20$ and $\rho = 0.00016$. Based on that theorem, the accuracy of Q on the population (from which $\mathcal{T}$ is drawn) falls in the interval 0.83 ± 0.38 with 80% confidence.

In reviewing the experiment, the expert remarked that seeing lists of rejected EHRs did indeed help to improve recall by exposing terms that were missed. She realized during the practice session that the synonym function was not very user-friendly, especially as the number of terms in a query increased: many items were frequently repeated in several lists, making it quite inefficient to browse and to remember which had been already included in the expression and which were not. Negatives were perceived to be a problem for IQRef: for example, how to eliminate records that include “glaucoma” without eliminating ones stating “no signs of glaucoma.” Furthermore, a phrase that indicates the positive or negative use of a term was sometimes longer than what could be captured by trigrams. The strategy used to increase recall is to combine groups in the aggregation phase, but including any group with low precision could cause the overall precision to decrease. In retrospect, the expert hypothesized that it might have been better to try to find more specific negation terms and to form fewer groups. Clearly, we had

⁸As it turned out, in spite of our original intention, the same expert (the second author) served as the annotator and the learner, but with a 2-year gap between annotating the sample in 2018 and serving as the learner for IQREF in 2020. We are convinced that this separation in time obliterated any knowledge the expert had of the characteristics of specific records in the hold-out sample S . On the other hand, we are more assured that the annotator and the learner share a common understanding of the inclusion and exclusion criteria for the cohort.

⁹As a further point of reference, it is estimated that the original study, starting from paper records, took approximately 600 hours of expert time to identify the cohort.

Fig. 7. Statistics reported on S.

overestimated the utility of the synonym function and should instead have learned from the TREC experiments to normalize the data in a pre-processing phase.

5.3.3 Comparison to Active Learning. As part of our evaluation, we compared using IQRef to using conventional active learning. Following the design decisions made by others [10, 17, 31, 37, 40, 41], we used SVMs as our machine learning paradigm, with unigrams and negated terms (as determined by an adapted version of NegEx [10]) as features, a linear kernel, and uncertainty sampling. We used three possible stopping criteria: sufficient accuracy as determined by the expert using the same hold-out sample and feedback as for IQRef,¹⁰ an annotation budget of 459 items, and an annotation time not to exceed the time used for building the classifier using IQRef.

The expert annotator spent 1 hour and 36 minutes to complete 26 active learning iterations (with five records to annotate in each iteration) from which the first four sessions were to annotate the seeds for the initial classifier. She examined records from $\mathcal{T}$ that were classified as “in cohort” or “not in cohort” and requested 13 (noisy) accuracy estimates over the held-out sample. She voluntarily stopped active learning after 199 records, when she observed that the accuracy fluctuates and she could witness no more improvements in the lists of accepted and rejected records. As for IQRef, the recall was always significantly below precision, as shown in Figure 8. Statistics of the final classifier, f_n , on S and WatES can be seen in Figure 9, and the values are surprisingly close

¹⁰Because the SVM is not given any information about the hold-out sample and the expert cannot influence the behavior of the SVM other than by annotating additional records, the only leakage of information about that sample to the active learning run that can lead to overfitting is in deciding when to stop iterating. Thus, using the same hold-out sample for both parts of the experiment is not problematic.

Fig. 8. Active learning statistics reported on S .

to the statistics for the IQRef classifier. As for IQRef, we set $\beta = 0.20$ and $\rho = 0.00016$, to determine from Theorem 4.1 that the accuracy on the population falls in the interval 0.83 ± 0.36 with 80% confidence.

As can be seen in Figure 9, the classifiers created by IQRef and active learning performed surprisingly similarly on all measures. In fact, they agreed more than 80% of the time (Figure 10).

As a final step, we compared the quality of these two classifiers by looking at all records in $S \cup T$ in which they differed. Out of the 88 records in S on which the classifiers disagreed, the expert agreed with the classifier produced by IQRef 47.7% of the time. Without being told which classifier included each of those records, the expert was asked to annotate an additional sample $S_\delta \subset T$ consisting of 50 records of the 1,514 on which the classifiers disagreed. For S_δ , the expert agreed with the classifier produced by IQRef 40% of the time.

A qualitative examination of the records that the classifier produced by IQRef erroneously accepted shows that the sub-queries did not include sufficiently many negated terms: some variants were missed (refit was specified, but not re-fit), some terms were on the negated list of one sub-query but omitted from the negated list of another sub-query, and a few terms were erroneously not included on any of the negated lists. Similarly, records erroneously rejected by the classifier produced by IQRef confirmed that the expert had difficulty specifying the distinction between statements of problems (e.g., “complains of dry eyes”) and statements regarding the absence of problems (e.g., “had dry eyes, but pt doesn’t notice dryness anymore”).

When asked to comment on her experience, the expert observed that active learning seemed to handle spelling mistakes and synonyms well, but she was quite surprised to see (after the experiments concluded) that the overall performance metrics were as good as those for IQRef. She found two aspects of active learning very frustrating: helping to classify and knowing when to stop. For example, the expert could see from the lists that the classifier had difficulty dealing properly with mentions of diseases such as diabetes. Even when prompted to annotate some records that included the term diabetes, the classifier “didn’t seem to be interested in that part of the record.” Even with knowing the accuracy of the classifier after each iteration and being able to examine records

measure	formula	IQRef			active learning		
		wrt S	wrt $\mathcal{T}$	wrt WatES	wrt S	wrt $\mathcal{T}$	wrt WatES
fraction matched	$(TP+FP)/(TP+FP+TN+FN)$	0.38	0.34	0.52	0.37	0.32	0.57
precision	$TP/(TP+FP)$	0.81		0.77	0.82		0.70
recall $\equiv$ sensitivity	$TP/(TP+FN)$	0.75		0.96	0.75		0.97
specificity	$TN/(TN+FP)$	0.88		0.80	0.89		0.70
accuracy	$(TP+TN)/(TP+FP+TN+FN)$	0.82		0.86	0.83		0.81

Fig. 9. Statistics for final classifiers.

measure	wrt S	wrt $\mathcal{T}$
inter-system agreement	0.81	0.82
expert-IQRef agreement when systems disagree	0.48	0.40 [†]

[†] expert-IQRef agreement wrt $\mathcal{T}$ on a random sample of 50 out of 1514 records

Fig. 10. Agreement for final classifiers.

that matched and did not match the classifier, the stopping point seemed to her to be more arbitrary than for IQRef. In fact, she decided to stop when the accuracy fluctuated and she was afraid that continuing would only make things worse. This was in spite of the fact that she knew that the classifier was still making obvious mistakes.

In comparison to IQRef, the expert recognizes that active learning took much less time to get to similar statistical accuracy. Nevertheless, she feels less confident in the outcome. This is supported by accepted observations on human-in-the-loop systems that the human needs "to understand the agent's behavior and responses enough to participate in the mixed-initiative execution process and to adjust the autonomy inherent in such systems. The user would also need to trust the reasoning and actions performed by the agent" [23]. Our expert's *perception* at the end of the experiment was that IQRef had produced a classifier with higher precision than active learning did, and that the classifier produced by active learning would have higher recall. She was very surprised to find that both classifiers had very similar performance metrics. In retrospect, she reported: "Looking at these data I would say that overall IQRef does a better job or could be made to do a better job than active learning."

As a further experiment, we also investigated how active learning would have performed had we adopted the stopping condition derived by Cormack and Grossman [12] using a variant of "mark and recapture" [9]. The idea is to identify 10 relevant documents from the collection (the marked set) and then randomly select documents until those 10 documents are retrieved ("recaptured"); with 95% probability, this will recall at least 70% of the relevant records. We examine the cost of this approach using simulations based on our experimental setup and the results of the active learning approach described above.

We start by creating a model collection of records M that mirrors the relevance judgments for records in $\mathcal{T}$. For record $i \in M$, where $1 \leq i \leq |\mathcal{T}|$, let the key be i and assign it to being "in the cohort" with probability 186/454 and "out of the cohort" with probability 268/454, where these probabilities are taken from the assessment for records in $|S|$. Doing so in our experiment results in 3,412 of the 8,587 records being designated as "in the cohort," which is within 0.4% of the corresponding fraction in S .

Next, for each run of the simulation, randomly select records from M with uniform distribution and without replacement until 10 records designated as being "in the cohort" are retrieved, and denote the subset of records retrieved M_1 and the 10 "marked" records M_r . Next, again randomly select records from M with uniform distribution and without replacement, stopping when all 10 records in M_r have been retrieved, and denote this set of selected records as M_2 .

Based on 10 runs of the simulation, an expert would have to annotate on average $|M_1 \cup M_2| = 7,742$ records, with a standard deviation of 915. Estimating the time to annotate a record to be 29 seconds (the average time the expert actually spent per record during active learning), the expected time to complete this task would be 62.4 ± 2.3 hours.

However, in order to require at least 70% of the records “in cohort” to be retrieved, Cormack and Grossman’s stopping criteria require that most or all of those records will be annotated (as well as most of the other records) [9]. If we instead relax the criterion so that we stop when any 7 of the 10 records in M_r are retrieved in the recapture phase, we can expect a recall (i.e., sensitivity) close to the one the expert actually achieved with IQRef (Figure 9). Using the same simulation as above, but stopping when the first 7 marked records are re-captured, the expert would have to annotate on average 6,280 records, with a standard deviation of 733, and an average recall of 0.73 (as anticipated). This makes the expected time for the task to be 50.6 ± 1.9 hours, which is approximately three times as long as the total time used by the two experts for annotating S and learning the Boolean discriminator with IQRef to achieve a similar recall.

6 DISCUSSION

6.1 Reflections on IQRef

The design of IQRef was motivated by several factors:

- Because data labeling is an expensive process, the annotated sample needs to be fairly small, respecting an arbitrarily imposed budget.
- If we tune a classifier to perform well on this sample, we must be able to expect acceptable performance in isolating members of the cohort from whatever subset of the population we wish to study. That is, performance on the sample must (provably) be indicative of performance on the population as a whole.
- We must be able to decide when to stop training the classifier, based on its performance on the sample.
- Health and medical researchers will be more comfortable with automatic cohort identification if they can understand on what basis the classifier determines which EHRs belong to the cohort.
- Health and medical researchers can interpret predicates over words and short phrases, and expressing the predicates using Boolean operators only (i.e., without resorting to proximity or other operators) is sufficient to distinguish EHRs that belong to the cohort from those that do not.
- Developing classification predicates in two stages, one that tries to improve precision and another that tries to improve recall, is simpler than attempting to formulate an accurate classifier directly.
- Active learning environments can also use a hold-out sample with Gaussian noise to determine suitable halting conditions.

The resulting system showed promise in the case study. The classifier developed by the expert performed similarly to one produced by active learning, albeit with more effort required. That effort, however, resulted in a classifier that was understandable to the expert and explainable to other experts.

We have proposed that active learning environments can also use a hold-out sample with Gaussian noise to determine suitable halting conditions. However, when the expert decided that improvements to accuracy were no longer being made, she felt frustrated that the final classifier was still making many “obvious” mistakes. Even with the knowledge that it performed as well as IQRef, she does not trust its decision. “Interpretability in a medical setting is of utmost importance” [6], and we hypothesize that no explanation of the classifier produced by active learning will serve to increase trust in the eyes of the expert sufficiently well to convince her to rely on that classifier in the absence of the one produced by IQRef.

The expert has now concluded that the two systems make different kinds of mistakes, and she feels she can use this to her advantage. For carrying on with her optometric research (the reason for conducting this study in the first place), she expects to use both classifiers: where they agree, she will accept the decision, and where

they disagree, she will examine the records herself. If the statistics from this experiment generalize, as they should because of the way in which we incorporated information from the hold-out sample, she then expects to have reduced her classification work to examining about 20% of the data she needs for her study, resulting in a significant savings of time and effort.

Finally, it should be noted that the problem of selecting records for asymptomatic routine eye exams is a particularly difficult cohort identification problem: being in the cohort depends at least as much on what is *not* written in the EHR as what actually forms part of the record. We believe that using our approach to find cohorts having diabetes, glaucoma convergence insufficiency, blurred vision, or contact lens wear would have been more straightforward and more successful, being able to build on other work in disease phenotyping (such as the recent paper by De Freitas et al. [15]). Identifying cohorts for dry eye syndrome or asthenopia (eye strain) falls somewhere in between, because of the many possible descriptive terms to explore. Furthermore, as has been recognized elsewhere, ICD-9 codes provide valuable information to find cohorts associated with specific diseases, but they are not useful to capture characteristics of the population (because they are not used consistently enough) or those exhibiting lesser symptoms, such as “headaches” (because other diagnostics would most likely take priority and ICD-9 codes for the lesser symptoms would not be entered).

6.2 Representative Sampling vs. Simple Random Sampling

When we began to design our system, we were tempted to try to select records to be annotated by being careful to include some that were clearly within the cohort, clearly not in the cohort, and close to the “line” distinguishing cohort records from the rest. Our reasoning was that having such records in our hold-out sample would allow us to craft a more precise classifier. This led us to consider various approaches to defining a representative sample.

As explained in the introduction, cohort identification is often the prelude to studying certain statistical properties of particular subsets of the patient population. Therefore, our aim is to identify a subset of the population for which those statistics can be accurately obtained, which fits well into representative sampling. This does not necessarily require identifying members of the target class with high recall and high precision. What we need is a representative subpopulation on which the target statistics match those statistics on the population as a whole [65]. We call such a system an *analytics-aware system*.

IQRef obtains its hold-out subset by simple random sampling. There are several common alternative approaches to probability sampling [64]: systematic sampling, sampling with **probability proportional to size (PPS)**, stratified sampling, and cluster sampling. In other work, Zhang et al. investigate various sampling algorithms, such as stratified sampling, to estimate the number of relevant documents in a finite collection [76]. Podgurski et al. apply stratified sampling to estimate software reliability [51]. Their random sample is reviewed by experts to give an estimate of the occurrences of failures in the entire population, e.g., a collection of execution profiles from a *beta test*.

For cohort identification, we assume that the statistics of interest to the health researcher, and thus the statistics to be computed from the representative sample, are known *a priori*. Performance with respect to those specific metrics is then indicative of the representativeness of the sample with regard to a Boolean predicate. Because our predicates are sensitive to the absence or presence of terms in $\mathcal{E}$ only, the ideal for us is for the sample to preserve the representativeness of the sets of terms and any other attributes used for computing the statistics of interest to the health researcher.

In the context of market basket analyses, Brönnimann et al. [7] have shown how to achieve this goal (a specific form of ϵ -approximation [47]) by carefully selecting a sub-sample S_r from sample $S \subseteq \mathcal{D}$, where $|S_r| \approx |S|/2$. Letting $\mathcal{E}$ be the set of items that can appear in a basket, for every $e_i \in \mathcal{E}$, they define a set, $s_i \subseteq S$, containing all the baskets in S in which e_i appears. Brönnimann et al. have devised a penalty function to decide whether or not to keep a market basket in S_r so that the fraction of all one-item sets in S_r approximately equals the fraction in S . They further showed that using their penalty function for each s_i , they can halve S while maintaining a

min support	SRS	Sept	ϵ -approximation, from			
			D	D_{50}	D_{500}	D_{1000}
5%	0.30	0.25	0.40	0.40	0.45	0.47
7%	0.50	0.30	0.50	0.57	0.50	0.59
10%	0.60	0.30	0.60	0.66	0.57	0.60
15%	0.80	0.30	0.75	0.80	0.77	0.78

Fig. 11. Accuracy of preserving maximal frequent itemsets when $|S| \approx 450$.

bound on the discrepancy between S_r and S for all one-item sets:

$$|f(e_i, S_r) - f(e_i, S)| \leq \epsilon(n, g), \quad (2)$$

where $f(e_i, S)$ is the fraction of S that contains e_i , $\epsilon(n, g) = O(\sqrt{\frac{\log(2g)}{n}})$, g is the number of possible items in a basket, and $n = |S|$. Starting with any subset $D \subseteq \mathcal{D}$, their approach can be used repeatedly to halve sample sizes (or, in fact, used in a compounded manner) in order to obtain a “representative” sample $S_0 \subseteq D$ of approximately any desired size.

In our setting, each EHR represents a “basket of terms” in a g -dimensional space, where $g = |\mathcal{E}|$, the number of possible terms (words or short phrases). In order to guarantee the bound in Equation (2), however, g needs to be much less than n , the number of EHRs in D . If we represent the data with a binary matrix, with rows representing EHRs and columns representing terms (e_i), the number of columns is much higher than the number of rows. However, we can use dimensionality reduction techniques, such as singular value decomposition [24], to produce a narrower matrix representing almost the same information as in the original matrix.

We compared simple random sampling to the ϵ -approximation algorithm proposed by Brönnimann et al. when applied to the optometry EHRs. For our experiments, our original dataset D is such that $n = 7,268$ and $g = 280,888$, where $\mathcal{E}$ is all unigrams, bigrams, and trigrams found in the EHRs. We chose our sample set size $|S_0|$ to be $n/16 = 454$.¹¹ To better understand the effectiveness of the ϵ -approximation method, we applied it to D and to derived datasets D_{50} , D_{500} , and D_{1000} produced by applying singular value decomposition to D in order to reduce g to be 50, 500, and 1,000, respectively. As a further point of comparison, we also examined the results of using all the EHRs from a single “representative” month (September) that had approximately 450 visits.

Our measure of comparison is the accuracy of the sample in preserving maximal frequent itemsets, as used in the study by Brönnimann et al. to validate their approach. This is defined as

$$accuracy = 1 - \frac{|L(D) - L(S_0)| + |L(S_0) - L(D)|}{|L(S_0)| + |L(D)|}, \quad (3)$$

where $L(X)$ is the set of maximal frequent itemsets in X . By experiment, Brönnimann et al. showed that by guaranteeing a bound on the discrepancy in support of all one-item sets, it preserves the support for larger itemsets as well. The results of our experiment (Figure 11), however, show that a carefully chosen set of records using the ϵ -approximation algorithm, with or without dimensionality reduction, does not offer significantly better accuracy than choosing records by simple random sampling (but that choosing all the records from 1 month is a poor strategy).

Thus, because theoretical bounds on sampling error are available for simple random sampling only and because using ϵ -approximation instead to determine membership in the sample does not seem to yield significant advantage in practice, we recommend using simple random sampling to select which EHRs to be annotated for the hold-out sample.

¹¹In fact, this is historically the origin of the sample size used in our case study in Section 5.

6.3 Potential for Automatic Query Expansion

When the IQRef system was designed, it was thought that idiosyncratic terminology and abbreviations, as well as spelling errors, could be easily accommodated by the expert through the synonym panel. However, that function was under-utilized because the expert found it cumbersome to use. Clearly, the usability of the synonym panel could be significantly improved, but it would be even better if we could save the expert's time by not requiring such effort at all.

There are two approaches to address the problem of mismatched vocabulary: normalizing the documents (or the document index) or broadening the query via query expansion.

To follow the first approach, it would likely have been helpful to pre-process all text fields to remove obvious term variations and to normalize for negated terms [10, 41] *as was done for active learning*. This hypothesis is supported by the fact that recall on the more constrained WatES records was far higher than for the EHRs in the hold-out sample. Furthermore, a detailed examination of many of the mis-classified records reveals several instances where normalization might have avoided those errors. Although normalization would reduce the dependence on finding alternative terms, not all synonymous terms would be unified, and therefore some support for query expansion would still be required.

Rather than relying on the expert to find synonyms, expanded forms for abbreviations, and alternative spellings, IQRef could employ automatic query expansion [4]. Query reformulation techniques that rely on relevance or pseudo-relevance feedback require ranked retrieval to identify the closest matches to a query in order to identify the documents from which to draw additional terms; such an approach is not compatible with our set-based Boolean retrieval approach to cohort identification. As an alternative, Roy et al. [54] have suggested using word embeddings to expand the set of query terms automatically.

The case study is extremely costly to be repeated, but in an attempt to investigate the effects that automatic query expansion *might* have had on our results, we augment the query formulation conducted by the expert as if there were a possibility of automatic query expansion. In the remainder of this section, we describe an experiment in which, as each query is refined, additional terms are provided by automatic query expansion, and we hypothesize an oracle who can intercede to stop the expert from further work because the query has reached its best performance.

Query Description: As stated earlier, the final query group created by the expert learner included six Boolean sub-queries. Query logs revealed that four of the sub-queries were developed gradually (Figure 6) and the other two were not refined beyond their original creation. Note that both the lists of terms to include (disjuncts) and the lists of terms to exclude (negations) are subject to query expansion.

Expansion algorithm: We adopt the best-performing approach suggested by Roy et al. [54], i.e., *Pre-retrieval incremental kNN* matching with *term composition*. Our queries include phrases as well as individual terms, so we apply term composition to generate bigrams and trigrams to match each phrase's consecutive terms. For instance, for the phrase "Regular Eye Exam," the embedding vectors for "Regular Eye," "Eye Exam," and "Regular Eye Exam" are computed and added to the set of embeddings associated with unigrams to create a set of vectors denoted by V . To create the expansion candidates, for every member of $v \in V$, we start with the 300 most similar word embeddings generated from the word2vec model that we had trained for the synonym panel. We choose the most similar term t_1 , prune the 30 least similar terms, and then reorder the remaining terms by similarity to t_1 . We repeat this (choose, prune, reorder) eight times to end with nine candidate term vectors for each $v \in V$. Finally, we pick the terms corresponding to the 30 vectors with the highest mean cosine similarity to all the vectors in V .

Experiment: The goal of this experiment is to see whether we can find a point in the evolution of each query's inclusion or exclusion lists where applying automatic expansion would produce a query that performs at least as well as the final query created by the expert. Thus, each sub-query gives us two possible applications of automatic query expansion, either of which could be applied at any step during the sub-query's evolution.

Query	init. ORs	init. NOTs	MQE ORs	MQE NOTs	prec'n	time (mins)	stop ORs	stop NOTs	AQE ORs	AQE NOTs	prec'n	saved (mins)
Q_1	18	7	59	94	0.82	43	-	78	-	108	0.84	27
Q_2	2	0	-	18	0.59	7	-	-	-	-	-	0
Q_3	5	0	-	17	0.59	9	-	-	-	-	-	0
Q_4	6	0	-	177	0.71	130	-	138	-	-	0.71	22
Q_5	6	33	-	-	0.00	<2	-	-	36	63	0.33	0
Q_6	6	79	-	-	1.00	<2	-	-	-	-	-	0

Fig. 12. Manual query expansion vs. AQE with Oracle.

We repeat the behavior of the user during the exploration phase (building high precision queries), but also test the effect of automatic query expansion on either or both lists. Thus, for every step during a sub-query's evolution depicted in Figure 6, we examine four possible scenarios:

- exclusion and inclusion lists do not get expanded automatically at that point,
- exclusion list is expanded but inclusion list is not expanded automatically at that point,
- inclusion list is expanded but exclusion list is not expanded automatically at that point, and
- exclusion and inclusion lists both get expanded automatically at that point.

Then, for each sub-query we stop tracing the steps actually taken by the expert if any of the four scenarios outperforms the final precision actually achieved by the expert.

The results in Figure 12 show that an omniscient learner could have ended the manual exploration of queries early in 2 of the 12 opportunities for query expansion and that automatic query expansion would have yielded some benefit in only 3 of the 12 opportunities. Of course, not knowing the precision that she would ultimately achieve and without evaluating each query (and thus significantly deteriorating the value of the hold-out sample), the expert learner could not possibly know when to choose to stop her manual effort and whether or not to apply automatic query expansion at that point. Furthermore, had automatic query expansion been available to our expert learner, she might well have pursued some other query alternatives entirely, perhaps improving her final result and perhaps saving some time. Other case studies will be required to determine how or whether to apply automatic query expansion in practice.

Recall that our naive approach merely presented the 10 nearest terms to each query term, which became particularly cumbersome and redundant as disjunction or negation lists grew. In retrospect, we probably should have used pre-retrieval incremental kNN matching with term composition to replace the synonym table with a much improved list of suggested terms.

6.4 Potential for Deep Learning

For many applications of machine learning, there has been significant success in using deep learning [75] and, more recently, in approaches based on BERT [42]. We are unaware of attempts to use deep learning for cohort identification, but because of successes for other applications using information retrieval, there is some hope for successful results for this application as well.

A major drawback in applying deep learning to our problem is the lack of training data. Lin et al. [42], for example, report that the training set sizes for the reranking task range from over 9,000 annotated records (for Robust04) to over 350K records (for TREC2019) and over 500K records (for MSMarco). The availability of only 454 annotated records for our corpus and application is exceedingly small by comparison.

We are also concerned about the black-box nature of deep learning and how domain experts in optometry will come to trust the results of cohort identification based only on BERT. “If the explainability of the AI application is poor or missing, trust is affected” [63], We therefore leave it to other researchers to explore this direction.

7 CONCLUSIONS

Cohort identification requires considerable effort on the part of experts but can be assisted by expert-in-the-loop learning. IQRef complements active learning in providing a mechanism to benefit from human expertise, both producing high accuracy. The advantage of IQRef lies in the understandability of the classifier; the advantage of active learning is that it demands less time from the expert, but it leaves the expert feeling much less empowered in helping to avoid “obvious” mis-classifications.

Both approaches are well served by access to a hold-out sample to help the expert to decide when to stop the learning process. By adding some noise to the answers provided by the hold-out sample, suitable guidance can be provided to the expert in determining how well the performance on the sample generalizes to the population more generally.

We are currently planning to use this approach to identify glaucoma patients for inclusion in a research study. As mentioned earlier, we expect this to be a simpler problem, as there are fewer possible terms, and we need to identify only sufficiently many patients with glaucoma to do the study. Given the system’s success with the challenging problem of identifying patient visits that can be classified as regular eye exams, we are optimistic of success.

We hope that other medical researchers can similarly benefit from this approach. For example, perhaps the adoption of IQRef could benefit someone searching for potential COVID-19 deaths that are not attributed to the coronavirus because of either the lack of testing or false negatives.

Most experts will likely be willing to invest time in using both methods, and the complementary classifiers produced by the two methods can help them to decide how best to choose the final cohort that is required for carrying out their health analyses. For example, it might be appropriate to use only those records that meet the conditions imposed by both classifiers. Alternatively, if a more complete cohort is required, every record on which the classifiers disagree can be reviewed by the expert to determine whether or not to add it to the set of EHRs on which they both agree. More generally, fusing results from several classifiers, including those that use “fill-in-the-blank” fields, images, and other forms of data, can prove to be extremely useful [29, 48].

Using implication as a proxy for precision solved the technical problem of requiring that statistical queries be posed against the hold-out data. However, it seems that toward the end of developing the classifier, the expert was misled into thinking that a sub-query had high precision based on seeing high implication (Figures 7 and 12), when instead the reason was low selectivity. Being able to report such information back to the expert, without compromising the hold-out data, would help her to formulate more accurate classifiers. We recommend that as well as reporting implication during the exploratory phase, the system should also report selectivity (which is a statistical query) so that the expert can avoid unknowingly including a sub-query with low precision. The cost of this approach is to double the number of queries to the hold-out data during the exploratory phase,¹² thus requiring a small increase in noise to avoid accidental over-fitting.

In summary, these are the major lessons learned for cohort selection from EHRs:

- Effort should be devoted in advance to normalize the language used in free text, including a representation for negated language.
- Expert selection of terms for inclusion and exclusion criteria is invaluable but can likely be improved (and time can likely be saved) by suggesting terms using automatic query expansion techniques.

¹²Had we also reported an estimate of selectivity in our experiment, we would have required five more queries to the hold-out data.

- As long as the cohort constitutes a reasonable fraction of the data (at least 10% perhaps), a simple random sample of records should be annotated in advance and used as a hold-out set with added noise to help the expert evaluate progress and determine when to stop without overfitting.
- Using IQRef in conjunction with active learning will produce results that will be acceptable by medical and healthcare professionals.

Experience has now confirmed our initial impression that achieving high recall might be difficult with IQRef. One possible improvement would be to allow the expert learner to browse or search through a list of terms (unigrams, bigrams, and trigrams) that appear in $\mathcal{T}$, ordered by frequency and perhaps linked by semantic similarity. Being informed of how many times various terms appear might help the expert to find new avenues to explore to capture additional EHRs that should be included in the cohort.

Finally, it would be interesting to examine how well any set of rules extracted from the classifier produced by active learning compare to the Boolean formula produced by IQRef. In particular, experiments supporting the rule-extraction method proposed by Fung et al. [21] and similar approaches to SVM model explanation have been conducted on datasets with hundreds of records in feature spaces with 9 to 70 dimensions. How well do these methods apply to our application, which has thousands of records in a feature space with hundreds of thousands of dimensions?

REFERENCES

- [1] Leonard Adelman. 1991. Experiments, quasi-experiments, and case studies: A review of empirical methods for evaluating decision support systems. *IEEE Transactions on Systems, Man, and Cybernetics* 21, 2 (1991), 293–301. <https://doi.org/10.1109/21.87078>
- [2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD'93)*. Association for Computing Machinery, New York, NY, 207–216. <https://doi.org/10.1145/170035.170072>
- [3] Michela Assale, Linda Greta Dui, Andrea Cina, Andrea Seveso, and Federico Cabitza. 2019. The revival of the notes field: Leveraging the unstructured content in eHealth records. *Frontiers in Medicine* 6 (2019), 66. <https://doi.org/10.3389/fmed.2019.00066>
- [4] Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: A survey. *Information Processing and Management* 56, 5 (2019), 1698–1735. <https://doi.org/10.1016/j.ipm.2019.05.009>
- [5] Steven Bedrick, Tracy Edinger, Aaron M. Cohen, and William R. Hersh. 2012. Identifying patients for clinical studies from electronic health records: TREC 2012 Medical Records Track at OHSU. In *Proceedings of the 21st Text REtrieval Conference (TREC'12) (NIST Special Publication)*, Ellen M. Voorhees and Lori P. Buckland (Eds.), Vol. 500-298. National Institute of Standards and Technology (NIST), 18 pages. <http://trec.nist.gov/pubs/trec21/papers/OHSU.medical.final.pdf>.
- [6] Alireza Borjali, Antonia F. Chen, Orhun K. Muratoglu, Mohammad A. Morid, and Kartik M. Varadarajan. 2020. Deep learning in orthopedics: How do we build trust in the machine? *Healthcare Transformation* (2020) 6 pages. <http://doi.org/10.1089/heat.2019.0006>
- [7] Hervé Brönnimann, Bin Chen, Manoranjan Dash, Peter J. Haas, and Peter Scheuermann. 2003. Efficient data reduction with EASE. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 59–68.
- [8] Steven R. Chamberlin, Steven D. Bedrick, Aaron M. Cohen, Yanshan Wang, Andrew We, Sijia Liu, Hongfang Liu, and William R. Hersh. 2020. Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. *JAIMA Open* 3, 3 (2020), 395–404. <https://doi.org/10.1093/jamiaopen/ooaa026>
- [9] Douglas G. Chapman. 1954. The estimation of biological populations. *Annals of Mathematical Statistics* 25, 1 (1954), 1–15. <https://doi.org/10.1214/aoms/1177728844>
- [10] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics* 34, 5 (2001), 301–310.
- [11] Yukun Chen, Robert J. Carroll, Eugenia R. McPeck Hinz, Anushi Shah, Anne E. Eyler, Joshua C. Denny, and Hua Xu. 2013. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association* 20, e2 (2013), e253–e259.
- [12] Gordon V. Cormack and Maura R. Grossman. 2016. Engineering quality and reliability in technology-assisted review. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 75–84. <https://doi.org/10.1145/2911451.2911510>
- [13] Licong Cui, Samden D. Lhatoo, Guo-Qiang Zhang, Satya Sanket Sahoo, and Alireza Bozorgi. 2012. EpiDEA: Extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. In *American Medical Informatics Association Annual Symposium (AIMA'12)*. AMLA. <http://knowledge.amia.org/amia-55142-a2012a-1.636547/t-003-1.640625/f-001-1.640626/a-134-1.640826/a-135-1.640823>.

- [14] Kim N. Danforth, Megan I. Early, Sharon Ngan, Anne E. Kosco, Chengyi Zheng, and Michael K. Gould. 2012. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. *Journal of Thoracic Oncology* 7, 8 (Aug. 2012), 1257–1262.
- [15] Jessica K. De Freitas, Kipp W. Johnson, Eddy Golden, Girish N. Nadkarni, Joel T. Dudley, Erwin P. Bottinger, Benjamin S. Glicksberg, and Riccardo Miotto. 2021. Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. *Patterns* 2, 100337 (September 10, 2021), <https://doi.org/10.1016/j.patter.2021.100337>
- [16] Dina Demner-Fushman, Swapna Abhyankar, Antonio Jimeno-Yepes, Russell F. Loane, François-Michel Lang, James G. Mork, Nicolas Ide, and Alan R. Aronson. 2012. NLM at TREC 2012 Medical Records Track. In *Proceedings of the 21st Text REtrieval Conference (TREC'12) (NIST Special Publication)*, Ellen M. Voorhees and Lori P. Buckland (Eds.), Vol. 500–298. National Institute of Standards and Technology (NIST), 5 pages. <http://trec.nist.gov/pubs/trec21/papers/NLM.medical.final.pdf>.
- [17] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management*. 148–155.
- [18] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.), 2350–2358. <http://papers.nips.cc/paper/5993-generalization-in-adaptive-data-analysis-and-holdout-reuse>.
- [19] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2017. Guilt-free data reuse. *Communications of the ACM* 60, 4 (2017), 86–93. <https://doi.org/10.1145/3051088>
- [20] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. 2015. Preserving statistical validity in adaptive data analysis. In *Proceedings of the 47th Annual ACM on Symposium on Theory of Computing (STOC'15)*, Rocco A. Servedio and Ronitt Rubinfeld (Eds.). ACM, 117–126. <https://doi.org/10.1145/2746539.2746580>
- [21] Glenn Fung, Sathyakama Sandilya, and R. Bharat Rao. 2005. Rule extraction from linear support vector machines. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD'05)*. 32–40.
- [22] Dominic Girardi, Josef Kueng, and Andreas Holzinger. 2015. A domain-expert centered process model for knowledge discovery in medical research: Putting the expert-in-the-loop. In *International Conference on Brain Informatics and Health*. Springer, 389–398.
- [23] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI'08)*. Association for Computing Machinery, New York, NY, 227–236. <https://doi.org/10.1145/1378773.1378804>
- [24] Gene H. Golub and Christian Reinsch. 1971. Singular value decomposition and least squares solutions. In *Handbook for Automatic Computation*. Springer, 134–151.
- [25] Maura R. Grossman, Gordon V. Cormack, and Adam Roegiest. 2016. TREC 2016 total recall track overview. In *Proceedings of the 25th Text REtrieval Conference (TREC'16) (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. 500–321. National Institute of Standards and Technology (NIST), 17 pages. <http://trec.nist.gov/pubs/trec25/papers/Overview-TR.pdf>.
- [26] Xuan Guo, Qi Yu, Rui Li, Cecilia Ovesdotter Alm, Cara Calvelli, Pengcheng Shi, and Anne Haake. 2016. An expert-in-the-loop paradigm for learning medical image grouping. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 477–488.
- [27] Joseph M. Hellerstein, Ron Avnur, Andy Chou, Christian Hidber, Chris Olston, Vijayshankar Raman, Tali Roth, and Peter J. Haas. 1999. Interactive data analysis: The control project. *Computer* 32, 8 (1999), 51–59.
- [28] Andreas Holzinger. 2016. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.
- [29] Andreas Holzinger, Bernd Malle, Anna Saranti, and Bastian Pfeifer. 2021. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Information Fusion* 71 (2021), 28–37.
- [30] Elizabeth L. Irving, Joel D. Harris, Carolyn M. Machan, Barbara E. Robinson, Patricia K. Hrynchak, Susan J. Leat, and Linda Lillakas. 2016. Value of routine eye examinations in asymptomatic patients. *Optometry and Vision Science: Official Publication of the American Academy of Optometry* 93, 7 (July 2016), 660–666.
- [31] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*. Springer, 137–142.
- [32] John R. Josephson and Susan G. Josephson. 1996. *Abductive Inference: Computation, Philosophy, Technology*. Cambridge University Press.
- [33] Brendan Juba. 2016. Learning abductive reasoning using random examples. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Dale Schuurmans and Michael P. Wellman (Eds.). AAAI Press, 999–1007. <http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12186>.
- [34] Michael Kearns. 1998. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)* 45, 6 (1998), 983–1006.
- [35] Vishesh Kumar, Katherine Liao, Su-Chun Cheng, Sheng Yu, Uri Kartoun, Ari Brettman, Vivian Gainer, Andrew Cagan, Shawn Murphy, Guergana Savova, Pei Chen, Peter Szolovits, Zongqi Xia, Elizabeth Karlson, Robert Plenge, Ashwin Ananthakrishnan, Susanne Churchill, Tianxi Cai, Isaac Kohane, and Stanley Shaw. 2014. Natural language processing improves phenotypic accuracy in an

- electronic medical record cohort of type 2 diabetes and cardiovascular disease. *Journal of the American College of Cardiology* 63, 12 Supplement (2014), A1359.
- [36] Leonard T Kurland and Craig A Molgaard. 1981. The patient record in epidemiology. *Scientific American* 245, 4 (1981), 54–63.
 - [37] Edda Leopold and Jörg Kindermann. 2002. Text categorization with support vector machines. How to represent texts in input space? *Machine Learning* 46, 1–3 (2002), 423–444.
 - [38] David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (Special Issue of the SIGIR Forum)*. 3–12.
 - [39] Katherine P. Liao, Tianxi Cai, Guergana K. Savova, Shawn N. Murphy, Elizabeth W. Karlson, Ashwin N. Ananthakrishnan, Vivian S. Gainer, Stanley Y. Shaw, Zongqi Xia, Peter Szolovits, Susanne Churchill, and Isaac Kohane. 2015. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* 350:h1885 (2015), 6 pages. <https://doi.org/10.1136/bmj.h1885>
 - [40] Nut Limsopatham, Craig Macdonald, Richard McCreadie, and Iadh Ounis. 2012. Exploiting term dependence while handling negation in medical search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1065–1066.
 - [41] Nut Limsopatham, Craig Macdonald, and Iadh Ounis. 2013. Learning to handle negated language in medical records search. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. 1431–1440.
 - [42] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. *Pretrained Transformers for Text Ranking: BERT and Beyond*. version 3, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, forthcoming. 204 pages. <https://arxiv.org/abs/2010.06467>
 - [43] Sijia Liu, Yanshan Wang, Andrew Wen, Liwei Wang, Na Hong, Feichen Shen, Steven Bedrick, William Hersh, and Hongfang Liu. 2019. Create: Cohort retrieval enhanced by analysis of text from electronic health records using OMOP common data model. (2019). arXiv:1901.07601
 - [44] Carolyn M. Machan, Patricia K. Hrynychak, and Elizabeth L. Irving. 2011. Waterloo Eye Study: Data abstraction and population representation. *Optometry and Vision Science* 88, 5 (2011), 613–620.
 - [45] Carolyn M. Machan, Patricia K. Hrynychak, and Elizabeth L. Irving. 2012. Modeling the prevalence of age-related cataract: Waterloo eye study. *Optometry and Vision Science: Official Publication of the American Academy of Optometry* 89, 2 (Feb. 2012), 130–136.
 - [46] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Proceedings*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 3111–3119. <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality>.
 - [47] Nabil H. Mustafa and Kasturi R. Varadarajan. 2017. Epsilon-approximations and epsilon-nets.
 - [48] Henry Friday Nweke, Ying Wah Teh, Ghulam Mujtaba, and Mohammed Ali Al-garadi. 2019. Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion* 46 (2019), 147–170.
 - [49] Serguei Pakhomov, Susan A. Weston, Steven J. Jacobsen, Christopher G. Chute, Ryan Meverden, and Véronique L. Roger. 2007. Electronic medical records for clinical research: Application to the identification of heart failure. *American Journal of Managed Care* 13, 6 Part 1 (2007), 281–288.
 - [50] Chiara Picardi, Richard Hawkins, Colin Paterson, and Ibrahim Habli. 2019. A pattern for arguing the assurance of machine learning in medical diagnosis systems. In *Computer Safety, Reliability, and Security - 38th International Conference (SAFECOMP'19), Proceedings (Lecture Notes in Computer Science)*, Alexander B. Romanovsky, Elena Troubitsyna, and Friedemann Bitsch (Eds.), Vol. 11698. Springer, 165–179. https://doi.org/10.1007/978-3-030-26601-1_12
 - [51] Andy Podgurski, Wassim Masri, Yolanda McCleese, Francis G. Wolff, and Charles Yang. 1999. Estimation of software reliability by stratified sampling. *ACM Transactions on Software Engineering and Methodology* 8, 3 (1999), 263–283.
 - [52] Adam Roegiest, Gordon V. Cormack, Charles L. A. Clarke, and Maura R. Grossman. 2015. TREC 2015 total recall track overview. In *Proceedings of the 24th Text REtrieval Conference (TREC'15) (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. 500–319. National Institute of Standards and Technology (NIST), 29 pages. <https://trec.nist.gov/pubs/trec24/papers/Overview-TR.pdf>.
 - [53] Ryan Rogers, Aaron Roth, Adam D. Smith, Nathan Srebro, Om Thakkar, and Blake E. Woodworth. 2020. Guaranteed validity for empirical approaches to adaptive data analysis. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS'20), Online (Proceedings of Machine Learning Research)*, Silvia Chiappa and Roberto Calandra (Eds.), Vol. 108. PMLR, 2830–2840. <http://proceedings.mlr.press/v108/rogers20a.html>.
 - [54] Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. 2016. Using word embeddings for automatic query expansion. In *The SIGIR 2016 Workshop on Neural Information Retrieval (Neu-IR'16)*. 5 pp. arXiv:1606.07608
 - [55] Daniel Russo and James Zou. 2020. How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory* 66, 1 (2020), 302–323. <https://doi.org/10.1109/TIT.2019.2945779>
 - [56] Yvonne Sada, Jason Hou, Peter Richardson, Hashem El-Serag, and Jessica Davila. 2016. Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. *Medical Care* 54, 2 (February 2016), e9–e14. <https://doi.org/10.1097/MLR.0b013e3182a30373>

- [57] Raymond Francis Sarmiento and Franck Dernoncourt. 2016. Improving patient cohort identification using natural language processing. In *Secondary Analysis of Electronic Health Records*. Springer, Chapter 28. https://doi.org/10.1007/978-3-319-43742-2_28
- [58] Burr Settles. 2012. *Active learning*. In *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers.
- [59] Burr Settles, Mark Craven, and Soumya Ray. 2007. Multiple-instance active learning. In *Advances in Neural Information Processing Systems 20: Proceedings of the 21st Annual Conference on Neural Information Processing Systems*. 1289–1296.
- [60] H. Sebastian Seung, Manfred Oppel, and Haim Sompolsky. 1992. Query by committee. In *Proceedings of the 5th Annual ACM Conference on Computational Learning Theory (COLT'92)*. 287–294.
- [61] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
- [62] Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J. Embi, Noemie Elhadad, Stephen B. Johnson, and Albert M. Lai. 2013. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association* 21, 2 (2013), 221–230. <https://doi.org/10.1136/amiajnl-2013-001935>
- [63] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.
- [64] Statistics Canada. 2001. Statistics: Power From Data! (2001). Retrieved April 30, 2020, from <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/toc-tdm/5214718-eng.htm>. Catalog no. 12-004-X.
- [65] Yves Tillé. 2006. *Sampling Algorithms*. Springer Science & Business Media.
- [66] Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research* 2 (Nov. 2001), 45–66.
- [67] Debbie A. Travers and Stephanie W. Haas. 2004. Evaluation of emergency medical text processor, a system for cleaning chief complaint text data. *Academic Emergency Medicine* 11, 11 (2004), 1170–1176. <https://doi.org/10.1197/j.aem.2004.08.012>
- [68] Ellen M. Voorhees. 2013. The TREC Medical Records Track. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (BCB'13)*. Association for Computing Machinery, New York, NY, 239–246. <https://doi.org/10.1145/2506583.2506624>
- [69] Ellen M. Voorhees. 2019. The evolution of Cranfield. In *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, Nicola Ferro and Carol Peters (Eds.). The Information Retrieval Series, Vol. 41. Springer, 45–69. https://doi.org/10.1007/978-3-030-22948-1_2
- [70] Ellen M. Voorhees and Lori P. Buckland (Eds.). 2011. *Proceedings of the 20th Text REtrieval Conference (TREC'11)*. Vol. Special Publication 500-296. National Institute of Standards and Technology (NIST). <https://trec.nist.gov/pubs/trec20/t20.proceedings.html>
- [71] Ellen M. Voorhees and Lori P. Buckland (Eds.). 2012. *Proceedings of the 21st Text REtrieval Conference (TREC'12)*. NIST Special Publication, Vol. 500-298. National Institute of Standards and Technology (NIST). <http://trec.nist.gov/pubs/trec21/t21.proceedings.html>
- [72] Alexander M. Walker, Xiaofeng Zhou, Ashwin N. Ananthakrishnan, Lisa S. Weiss, Rongjun Shen, Rachel E. Sobel, Andrew Bate, and Robert F. Reynolds. 2016. Computer-assisted expert case definition in electronic health records. *International Journal of Medical Informatics* 86 (2016), 62–70. <https://doi.org/10.1016/j.ijmedinf.2015.10.005>
- [73] William Webber. 2013. Approximate recall confidence intervals. *ACM Transactions on Information Systems* 31, 1 (2013), 2:1–2:33. <https://doi.org/10.1145/2414782.2414784>
- [74] Aolin Xu and Maxim Raginsky. 2017. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 2524–2533. <http://papers.nips.cc/paper/6846-information-theoretic-analysis-of-generalization-capability-of-learning-algorithms>.
- [75] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J Smola. 2021. Dive into Deep Learning, v2. 839. <https://arxiv.org/abs/2106.11342>.
- [76] Haotian Zhang, Jimmy J. Lin, Gordon V. Cormack, and Mark D. Smucker. 2016. Sampling strategies and active learning for volume estimation. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 981–984. <https://doi.org/10.1145/2911451.2914685>

Received February 2021; revised June 2021; accepted August 2021