Active Learning for Phenotyping Tasks

Dmitriy Dligach **Timothy A. Miller** **Guergana K. Savova**
Boston Children's Hospital and Harvard Medical School
firstname.lastname@childrens.harvard.edu

Abstract

Active learning is a popular research area in machine learning and general domain natural language processing (NLP) communities. However, its applications to the clinical domain have been studied very little and no work has been done on using active learning for phenotyping tasks. In this paper we experiment with a specific kind of active learning known as uncertainty sampling in the context of four phenotyping tasks. We demonstrate that it can lead to drastic reductions in the amount of manual labeling when compared to its passive counterpart.

1 Introduction

Several multi-year, multi-institutional translational science initiatives focus on combining large repositories of biological specimens and Electronic Health Records (EHR) data for high-throughput genetic research with the ultimate goal of transferring the knowledge to the point of care. Among them are Electronic Medical Records and Genomics (eMERGE)(McCarty et al., 2011), Pharmacogenomics Network (PGRN)(Long and Berg, 2011), Informatics for Integrating Biology and the Bedside (i2b2)(Kohane et al., 2012). In each of these initiatives there are a number of diseases or driving biology projects that are studied such as Rheumatoid Arthritis, Multiple Sclerosis, Inflammatory Bowel Disease, Autism Spectrum Disorder, Early Childhood Obesity, each defined as the phenotype of interest. To enable large cohort identification, phenotype-specific algorithms are developed, evaluated and run against multi-million-patient EHRs which are then matched against the biobanks for further genetic analysis. Efficient and accurate large-scale automated phenotyping is a key component of these efforts.

Supervised machine learning is widely used for phenotype cohort identification (Ananthakrishnan et al., 2012; Ananthakrishnan et al., 2013b; Ananthakrishnan et al., 2013a; Lin et al., 2012b; Lin et al., 2012a; Xia et al., 2012). However, the supervised learning approach is expensive due to the costs associated with gold standard creation. While large amounts of unlabeled data are available to the researchers in the form of EHRs, a significant manual effort is required to label them. In a typical phenotype creation project (Lin et al., 2012b; Lin et al., 2012a), a pool of patients is identified using some filtering criteria and a subset of patients is selected from that pool for subsequent expert annotation. During annotation, a domain expert examines the notes associated with a patient, assigning either the positive label (i.e. relevant for the given phenotype) or the negative one. A model is subsequently trained using the annotated data. This scenario is known as *passive learning*.

On the other hand, active learning (Settles, 2009; Olsson, 2009) is an efficient alternative to the traditionally used passive learning as it has the potential to reduce the amount of annotation that is required for training highly accurate machine learning models. Multiple studies have demonstrated that when active learning is used, machine learning models require significantly less training data and can still perform without any loss of accuracy. Active Learning is a popular research area in machine learning and general domain natural language processing (NLP) communities. However, its applications to the clinical domain have been little studied and no work has been done on using active learning for phenotyping tasks. In this paper we experiment with a specific kind of active learning known as *uncertainty sampling* in the context of four phenotyping tasks. We demonstrate that active learning can lead to drastic reductions in the amount of manual labeling without any loss of ac-

Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013, pages 1–8,
Hissar, Bulgaria, 13 September 2013.

curacy when compared to passive learning.

2 Background

2.1 Phenotyping as Document Classification

Phenotyping can be viewed as a document classification task in which a document consists of all EHR documents and other associated data (labs, ordered medications, etc.) for the given patient. Initial filtering is usually performed based on a set of inclusion and exclusion criteria (ICD-9 codes, CPT codes, laboratory results, medication orders). Within the eMERGE and PGRN, flowcharts outlining each phenotyping criterion and logical operators (AND, OR) are defined to constitute the phenotyping algorithm (Pacheco et al., 2009; Waudby et al., 2011; Kho et al., 2011; Kullo et al., 2010; Kho et al., 2012; Denny et al., 2010). The phenotyping within i2b2 takes a different approach – that of a machine learning patient-level classification task (Ananthakrishnan et al., 2012; Ananthakrishnan et al., 2013b; Ananthakrishnan et al., 2013a; Lin et al., 2012b; Lin et al., 2012a; Xia et al., 2012). Each patient is represented as a set of variables derived from the structured and unstructured part of the EHR (ICD-9 codes, lab results, relevant mentions in the clinical narrative along with their attributes) which are then passed to a machine learning algorithm. Whether the choice is a rule-based or machine learning approach, a fairly big sample of data needs to be labeled by experts which will then be used to derive the rules/train a classifier and to evaluate the performance.

2.2 Active Learning

Active learning is an approach to selecting unlabeled data for annotation that can potentially lead to large reductions in the amount of manual labeling that is necessary for training an accurate classifier. Unlike passive learning, where the data is sampled for annotation randomly, active learning delegates the data selection to the classifier. Active learning succeeds if it reaches the same performance as its passive counterpart but with fewer training examples.

Seung et. al. (Seung et al., 1992) present an active learning algorithm known as query by committee. In this algorithm, two classifiers are derived from the labeled data and used to label new examples. The instances where the two classifiers disagree are returned to a human annotator for labeling. Lewis and Gale (Lewis and Gale, 1994) pioneered the use of active learning for text categorization. Their scenario, known as pool-based active learning, corresponds to a setting where an abundant supply of text documents is available but only a small sample can be economically annotated by a human labeler. Pool-based active learning has since been explored for many problem domains such as text classification (McCallum and Nigam, 1998; Tong and Chang, 2001; Tong and Koller, 2002), word-sense disambiguation (Chen et al., 2006; Zhu and Hovy, 2007; Dligach and Palmer, 2011), information extraction (Thompson et al., 1999; Settles et al., 2008), and image classification (Tong and Chang, 2001; Hoi et al., 2006).

The pool-based scenario matches the setting in our phenotyping tasks where large supplies of unlabeled EHRs are available but only a small set can be manually reviewed at a reasonable cost. Pool-based active learning is typically an iterative process that operates by first training a classifier on a small sample of the data known as the seed set. The classifier is subsequently applied to a pool of unlabeled data with the purpose of selecting additional examples the classifier views as informative. The selected data is annotated and the cycle is repeated, allowing the learner to quickly refine the decision boundary between classes.

Little research exists on the applications of active learning to the clinical domain. Figueroa et al. (Figueroa et al., 2012) evaluate a Support Vector Machine (SVM) based active learning algorithm in the context of several text classification tasks and find that active learning did not always perform better than random sampling. The use of SVMs restricted their evaluation to binary classification only, limiting the applicability of their findings for many clinical NLP tasks. Chen et al. (Chen et al., 2011) investigate the use of active learning for assertion classification and show that active learning outperforms random sampling. Both of the above mentioned studies experiment with datasets that are quite different from ours in that they annotate relatively short snippets of text. Miller et al. (Miller et al., 2012) develop a series of active learning methods that are highly tailored to coreference resolution in clinical texts. Finally, Hahn et al. (Hahn et al., 2012) utilize active learning in practice for a corpus annotation task that involves labeling pathological phenomena in MEDLINE abstracts. Unfortunately they do not compare the performance of their active learning method to a passive learning baseline, so no conclusion about the effectiveness of active learning can be made. To the best of our knowledge, no work has been done on using active learning for phenotyping. In this work, we experiment with multi-class pool-based active learning in the context of four phenotyping tasks.

3 Methods

3.1 Data Representation

In a phenotyping task, the unit of classification is the patient chart. We represent each chart as a set of Unified Medical Language System (UMLS) (Bodenreider and McCray, 2003) concept identifiers (CUIs) which we extract from the patient records using Apache Clinical Text Analysis and Knowledge Extraction System[1] (cTAKES) (Savova et al., 2010). CUIs aim at abstracting our representations from the lexical variability of medical terminology and capturing the clinically relevant terms in a document leaving out the non-essential and potentially noisy lexical items. Each CUI can be either asserted or negated, as determined by the cTAKES negation module.

Although cTAKES is capable of extracting most CUIs that exist in the UMLS, we only include the CUIs that are listed in phenotype-specific dictionaries. The dictionaries are created manually by domain experts and define the terms that are relevant for each phenotype. Thus, we model each patient $\vec{x}$ as a vector of CUIs where each element n indicates the frequency of occurrence of the respective CUI_n in the records for this patient.

3.2 Models

To perform the classification and to estimate the informativeness of an instance during active learning, we need to evaluate the posterior probability $p(c_i|\vec{x})$, where c_i is the class indicating the relevance of the patient $\vec{x}$ for the given phenotype. For that purpose, we utilize a multinomial Naive Bayes model, which is widely used in document classification.

Naive Bayes classifiers possess several useful properties that make them particularly appropriate for active learning: (1) training and classification speed, (2) ability to produce a probability distribution over the target classes, and (3) ability to perform multi-class classification. Because active learning requires many rounds of retraining (potentially as many as the number of training examples), the first property is crucial for using active learning in practice. The second property is desirable for evaluating the level of uncertainty of the learner over the class predictions. Finally, the third property is important since some of our datasets include more than two classes.

We model the posterior probability as follows:

$$p(c_i|\vec{x}) = \frac{1}{Z} p(c_i) \prod_{n=1}^{N} p(CUI_n|c_i)^{x_n} \qquad (1)$$

Where $p(c_i)$ is the prior probability of class c_i, N is the number of CUIs in the phenotype-specific dictionary, CUI_n is the nth CUI in that dictionary, x_n is the frequency of CUI_n in $\vec{x}$, and Z (evidence) is the scaling factor. We determine the model parameters, $p(c_i)$ and $p(CUI_n|c_i)$, using maximum likelihood estimation with Laplace smoothing from the training data. For classification we predict the label c as:

$$c = \arg\max_i p(c_i|\vec{x}) \qquad (2)$$

For active learning, we utilize a framework known as *uncertainty sampling* (Lewis and Gale, 1994; Schein and Ungar, 2007). In this framework, the learner requests a label for the instance it is most uncertain how to label. We evaluate the level of uncertainty using the prediction margin metric (Schein and Ungar, 2007) which is defined as:

$$prediction\ margin = |p(c_1|\vec{x}) - p(c_2|\vec{x})| \qquad (3)$$

Where c_1 and c_2 are the two most probable classes for the patient $\vec{x}$ according to the model.

3.3 Datasets

In this work we utilize four datasets all of which were created within the i2b2 initiative (Ananthakrishnan et al., 2012; Ananthakrishnan et al., 2013b; Ananthakrishnan et al., 2013a; Xia et al., 2012). We show various important characteristics of our datasets in Table 1. Domain experts defined the ICD-9 codes relevant for each phenotype. These were then used to create the initial cohort from the 6 million+ patient EHR of the Partners Healthcare System. From that initial cohort, 600 patients were randomly chosen for manual labeling.

[1] http://incubator.apache.org/ctakes/

Each patient chart was reviewed by a domain expert and labeled at the patient level for CASE or NON-CASE (2-way labeling) for Ulcerative Colitis and Crohn's Disease; CASE, NON-CASE, or UNKNOWN (3-way labeling) for Type II Diabetes; CASE, NON-CASE, PROBABLE, UNKNOWN, or IRRELEVANT (5-way labeling) for Multiple Sclerosis. In our experiments, we used only the clinical narrative data, not a combination of structured and unstructured data. The predominant class for each phenotype was CASE.

3.4 Experimental Setup

Active learning is typically evaluated by comparing the learning curves for passive and active learning-based data selection methods. We generated the learning curves in the style of N fold cross validation ($N = 10$). Within each fold, we have a held out test set and a pool of unlabeled examples. We begin by randomly selecting the seed set of size S, removing it from the pool, and training a model. To produce a point of the active learning curve, we apply the model to the pool of remaining unlabeled data and select the most informative example using the prediction margin metric defined in Equation 3. We move the selected example to the training set, retrain the model, and evaluate its performance on the held out test set. In parallel, to produce a point of the passive learning curve, we select a single example from the pool randomly. We continue this process in an iterative fashion until the pool is exhausted. We repeat this for each of the ten folds and average the resulting learning curves.

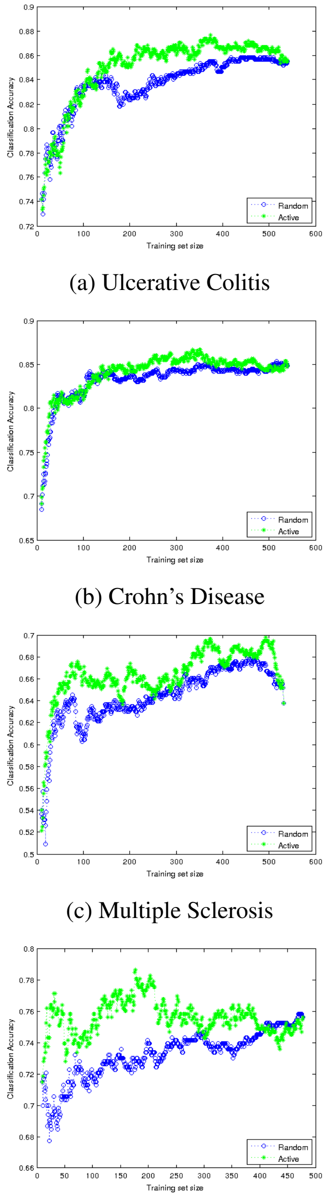
In addition, we conduct a series of experiments for each phenotype in which we vary the size of the seed set S. Our motivation is to explore the sensitivity of active learning to the size of the initial seed set. We only try several relatively small seed set sizes. Larger seed sets may erase the gains that could otherwise be obtained by active learning.

In this work, we do not compare the performance of the models accross different phenotypes. Instead, we focus on comparing the performance of active learning against the passive learning baseline.

In practice, active learning is used for selecting examples for subsequent labeling from the pool of unlabeled data. This scenario is simulated in our experiments – we utilize the gold standard data, but we hide the labels from the model. The label is revealed only after the instance is selected and is ready to be added to the training set. This is a common practice used in most published studies of active learning.

4 Results

For each phenotype, we construct the learning curves for different sizes of the seed set. The results are shown in Figures 1, 2, and 3, which include the learning curves for seed sizes $S = 10$, $S = 30$, and $S = 50$ respectively.



(a) Ulcerative Colitis

(b) Crohn's Disease

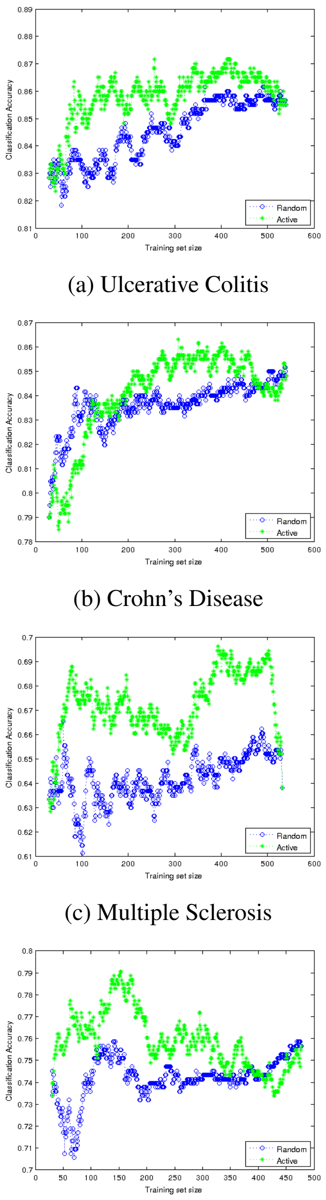
(c) Multiple Sclerosis

(d) Type II Diabetes

Figure 1: Passive vs. active learning performance on held-out data ($S = 10$)

Phenotype	Total Instances	Number of Classes	Proportion of Predominant Class
Ulcerative Colitis	600	2	0.630
Crohn’s Disease	600	2	0.665
Multiple Sclerosis	595	5	0.395
Type II Diabetes	600	3	0.583

Table 1: Dataset Characteristics



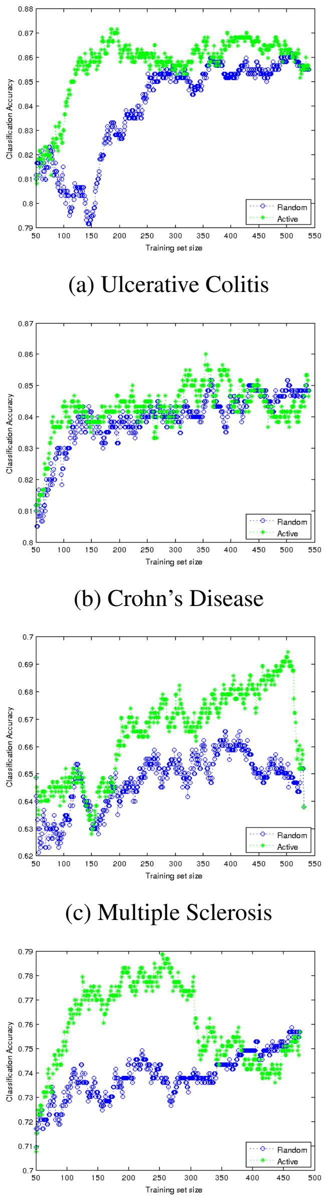
(a) Ulcerative Colitis

(b) Crohn’s Disease

(c) Multiple Sclerosis

(d) Type II Diabetes

Figure 2: Passive vs. active learning performance on held-out data ($S = 30$)



(a) Ulcerative Colitis

(b) Crohn’s Disease

(c) Multiple Sclerosis

(d) Type II Diabetes

Figure 3: Passive vs. active learning performance on held-out data ($S = 50$)

For each plot we also compute the area under the active learning and passive learning curves. We report the difference between the two curves in Table 2.

5 Discussion and Conclusion

As we see in Figures 1, 2, and 3, for all phenotypes, active learning curves lie above passive

Seed Size	Ulcerative Colitis	Crohn's Disease	Multiple Sclerosis	Type II Diabetes
10	6.90	4.17	10.50	11.05
30	6.64	2.21	15.43	7.49
50	8.63	1.75	8.61	8.90

Table 2: Difference between areas under the curve (Active - Passive)

learning curves for most sizes of the training sets. This means that the models trained on the data selected via active learning typically perform better than the models trained using random sampling. This result is also supported by the fact that the difference between the areas under the active and passive learning curves in Table 2 was positive for all of our experiments.

For most models, active learning reaches the level of the best random sampling performance with fewer than 200 examples, or about 1/3 of the data. This potentially translates into manual annotation savings of about 2/3. Moreover, the best active learning performance is often *above* that of the random sampling baseline. For example, consider Figure 3d. The sizes of the training set in the range between approximately 150 and 350 produce better performance than the best performance of the model trained on the randomly sampled data. At about 260 training examples, the performance of active learning approaches 0.79, which is at least 3 percentage points higher than the performance of the passive learning baseline that it achieves with the *entire* training set.

Although active learning consistently outperformed the passive learning baseline in most of our experiments, occasionally active learning performed worse at certain training set sizes. Consider Figure 2b. During early stages of learning (training set sizes of about 50-130), the passive curve lies above the active learning curve (although active learning recovers later on). We hypothesize that the reason for this behavior lies in outlier selection. Because outliers often do not fit into one of the predefined classes, the classifier is often uncertain about their labels, recommending their selection during active learning. At the same time, the outliers do not help to clarify the decision boundary, negatively affecting the performance. We leave a further investigation into the nature of this behavior for future work.

In other cases, the active learning briefly dips below the passive learning curve at the very end of the selection process. Although this behavior is observed in several cases (e.g. 1d, 2b, 3d), it is unlikely to be of consequence in practice. Active learning would typically be stopped at a much earlier stage, e.g. when 1/3 or 1/2 of the data has been annotated. Nevertheless, it would still be interesting to uncover the conditions leading to this behavior and we leave this investigation for future work.

Both of these scenarios, where active learning performed worse than random sampling, highlight the need for developing stopping criteria for active learning such as (Laws and Schätze, 2008; Bloodgood and Vijay-Shanker, 2009). In a practical application of active learning, a held-out test set is unlikely to be available and some automated means of tracking the progress of active learning is needed. We plan to pursue this avenue of research in the future. In addition to that, we plan to explore the portability of the models trained via active learning. It would also be interesting to investigate the effect of swapping the base classifier: in this work we collect the data for annotation using a multinomial Naive Bayes model. It is still not clear whether the gains obtained by active learning would be preserved if a model was trained on the selected data using a different classifier (e.g. SVM).

Finally, in addition to investigating the performance of active learning across different phenotypes, we also looked at the effects of varying the size of the seed set S. We did not find a clear correlation between the size of the seed set and active learning performance. However, the relationship may exist and could potentially be uncovered if a larger set of seed set sizes was used. We leave the further investigation in this area for future work.

In this work, we explored the use of active learning for several phenotyping tasks. Supervised learning is frequently used for phenotype creation, but the manual annotation that is required for model training is expensive. Active learning offers a way to reduce the annotation costs by involving the classifier in the data selection process. During active learning, the classifier chooses the unlabeled examples it views as informative, thus eliminating the need to annotate the examples that do not contribute to determining the decision boundary. We demonstrated that active learning outperforms the traditionally used passive learning baseline, potentially producing annotation cost savings of up to two-thirds of what is required by the passive baseline.

Acknowledgements

The work described was supported by Temporal History of Your Medical Events (THYME) NLM R01LM010090 and Integrating Informatics and Biology to the Bedside (i2b2) NCBO U54LM008748. Thanks to the anonymous reviewers for thorough and insightful comments. Guergana K. Savova is on the Advisory Board of Wired Informatics, LLC which provides services and products for clinical NLP applications.

References

A.N. Ananthakrishnan, T. Cai, S. Cheng, P.J. Chen, G. Savova, R.G. Perez, V.S. Gainer, S.N. Murphy, P. Szolovits, K. Liao, et al. 2012. Improving case definition of crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Gastroenterology*, 142(5):S–791.

A. Ananthakrishnan, V. Gainer, T. Cai, Guzman P.R., S. Cheng, G. Savova, P. Chen, P. Szolovits, Z. Xia, P. De Jager, S. Shaw, S. Churchill, E. Karlson, and I. Kohane. 2013a. Similar risk of depression and anxiety following surgery or hospitalization for crohn's disease and ulcerative colitis. *Am J Gastroenterol*.

A.N. Ananthakrishnan, V.S. Gainer, R.G. Perez, T. Cai, S.C. Cheng, G. Savova, P. Chen, P. Szolovits, Z. Xia, P.L. Jager, et al. 2013b. Psychiatric co-morbidity is associated with increased risk of surgery in crohn's disease. *Alimentary pharmacology & therapeutics*, 37(4):445–454.

M. Bloodgood and K. Vijay-Shanker. 2009. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 39–47. Association for Computational Linguistics.

O. Bodenreider and A.T. McCray. 2003. Exploring semantic groups through visual approaches. *Journal of biomedical informatics*, 36(6):414.

J. Chen, A. Schein, L. Ungar, and M. Palmer. 2006. An empirical study of the behavior of active learning for word sense disambiguation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 120–127, Morristown, NJ, USA. Association for Computational Linguistics.

Y. Chen, S. Mani, and H. Xu. 2011. Applying active learning to assertion classification of concepts in clinical text. *Journal of Biomedical Informatics*.

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

D. Dligach and M. Palmer. 2011. Good seed makes a good crop: accelerating active learning using language modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 2. Association for Computational Linguistics.

R.L. Figueroa, Q. Zeng-Treitler, L.H. Ngo, S. Goryachev, and E.P. Wiechmann. 2012. Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association*, 19(5):809–816.

U. Hahn, E. Beisswanger, E. Buyko, and E. Faessler. 2012. Active learning-based corpus annotationâĭthe pathojen experience. In *AMIA Annual Symposium Proceedings*, volume 2012, page 301. American Medical Informatics Association.

S.C.H. Hoi, R. Jin, J. Zhu, and M.R. Lyu. 2006. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424. ACM.

A.N. Kho, J.A. Pacheco, P.L. Peissig, L. Rasmussen, K.M. Newton, N. Weston, P.K. Crane, J. Pathak, C.G. Chute, S.J. Bielinski, et al. 2011. Electronic medical records for genetic research: results of the emerge consortium. *Sci Transl Med*, 3(79):79rel.

A.N. Kho, M.G. Hayes, L. Rasmussen-Torvik, J.A Pacheco, W.K. Thompson, L.L. Armstrong, J.C. Denny, P.L Peissig, A.W. Miller, W. Wei, et al. 2012. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association*, 19(2):212–218.

I.S. Kohane, S.E. Churchill, and S.N. Murphy. 2012. A translational engine at the national scale: informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association*, 19(2):181–185.

I.J. Kullo, J. Fan, J. Pathak, G.K. Savova, Z. Ali, and C.G. Chute. 2010. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *Journal of the American Medical Informatics Association*, 17(5):568–574.

F. Laws and H. Schätze. 2008. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pages 465–472. Association for Computational Linguistics.

D.D. Lewis and W.A. Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, New York, NY, USA. Springer-Verlag New York, Inc.

C. Lin, H. Canhao, T. Miller, D. Dligach, R.M. Plenge, E.W. Karlson, and G. Savova. 2012a. Maximal information coefficient for feature selection for clinical document classification. In *ICML Workshop on Machine Learningfor Clinical Data.*

C. Lin, H. Canhao, T. Miller, D. Dligach, R.M. Plenge, E.W. Karlson, and G.K. Savova. 2012b. Feature engineering and selection for rheumatoid arthritis disease activity classification using electronic medical records. In *ICML Workshop on Machine Learning for Clinical Data Analysis.*

R.M. Long and J.M. Berg. 2011. What to expect from the pharmacogenomics research network. *Clinical Pharmacology & Therapeutics*, 89(3):339–341.

A. McCallum and K. Nigam. 1998. Employing em and pool-based active learning for text classification. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 350–358, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

C.A. McCarty, R.L. Chisholm, C.G. Chute, I.J. Kullo, G.P. Jarvik, E.B. Larson, R. Li, D.R. Masys, M.D. Ritchie, D.M. Roden, et al. 2011. The emerge network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics*, 4(1):13.

T. Miller, D. Dligach, and G Savova. 2012. Active learning for coreference resolution. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 73–81, Montréal, Canada, June. Association for Computational Linguistics.

F. Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. In *Technical Report, Swedish Institute of Computer Science*.

J.A. Pacheco, P.C. Avila, J.A. Thompson, M. Law, J.A. Quraishi, Alyssa K. Greiman, E.M. Just, and A. Kho. 2009. A highly specific algorithm for identifying asthma cases and controls for genome-wide association studies. In *AMIA Annual Symposium Proceedings*, volume 2009, page 497. American Medical Informatics Association.

G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, and C.G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

A.I. Schein and L.H. Ungar. 2007. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265.

B. Settles, M. Craven, and S. Ray. 2008. Multiple-instance active learning. *Advances in Neural Information Processing Systems (NIPS)*, 20:1289–1296.

B. Settles. 2009. Active learning literature survey. In *Computer Sciences Technical Report 1648 University of Wisconsin-Madison.*

H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, New York, NY, USA. ACM.

C.A. Thompson, M.E. Califf, and R.J. Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the Sixteenth International Conference on Machine Learnin*, pages 406–414. Citeseer.

S. Tong and E. Chang. 2001. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM New York, NY, USA.

S. Tong and D. Koller. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

C. Waudby, R. Berg, J. Linneman, L. Rasmussen, P. Peissig, L. Chen, and C. McCarty. 2011. Cataract research using electronic health records. *BMC ophthalmology*, 11(1):32.

Z. Xia, R. Bove, T. Cai, S. Cheng, R.N.G. Perez, V.S. Gainer, S.N. Murphy, P. Chen, G.K. Savova, K. Liao, E.W. Karlson, S. Shaw, S. Ananthakrishnan, P. Szolovits, S. Churchill, I.S. Kohane, R.M. Plenge, and Philip L.D. 2012. Leveraging electronic health records for research in multiple sclerosis. In *European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS).*

J. Zhu and E. Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 783–790.