ACR Open Rheumatology
Vol. 3, No. 9, September 2021, pp 593–600
DOI 10.1002/acr2.11289

American College of Rheumatology
Empowering Rheumatology Professionals

Improving the Efficiency of Clinical Trial Recruitment Using an Ensemble Machine Learning to Assist With Eligibility Screening

Tianrun Cai,[1] Fiona Cai,[2] Kumar P. Dahal,[1] Gabrielle Cremone,[1] Ethan Lam,[1] Charlotte Golnik,[1] Thany Seyok,[1] Chuan Hong,[3] Tianxi Cai,[3] and Katherine P. Liao[4]

Objective. Efficiently identifying eligible patients is a crucial first step for a successful clinical trial. The objective of this study was to test whether an approach using electronic health record (EHR) data and an ensemble machine learning algorithm incorporating billing codes and data from clinical notes processed by natural language processing (NLP) can improve the efficiency of eligibility screening.

Methods. We studied patients screened for a clinical trial of rheumatoid arthritis (RA) with one or more International Classification of Diseases (ICD) code for RA and age greater than 35 years, from a tertiary care center and a community hospital. The following three groups of EHR features were considered for the algorithm: *1*) structured features, *2*) the counts of NLP concepts from notes, *3*) health care utilization. All features were linked to dates. We applied random forest and logistic regression with least absolute shrinkage and selection operator penalty against the following two standard approaches: *1*) one or more RA ICD code and no ICD codes related to exclusion criteria ($\text{Screen}_{\text{RAICD1+EX}}$) and *2*) two or more RA ICD codes ($\text{Screen}_{\text{RAICD2}}$). To test the portability, we trained the algorithm at one institution and tested it at the other.

Results. In total, 3359 patients at Brigham and Women's Hospital (BWH) and 642 patients at Faulkner Hospital (FH) were studied, with 461 (13.7%) eligible patients at BWH and 84 (13.4%) at FH. The application of the algorithm reduced ineligible patients from chart review by 40.5% at the tertiary care center and by 57.0% at the community hospital. In contrast, $\text{Screen}_{\text{RAICD2}}$ reduced patients for chart review by 2.7% to 11.3%; $\text{Screen}_{\text{RAICD1+EX}}$ reduced patients for chart review by 63% to 65% but excluded 22% to 27% of eligible patients.

Conclusion. The ensemble machine learning algorithm incorporating billing codes and NLP data increased the efficiency of eligibility screening by reducing the number of patients requiring chart review while not excluding eligible patients. Moreover, this approach can be trained at one institution and applied at another for multicenter clinical trials.

INTRODUCTION

One of the major challenges of a successful clinical trial is efficiently identifying eligible patients (1). Drug development costs in many cases exceed $1 billion, and clinical trials often account for more than one-third of this cost (2,3); more than half of these expenses are associated with delays due to prolonged recruitment (4). With the increasing availability of electronic health record (EHR) systems, there is growing interest in using EHR data to assist with patient recruitment for clinical trials (5). Patients satisfying the eligibility criteria of a trial can be identified via eligibility screening (ES). ES typically requires labor-intensive and costly reviews of patients' medical histories (6-8). Furthermore, manual ES limits the number of patients who can be evaluated, potentially leading to an insufficient number of participants enrolled in trials within their allocated time frames. Carlisle et al found that 19% of registered trials terminated either failed to meet accrual goals or were completed with less than 85% of expected enrollment (9). Thus, improvements in ES efficiency could help lower costs for drug discovery and expedite the process of patient recruitment (10).

Supported by the National Institutes of Health (R01HL127118 and P30 AR072577).

[1]Tianrun Cai, MD, Kumar P. Dahal, MS, Gabrielle Cremone, BS, Ethan Lam, BS, Charlotte Golnik, BS, Thany Seyok, BS: Brigham and Women's Hospital, Boston, Massachusetts; [2]Fiona Cai: Massachusetts Institute of Technology, Cambridge, Massachusetts; [3]Chuan Hong, PhD, Tianxi Cai, ScD: Harvard University, Boston, Massachusetts; [4]Katherine P. Liao, MD, MPH: Brigham and Women's Hospital, Harvard University, and Veterans Affairs Boston Healthcare System, Boston, Massachusetts.

Drs. Tianxi Cai and Liao contributed equally.

No potential conflicts of interest relevant to this article were reported.

Address correspondence to Tianrun Cai, MD, Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, 60 Fenwood Road, Boston, MA 02115. Email: tcai1@bwh.harvard.edu.

Submitted for publication November 30, 2020; accepted in revised form May 18, 2021.

Leveraging EHR data to devise more efficient clinical trial ES approaches is currently an active area of research (11). Approaches have included the use of natural language processing (NLP), information extraction, and machine learning approaches in an emergency department setting (12,13). Other approaches, such as random forest (RF) and the use of cosine similarities, have also been applied post hoc to assess whether EHR data could have enhanced recruitment in completed randomized controlled trials (14,15). However, there are still many challenges on leveraging EHR data for clinical trial ES. Butler et al performed text-mining on Alzheimer disease clinical trials on clinicaltrials.gov and found that 40% of the most commonly used eligibility criteria were not available in the EHR data of patients with Alzheimer disease (16).

The challenges to clinical trial recruitment highlight the need for a more systematic approach toward developing an automated ES while taking advantage of the structured and unstructured data in the EHR system. Currently, within the LiiRA (Lipids, Inflammation, and Cardiovascular Risk in Rheumatoid Arthritis) study (17), the most time-consuming aspect of patient recruitment is identifying eligible patients from manual chart reviews when using EHRs to support patient recruitment. The LiiRA study team incorporated rule-based filters to reduce the number of patients requiring chart review to identify eligible patients. Their filters rely on available structured data such as demographic information and International Classification of Diseases (ICD) codes (ICD-9 and ICD-10, signifying the ninth and tenth revisions of the classification system, respectively). Most commonly, patients are excluded from the LiiRA study if their count of ICD codes for rheumatoid arthritis (RA), denoted as "RA_{ICD}," is less than one. However, this approach of minimizing manual labor in ES still results in a substantial amount of chart review performed.

In this study, we tested whether screening by ensemble machine learning algorithm (SMALL) could improve the efficiency of clinical trial recruitment using data from the LiiRA study. We hypothesize that incorporating data more broadly from EHRs to develop a screening algorithm can significantly reduce the number of patients requiring chart review while not screening out eligible patients.

PATIENTS AND METHODS

Study population. This study was conducted on patients from two health care centers, Brigham and Women's Hospital (BWH; a tertiary care center), and Faulkner Hospital (FH; a community hospital), both in Boston, Massachusetts. Both are sites for the LiiRA study. Our proposed method was built and evaluated with an EHR cohort of 3359 patients at BWH and 642 patients at FH who underwent chart review. All patients had one or more RA_{ICD}, including ICD-9 714* (excluding 714.3), ICD-10 M05*, and ICD-10 M06*. Using the LiiRA recruitment criteria, the eligibility status of all patients was determined via manual chart review by the LiiRA study team. At both institutions, manual ES was performed for each patient from 2016 to 2020. For inclusion in the LiiRA study, patients must meet all of the following conditions: diagnosis of RA, age of more than 35 years, and fluency in English. We excluded patients with any of following conditions based on the LiiRA exclusion criteria: *1*) receipt of a statin or biologic disease-modifying antirheumatic drugs within 6 months before the chart review date, *2*) history of melanoma, *3*) history of psoriatic arthritis, *4*) history of lymphoma within 5 years before chart review date, *5*) pregnancy within 1 year before chart review date, *6*) asthma with active wheezing, *7*) active human immunodeficiency virus, *8*) active hepatitis B or C, and *9*) active tuberculosis.

Data source and features. EHR data were requested from the Partners HealthCare System Research Patient Data Registry, which contains comprehensive patient data, including structured data such as demographic information and billing codes for diagnosis, procedure, medication prescription, and laboratory test and unstructured clinical notes (18).

The following three types of features were considered for algorithm training in our study: structured features, NLP features, and health care utilization. The structured features include demographic information, diagnosis codes of ICD-9 and ICD-10 for disease status, and medication prescription codes. For demographic information such as age and English fluency, we only used structured features. For the remaining criteria, we generated both structured and unstructured features with different time frames relevant to approximate the inclusion or exclusion criteria (Table 1). For example, the criterion "history of lymphoma within 5 years" was emulated using the EHR data by generating features for both the structured feature (the count of lymphoma ICD codes) and the NLP feature (lymphoma concepts present in narrative notes) within 5 years. For criteria with no time frame, such as "history of melanoma," we generated 1-year, 2-year, and total counts for both structured and NLP features for each patient. The total count of notes was used as a proxy for health care utilization.

To extract NLP concepts, we created a dictionary listing the concepts for each of the relevant recruitment criteria using the Unified Medical Language System (UMLS) (19). Specifically, each dictionary contained a list of clinical terms and synonyms to represent the concept of each criteria item. Each concept was then mapped to a concept unique identifier in UMLS. For example, based on the exclusion criteria "lymphoma diagnosis within 5 years," we created a dictionary for the concept of "lymphoma,", including the term "lymphoma" and all its synonyms existing in UMLS, such as "germinoblastoma". We processed the clinical notes with a previously developed NLP tool, Narrative Information Linear Extraction (NILE) (20,21) to obtain the number of times the concept is mentioned in the notes. NILE can also distinguish positive mentions (eg, the mention of "melanoma" in a sentence such as "new melanoma of the left upper medial shoulder"), uncertain mentions (eg, the mention of "melanoma" in a sentence such as "there was a broad field that was suspicious for melanoma in

Table 1. EHR features approximating the inclusion and exclusion criteria for LiiRA used in model training for BWH and FH

	Structured Feature	Unstructured Feature	Timeframe				
			6 months	1 year	2 years	5 years	All years
Inclusion	Age						
	English-fluency						
	RA_{ICD}	RA_{NLP}		✓	✓		✓
Exclusion	JRA_{ICD}	JRA_{NLP}		✓	✓		✓
	PsA_{ICD}	PsA_{NLP}		✓	✓		✓
	$Melanoma_{ICD}$	$Melanoma_{NLP}$		✓	✓		✓
	TB_{ICD}	TB_{NLP}		✓	✓		✓
	$Asthma_{ICD}$	$Asthma_{NLP}$		✓	✓		✓
	$HepatitisB_{ICD}$	$HepatitisB_{NLP}$		✓	✓		✓
	$HepatitisC_{ICD}$	$HepatitisC_{NLP}$		✓	✓		✓
	HIV_{ICD}	HIV_{NLP}		✓	✓		✓
	$bDMARDs_{Med}$	$bDMARDs_{NLP}$	✓				
	$Statin_{Med}$	$Statin_{NLP}$	✓				
	$Pregnancy_{ICD}$	$Pregnancy_{NLP}$		✓			
	$Lymphoma_{ICD}$	$Lymphoma_{NLP}$				✓	
HU		*Note Count*					✓

Abbreviation: $*_{ICD}$, sum of ICD9 and ICD10 code counts; $*_{Med}$, sum of medication code count; $*_{NLP}$, concept count from notes; bDMARD, biologic disease-modifying antirheumatic drug; BWH, Brigham and Women's Hospital; EHR, electronic health record; FH, Faulkner Hospital; HIV, human immunodeficiency virus; HU, healthcare utilization; ICD, International Classification of Diseases; JRA, juvenile rheumatoid arthritis; LiiRA, Lipids, Inflammation, and Cardiovascular Risk in Rheumatoid Arthritis; NLP, natural language processing; PsA, psoriatic arthritis; RA, rheumatoid arthritis; TB, tuberculosis.

situ"), and negations (eg, the mention of "melanoma" in a sentence such as "there was no evidence of melanoma") for the concepts. In this study, only the positive mentions of concepts were collapsed into counts. The uncertain and negated mentions were not counted.

The gold-standard labels were obtained using manual chart review with information for patients who were eligible for enrollment into LiiRA, fulfilling all the inclusion/exclusion criteria.

The final dataset for model training included the gold-standard eligibility labels and three categories of 65 features, denoted as "F_{select}," as presented in Table 1. For features with timeframes defined in criteria, we only extracted data for both structured and NLP features in the specified timeframes. For other features except for age and English fluency, we obtained the number of counts in the following three timeframes: 1 year, 2 years, and all years. In other words, the RA_{ICD} and the count of RA concept (RA_{NLP}) in a 1-year time frame, 2-year time frame, and all years were six independent features for the criterion "diagnosis of RA." Because most features were represented in counts and tend to be highly skewed, a transformation by $x \rightarrow \log(x+1)$ was applied before fitting our model to improve model training. We use F to denote the full feature vector and F_c to denote the subvector of features that directly correspond with the inclusion/exclusion criteria. See the supplementary materials for details on F_c.

SMALL algorithm. The proposed SMALL algorithm is an ensemble classification algorithm that takes a model average of two algorithms, RF and logistic regression with least absolute shrinkage and selection operator penalty (logistic LASSO), to determine patients requiring chart review. The ensemble approach, by averaging over multiple algorithms, reduces the variation in performance across different datasets (22). The flow chart of the SMALL algorithm is presented in Figure 1.

The SMALL algorithm first fit a logistic LASSO with the full feature vector F and produce a predicted probability of being eligible as $p_{lasso}(F)$. It then fits an RF with the predicted risk from the LASSO fitting and the features F_c to produce another predicted probability $p_{rf}(F)$. The final predicted probability of being eligible for a patient with feature vector F is $p_{SMALL}(F) = expit\left[0.5logit\left\{p_{rf}(F)\right\} + 0.5logit\{p_{lasso}(F)\}\right]$, in which $expit(x)$ is the inverse of $logit(x)$. We use fewer features in the RF algorithm to control for overfitting, which is particularly important when using a small number of training samples.

To validate the algorithms, we randomly split the data into training sets for algorithms training and validation sets for evaluating the performance of the algorithms. To improve the stability of the parameter estimates, we repeatedly split the data points into training and validation sets for multiple times and reported average estimates. We used 1400 and 200 samples for validation at BWH and FH. We trained the algorithms using a range of training sample sizes from 50 to 1900 for BWH and 442 for FH, which allowed us to investigate the smallest label size needed to ensure satisfactory performance of the algorithms. To select an optimal cutoff value c for classifying eligibility status and to evaluate the performance of SMALL, we estimated the receiver operating characteristic curve for the algorithm using the validation data. Specifically, for a given c, we estimated the sensitivity level against one specificity level of $p_{SMALL} \geq c$ for across a range of c. We chose an optimal cutoff value, $\widehat{cut}_{valid}$, as the largest threshold value such that the corresponding sensitivity is at least 0.98 for

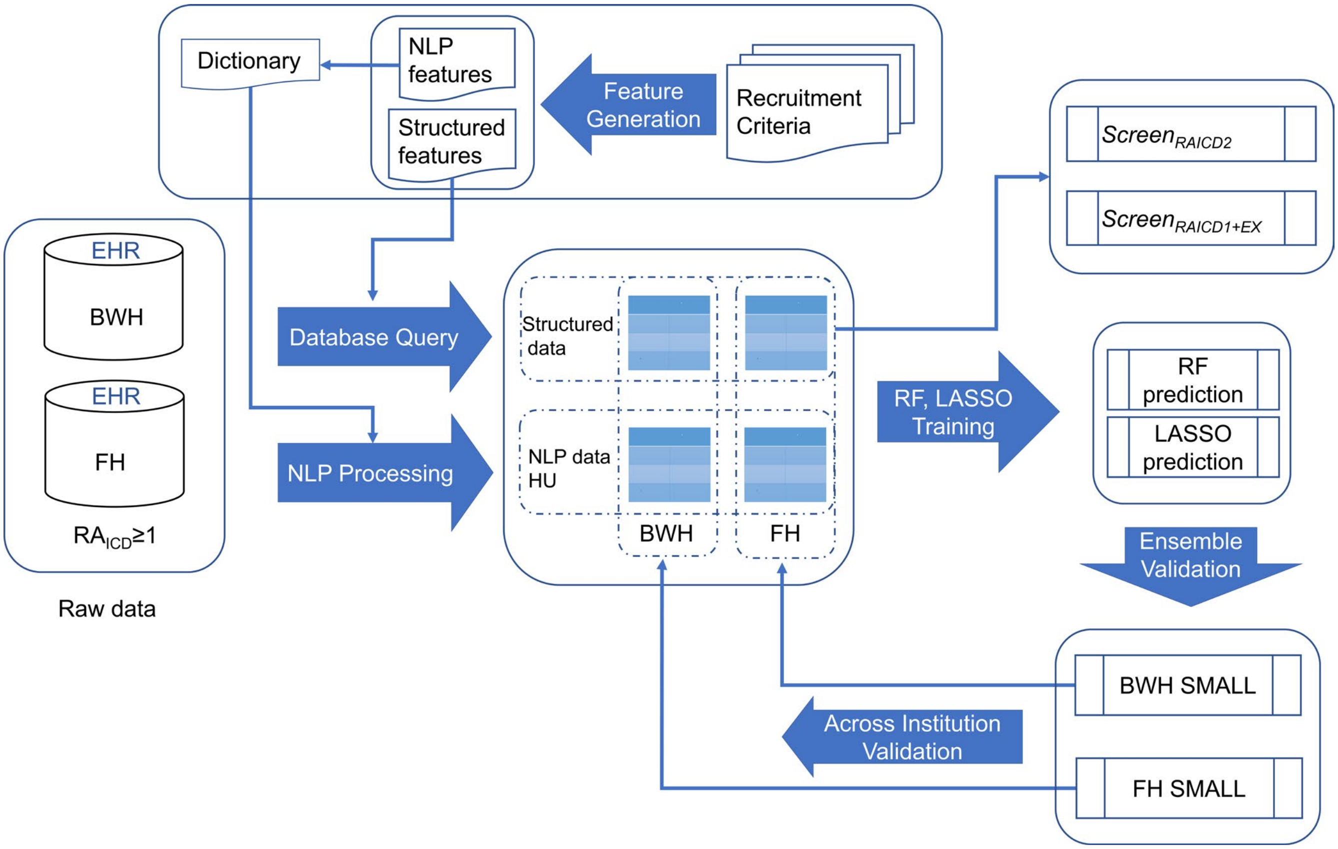

Figure 1. Overview of study design (created using Photoshop). Data were extracted from electronic health record (EHRs) of Brigham and Women's Hospital (BWH) and Faulkner Hospital (FH), resulting in sets of structured data and natural language processing (NLP) data. Data were trained against patients considered eligible by chart review using both random forest and logistic least absolute shrinkage and selection operator (LASSO) at each institution and validated at the other. Performance of ensemble algorithm compared with rule-based International Classification of Diseases (ICD) screens (screening with at least two ICD codes for rheumatoid arthritis [RA] [$\mathrm{Screen}_{\mathrm{RAICD2}}$] and screening with at least one ICD code for RA and no ICD codes of any exclusionary codes [$\mathrm{Screen}_{\mathrm{RAICD1+EX}}$]). RA_{ICD}≥1, at least one count of RA ICD codes; RF, random forest; PPV, positive predictive value, SE, sensitivity; SMALL, screening by ensemble machine learning algorithm.

both BWH and FH. This threshold represents the optimal value to minimize the number of charts needed to review while maintaining a high sensitivity. Patients with $p_{SMALL} \geq \widehat{cut}_{valid}$ were then classified as potentially eligible for chart review. We reported the performance of SMALL based on the positive predictive value (PPV) and the percentage of patients needed to review for the rule $p_{SMALL} \geq \widehat{cut}_{valid}$.

Comparison of proposed method and benchmarks. For comparison, the following two rule-based benchmark methods were applied to the data: *1*) at least two ICD codes for RA ($\mathrm{Screen}_{\mathrm{RAICD2}}$) and *2*) at least one ICD code for RA and no ICD codes of any exclusionary codes ($\mathrm{Screen}_{\mathrm{RAICD1+EX}}$). The goal of the algorithm was to achieve high sensitivity to not exclude potentially eligible patients while screening out patients with a low probability for eligibility based on billing codes alone. For the latter, only a modest improvement in PPV would be required. We also compared this with the logistic LASSO alone and the RF algorithm alone trained with the full feature set.

Porting algorithm across institutions. To study the portability of the proposed algorithm SMALL for ES from one institution to another, we applied the algorithm trained with BWH data to the FH data. This was additionally done using the FH data to train the algorithm and subsequently validated with the BWH data. All computation was conducted in R versions 3.5.0 and 4.0.3 (23). In particular, the *randomForest* package was used to build the RF model, and *glmnet* was used for LASSO (24,25).

RESULTS

LiiRA clinical study at BWH and FH. A total of 3359 subjects at BWH and 642 subjects at FH had at least one RA_{ICD} (Table 2). For patients at FH, the median RA_{ICD} was 21 and the maximum was 542. At BWH, the median RA_{ICD} was 76 and the maximum was 2119. At both institutions, manual ES was performed for each patient from 2016 to 2020, and eligibility was categorized as 0 (ineligible) or 1 (eligible).

Table 2. LiiRA patient statistics at BWH and FH

	BWH	*FH*
Total patients, n	3359	642
Eligible patients, n	461	84
Eligibility prevalence, %	13.7	13.4

Abbreviation: BWH, Brigham and Women's Hospital; FH, Faulkner Hospital; LiiRA, Lipids, Inflammation, and Cardiovascular Risk in Rheumatoid Arthritis.

The results of the SMALL algorithm compared with rule-based methods are presented in Table 3. Both rule-based methods had either a low PPV, which leads to a high percentage of ineligible patients reviewed, or a low sensitivity, which leads to missing a significant fraction of eligible patients. In contrast, using the SMALL algorithm on the BWH subjects, with modest improvement of PPV as well as improved sensitivity of 0.983, 40.5% of ineligible patients would be reduced from chart review, excluding only 1.7% of potentially eligible patients. At FH, while maintaining a sensitivity of 1, 57.0% of ineligible patients would be reduced from chart review.

Results for assessing the portability of the proposed method across institutions are displayed in Table 4. We found that by applying the SMALL algorithm trained using the BWH data to the FH data, the percentage of patients requiring chart review was 58.4%. Applying the SMALL algorithm trained at FH to the BWH data, the percentage of patients to screen was 63.9%.

DISCUSSION

In this study, we observed that the proposed method leveraging EHR data with NLP and application of the SMALL algorithm, which mirrored the trial's eligibility requirements, significantly reduced number of charts required for review to screen for eligibility. Additionally, the SMALL algorithm was able to uphold a high sensitivity value, eliminating patients who did not require chart review but maintaining almost all potentially eligible patients—98.3% at BWH and 100% at FH. The average chart review time required to determine a patient's eligibility for LiiRA was approximately 30 minutes. As seen in Table 3, using the proposed method, there would be 1176 fewer patients for chart reviews at BWH and 318 fewer patients for chart reviews at FH , saving an estimated 588 working hours (14.7 weeks) and 159 working hours (4.0 weeks) on the BWH and FH datasets, respectively.

As clinical trials are time intensive, one must also consider the additional time and resources required to train and develop an algorithm to screen for eligibility versus devoting all effort to reviewing charts. We provide the following estimates of the time and resources required based on our experience developing SMALL: *1*) building dictionaries and processing notes takes half a day, *2*) querying database for structured features takes half a day, and *3*) building and validating the algorithm takes 3 to 4 days. In total, approximately 1 week would be needed to obtain a new patient screening algorithm for one institution. A few more days would be required to apply an algorithm to another institution. The cohort size and the performance of database server and the application computer for processing notes and building the algorithm would potentially affect the total time required. We used Microsoft Structured Query Language (MSSQL) database server on a machine with 16 core central processing units (CPUs) with 32 gigabyte (GB) of memory and a regular PC with 64 GB of memory and 16 core CPUs for processing notes and building the SMALL algorithm. Furthermore, personnel trained in NLP and machine learning are required for replicating this method.

Prior to chart review, the rule-based approach does not consider other structured EHR data, such as the ICD counts for each exclusionary disease as negative predictors. For the RF and LASSO models, inclusion of the exclusionary features yielded a more accurate estimated probability of trial eligibility, even before manual screening. In addition, our method leveraged unstructured EHR data as well by using NLP on patient narrative notes to extract counts of concepts derived from recruitment criteria. This provided a more informative definition of each criterion, as billing code counts may be inaccurate. User-engineered features such as specific negations and temporal ordering of mentions may be useful to reflect specific inclusion or exclusion criteria. Incorporating total health care utilization as quantified by the total

Table 3. Performance comparison of SMALL based on 1900 and 442 training samples at BWH and FH versus rule-based approaches to screen for eligible patients

	SMALL		Screen$_{\text{RAICD2}}$		Screen$_{\text{RAICD1+EX}}$	
	BWH	FH	BWH	FH	BWH	FH
Sensitivity, %	98.0	98.0	99.6	96.4	78.1	84.5
PPV, %	21.8	24.3	14.0	15.3	25.8	25.5
Patients for review, %	65.0	50.4	86.8	97.9	44.6	46.5
Patients for review, n	2183	324	3289	545	1498	293
Patients incorrectly excluded, n	10	2	2	3	101	13

Abbreviation: BWH, Brigham and Women's Hospital; FH, Faulkner Hospital; PPV, positive predictive value; Screen$_{\text{RAICD1+EX}}$, Screening with at least one International Classification of Diseases code for rheumatoid arthritis and no International Classification of Diseases codes of any exclusionary codes; Screen$_{\text{RAICD2}}$, screening with at least two International Classification of Diseases codes for rheumatoid arthritis; SMALL, screening by ensemble machine learning algorithm.

Table 4. Performance characteristics demonstrating portability of data across institutions

	BWH → FH	FH → BWH
Sensitivity, %	99.1	98.3
PPV, %	23.2	21.6
Patients requiring chart review, %	58.4	63.9
Patients left to review, n	367	2146
Patients incorrectly excluded, n	1	8

Abbreviation: BWH, Brigham and Women's Hospital; FH, Faulkner Hospital; PPV, positive predictive value.
Training model using BWH data and validating with FH data versus training model using FH data and validating with BWH data. The sensitivity cutoff was set to 98%. The comparison of area under the curves (AUCs) of different algorithms using different training data size is displayed in Figure 2. At both BWH and FH, the screening by ensemble machine learning algorithm (SMALL) algorithm attained an AUC of approximately 0.9 when the training size reaches approximately 400 to 500. The SMALL algorithm performed similarly to logistic regression with least absolute shrinkage and selection operator penalty (logistic LASSO) at BWH but better than both logistic LASSO and random forest (RF) at FH. The RF algorithm generally attained lower performance, especially when the training samples are small, largely because of the higher model complexity.

number of clinical notes enhances the performance and accuracy of the model because the amount of utilization can vary dramatically across patients. Total health care utilization can affect the informativeness of an ICD code count. For example, a patient with 10 total visits, five of which are for RA, may be more likely to be a case than a patient with 1000 total visits, five of which are for RA.

The SMALL algorithm was shown to be reasonably portable across the FH and BWH patient cohorts. The algorithm trained at FH attained similar accuracy at BWH as the algorithm trained at BWH, whereas the BWH algorithm performed slightly worse than the FH algorithm on the FH data. This could be in part due to the difference in data recording between institutions. For example, one mention of juvenile RA (JRA) at one institution may be defined differently than one mention of JRA at a different institution.

It should be noted that high sensitivity is favored, specifically for clinical trials in which the population of eligible patients is small. This was the case for the LiiRA trial and most likely for other rheumatic conditions, in which the eligibility rate was 13.7% at BWH and 13.4% at FH. We found that increasing the sensitivity threshold at BWH would result in a tradeoff of a higher percentage of patient charts to review. This was also evident in the benchmark methods. If the population of eligible patients was larger, the proposed method would not be as practical because, according to the results of the model tested at BWH, 1.7% of eligible patients would be omitted by the algorithm. Furthermore, patient recruitment for clinical trials is often slow, and some studies may not be able to afford losing eligible patients.

Both RF and LASSO models were supervised with available gold-standard eligibility labels as response features. In previous studies, RFs have been shown to be useful for classification tasks, specifically in a binary decision such as assessing patient eligibility for clinical trials (12). LASSO, using shrinkage and feature importance, has also been successfully used in prediction models (26). Although ensemble learning did not lead to significant improvements in reducing patients for review at BWH using LASSO alone in this study, the ensemble method was beneficial in improving the screen at FH.

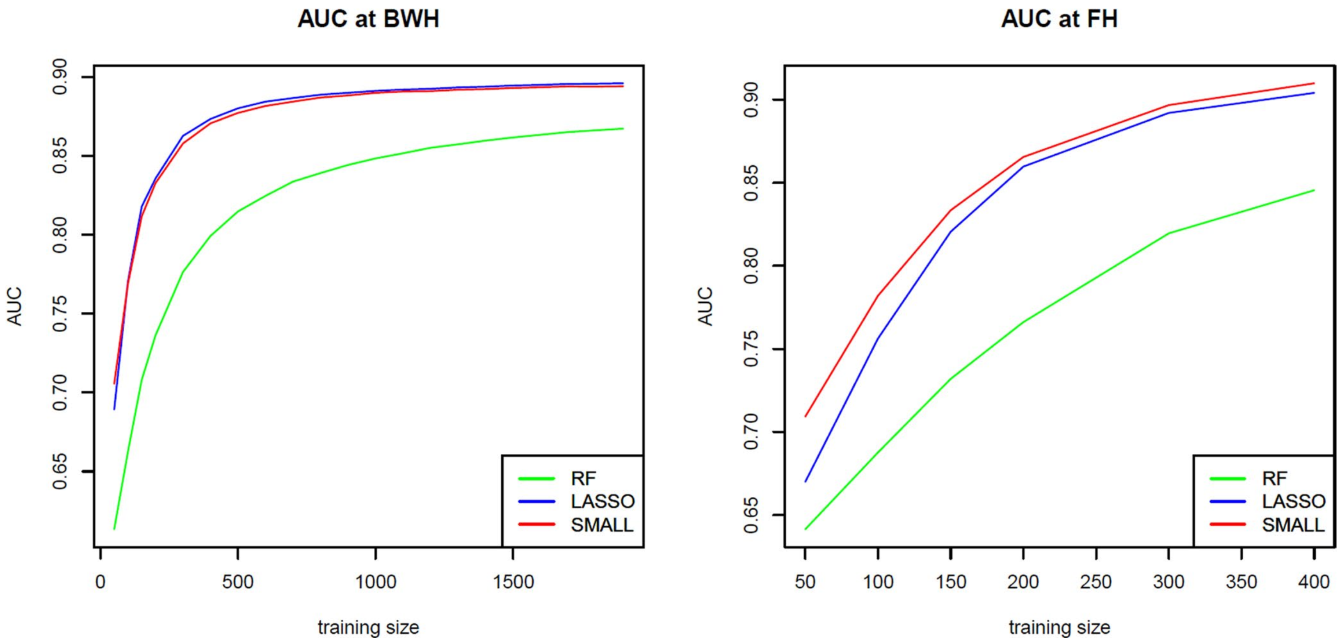

Figure 2. Comparison of areas under the curves (AUCs) of different algorithms with different training sizes. BWH, Brigham and Women's Hospital; FH, Faulkner Hospital, LASSO, least absolute shrinkage and selection operator; RF, random forest; SMALL, screening by ensemble machine learning algorithm.

The number of labels needed to train a stable algorithm inherently depend on the feature size and the model complexity. In our study, we observed that a training set of 400 to 500 subjects was required to achieve stable models using BWH data. The reported training sizes in the literature vary greatly. For example, Kopcke et al (14) sampled approximatley 30% (200 in number) of patients for building a rule-based algorithm with AUCs ranging from 0.81 to 0.99 for three different trials. Using a different method, Miotto et al used anywhere from 4 to 128 eligible patients for producing "target patients" for different trials (15). However, it was not clear from these studies whether the training sizes are sufficiently large to yield sufficiently stable algorithms.

Limitations to this study include the fact that BWH and FH share an administrative system. A few physicians practice at both sites, and thus there may be similarity in writing style and utilization of medical vocabularies for recording notes. In addition, they use the same EHR vendor. This reduced heterogeneity of the data may overestimate the transportability of the algorithm across the two institutions. As a proof-of-concept, we used one clinical trial as an example, and this approach will need to be tested in other conditions.

Future directions will include applying this workflow in real-time. Weng et al developed a real-time screening alert method to recruit patients for an ongoing study of post acute coronary syndrome (ACS) (27). Compared with different manual screening approaches, the proposed method had improved performance over time. However, this method mainly relied on a locally available decision support system, Vigilens, to generate a potentially eligible patient list using information derived from trial criteria such as ACS-specific medications and ICD-9 codes. Additional work will include considering methods that may be generalizable across systems.

Additional machine learning algorithms can be potentially used to further strengthen the efficiency of the model. More specifically, instead of simply averaging the RF and LASSO models, future work could implement a learning system that takes a weighted average of predicted probabilities from a larger number of supervising algorithms with weights reflecting the performance of the individual algorithm. In addition, further testing should be conducted through applying the method in other conditions and institutions. Finally, automating the steps for generating structured and NLP features to approximate the recruitment criteria using EHR data can further improve the efficiency of ES.

In conclusion, leveraging both unstructured and structure EHR data in conjunction with machine learning algorithms, the proposed method SMALL algorithm significantly reduced the number of charts required for ES, which is the most time-consuming aspect of patient recruitment in our study. Our approach has the potential to help researchers speed up patient identification and recruitment and move toward more efficient clinical trials.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it for intellectual content, and all authors approved the final version to be published. Dr. Tianrun Cai had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Tianrun Cai, Tianxi Cai, Liao.

Acquisition of data. Fiona Cai, Dahal, Cremone, Lam, Golnik, Seyok, Tianrun Cai.

Analysis and interpretation of data. Fiona Cai, Hong, Tianrun Cai, Liao, Tianxi Cai.

REFERENCES

1. Haidich A-B, Ioannidis JP. Patterns of patient enrollment in randomized controlled trials. J Clin Epidemiol 2001;54:877–83.
2. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ 2003;22: 151–85.
3. Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: a systematic review. Health Policy 2011;100:4–17.
4. Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. Nat Rev Drug Discov 2004;3:417–29.
5. Mc Cord KA, Ewald H, Ladanie A, Briel M, Speich B, Bucher HC, et al. Current use and costs of electronic health records for clinical trial research: a descriptive study. CMAJ Open 2019;7:E23–32.
6. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. Effort required in eligibility screening for clinical trials. J Oncol Pract 2012;8:365–70.
7. Wynn L, Miller S, Faughnan L, Luo Z, Debenham E, Adix L, et al. Recruitment of infants with sickle cell anemia to a phase III trial: data from the BABY HUG study. Contemp Clin Trials 2010;31:558–63.
8. Ibrahim GM, Chung C, Bernstein M. Competing for patients: an ethical framework for recruiting patients with brain tumors into clinical trials. J Neurooncol 2011;104:623–7.
9. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. Clin Trials 2015;12:77–83.
10. Beauharnais CC, Larkin ME, Zai AH, Boykin EC, Luttrell J, Wexler DJ. Efficacy and cost-effectiveness of an automated screening algorithm in an inpatient clinical trial. Clin Trials 2012;9:98–203.
11. Penberthy L, Brown R, Puma F, Dahman B. Automated matching software for clinical trials eligibility: Measuring efficiency and flexibility. Contemp Clin Trials 2010;31:207–17.
12. Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak 2015;15:28.
13. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015;22:166–78.
14. Köpcke F, Lubgan D, Fietkau R, Scholler A, Nau C, Stürzl M, et al. Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. BMC Med Inform Decis Mak 2013;13:134.
15. Miotto R, Weng C. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. J Am Med Inform Assoc 2015;22:e141–e50.
16. Butler A, Wei W, Yuan C, Kang T, Si Y, Weng C. The data gap in the ehr for clinical research eligibility screening. AMIA Jt Summits Transl Sci Proc 2018;2017:320–9.
17. Brigham and Women's Hospital, sponsor. Lipids, Inflammation and Cardiovascular Risk in Rheumatoid Arthritis (LiiRA). Clinicaltrials.gov identifier: NCT02714881; 2016.
18. Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. J Am Med Inform Assoc 2017;24:339–44.

19. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32:D267–70.
20. Yu S, Cai T, Cai T. NILE: Fast natural language processing for electronic health records. 2013. URL: https://arxiv.org/abs/1311.6063.
21. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc 2019;14:3426–44.
22. Zhang C, Ma Y, editors. Ensemble machine learning: methods and applications. New York, NY: Springer; 2012. 3426–44.
23. R core team. R: a language and environment for statistical computing. URL: https://www.R-project.org/.
24. Liaw A, Wiener M. Classification and regression by randomforest. R News 2002;2:18–22.
25. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
26. Lu F, Petkova E. A comparative study of variable selection methods in the context of developing psychiatric screening instruments. Stat Med 2014;33:401–21.
27. Weng C, Batres C, Borda T, Weiskopf NG, Wilcox AB, Bigger JT, et al. A real-time screening alert improves patient recruitment efficiency. AMIA Annu Symp Proc 2011;2011:1489–98.