Improving the Efficiency of Clinical Trial Recruitment using Electronic Health Record Data, Natural Language Processing, and Machine Learning

Tianrun Cai, Fiona Cai, Kumar P. Dahal, Chuan Hong, Katherine P. Liao

Abstract:

Efficiently identifying eligible patients is an important component of a successful clinical trial. Using billing codes from electronic health record data to screen for potential patients leads lots of unnecessary patients for chart review. Incorporating billing codes and data extracted from notes using natural language processing to build machine learning algorithm for patient screen could significantly improve the efficiency for identifying eligible patients for clinical trials.

Introduction:

Efficiently identifying eligible patients is an important component of a successful clinical trial. Billing codes from electronic health record (EHR) data are commonly used to first screen for potential patients, followed by labor-intensive chart review to identify the eligible patients by trial criteria. The objective of this study was to test whether a machine learning screening algorithm (ML-screen) incorporating ICD codes and data extracted from notes using natural language processing (NLP), could improve the efficiency for identifying eligible patients for an ongoing clinical trial.

Methods:

We studied EHR data used for a clinical recruitment study of rheumatoid arthritis (RA) and cardiovascular disease recruiting from a tertiary care center (TCC) and a community hospital (CH). The target population were RA patients, age >35, about to initiate a tumor necrosis factor inhibitor, and not on a statin. Prior to this study all patients with ≥1 RA ICD codes (RA_{ICD}) and age>35 years were selected for chart review. The CH and TCC data sets were both manually reviewed as gold standard labels including 642 and 2387 patients, respectively.

All notes were processed with NLP to obtain the number of mentions for the concept of RA and inflammatory arthritis. Three groups of features were considered for the ML-screen (**Table 1**): (1) inclusion criteria features, e.g. RA_{ICD}; (2) exclusion criteria features, e.g. # of electronic prescriptions for a statin; (3) the total # ICD codes as a proxy for healthcare utilization. For the ML-screen we considered features within a 2-year timeframe prior to the chart review as well as all years prior.

The ML-screen combined two ML methods, random forest (RF) and penalized logistic regression. The goal for the ML-screen was to reduce the number of patients requiring chart review without excluding potentially eligible patients. The ML-screen was compared to rule-based approaches using RA_{ICD}≥1, RA_{ICD}≥2, and RA_{ICD} ≥1+exclusion criteria features. To test whether the ML-screen can be successfully ported to other institutions, we trained at TCC and applied at CH, and vice versa.

Results:
The current method reviewing all charts with $RA_{ICD} \geq 1$ yielded 346 (14.5%) eligible patients out of 2387 at TCC, and 74 (11.5%) out of 642 at CH. Applying the ML-screen would result in reviewing 37.9% less ineligible patients in TCC and 45.4% less in CH, compared to $RA_{ICD} \geq 1$, without screening out potentially eligible patients (**Table 2**). In contrast, $RA_{ICD} \geq 2$ can keep sensitivity 0.93 and 0.98, but only reduce 11.3% and 2.7% of patients for chart review at CH and TCC respectively. The $RA_{ICD} \geq 1$+exclusion yielded a larger reduction of ineligible patients for review, 71.8% and 71.1%, however excluded approximately 27% and 22% of eligible patients from TCC and CH respectively. The ML-screen had good performance when trained on one institution and tested on the other (**Table 3**).

Conclusion:
The ML-screen incorporating EHR and NLP data can increase the efficiency of clinical trial recruitment by reducing the number of patients requiring chart review; importantly, this approach did not screen out eligible patients. Moreover, the ML-screen can be trained at one institution and applied at another for multi-center clinical trials.

Table 1. Features used in the ML-screen for clinical trial recruitment.

Category	Feature	Description
Inclusionary features	**RA_{ICD}**	# RA ICD codes
	RA_{NLP}	# mentions for the concept of RA in the narrative notes
	IA_{NLP}	# of mentions for the concept of inflammatory arthritis in the narrative notes
	$RA_{ICD+NLP}$	the sum of RA_{ICD} and RA_{NLP}
Exclusionary features	**JRA_{ICD}**	ICD codes for juvenile rheumatoid arthritis
	SLE_{ICD}	ICD codes for systematic lupus erythematosus
	PsA_{ICD}	ICD codes for psoriatic arthritis
	$Melanoma_{ICD}$	ICD codes for melanoma
	$bDMARD_{COD}$	electronic prescriptions for biologic disease modifying anti-rheumatic drugs
	$Statin_{COD}$	electronic prescriptions for statin
Other	**HU**	Health care utilization, total # of ICD codes

Table 2. Comparison of performance between a screen developed using machine learning vs ICD only screens

	ML-screen		$RA_{ICD} \geq 2$		$RA_{ICD} \geq 1$ & Exclusion		$RA_{ICD} \geq 1$ (REF)	
Institution	**TCC**	**CH**	**TCC**	**CH**	**TCC**	**CH**	**TCC**	**CH**
Sensitivity	0.98	1	0.98	0.93	0.73	0.78	1	1
PPV	0.22	0.29	0.15	0.17	0.3	0.36	0.15	0.16
# patients for review	1606	258	2322	569	828	222	2387	642
% ineligible patients reduced	37.9	45.4	2.8	11.9	71.8	71.1	-	-

Table 3. Comparison of performance for MLS algorithm across institutions

	TC	TC→ CH	CH	CH→ TC
Sensitivity	0.98	0.99	1.0	0.98
Positive predictive value	0.22	0.19	0.29	0.22
# patients for review	1606	355	258	1713
% ineligible patients reduced	37.9	28.7	45.4	32.8