Engage: An Open Source, Automated Clinical Trial Awareness Tool

William R. Haslett, PhD[1], Craig H. Ganoe, MS[1], Rebecca Faill[1], Steven B. Andrews, PhD[1]
[1]Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire

Abstract

Clinical researchers are increasingly focusing on community engagement, but public awareness of research participation opportunities can be improved. While clinicaltrials.gov is a comprehensive source, we have found issues with awareness and use. We created a clinical-trials awareness tool that imports clinicaltrials.gov data, categorizes it using a Bayesian classifier, and makes institutionally relevant data available to user via the Web or email alerts.

Introduction

Community engagement has been defined by the Centers for Disease Control and Prevention as "the process of working collaboratively with groups of people who are affiliated by geographic proximity, special interests, or similar situations with respect to issues affecting their wellbeing"[1]. In the present context, we wish to engage community members as participants in clinical research. Recruitment for clinical trials can be difficult, especially for trials that have eligibility criteria that are met by a small proportion of the population. Current recruitment strategies often rely on traditional marketing techniques such as telemarketing mailing campaigns[2], but there is evidence that online tools can improve enrollment[3].

Institution-specific web sites designed to promote community-level awareness of clinical trial participation opportunities are common, but they may require manual maintenance as new studies enter the recruitment phase and older studies exit the recruitment phase. These sites also face the challenge of describing and classifying studies using lay-friendly language. Static sites that list recruiting clinical trials are also limited in that community members must keep returning to the site in order to stay up-to-date regarding clinical trial participation opportunities.

To address these challenges, we built the Engage application. Engage has three integrated modules: a clinical trial importer that maintains an up-to-date list of an institution's recruiting clinical trials, a text classifier that categorizes clinical trials using lay-friendly terms, and a web application that allows community members to search for clinical trials that match their preferences. It also allows community members to register to receive email notifications when new studies that match their preferences appear. Engage performs all of these functions automatically, importing and classifying clinical trials, updating the web application, and sending email notifications, without human intervention. A schematic depicting the Engage system is shown in the figure.

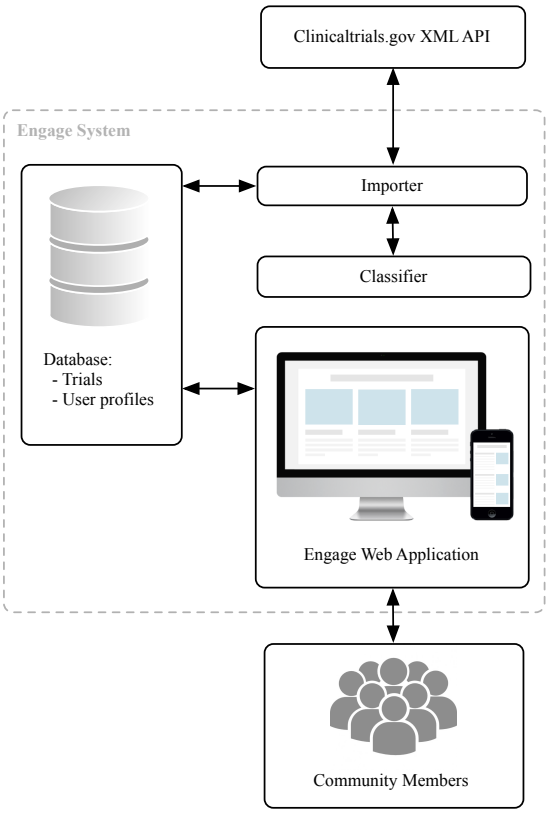

Figure: The Engage System

Methods

Engage uses clinicaltrials.gov as its source for clinical trial information. Clincialtrials.gov provides an XML interface that allows for programmatic access to its database[4]. Using this interface and a parser/importer built in the Python programming language, Engage performs nightly pulls from the clinicaltrials.gov and updates a local database of recruiting trials. During each database update, existing

records are checked for their status as still recruiting or not and new records are checked against existing records to avoid duplicates.

Records from clinicaltrials.gov contain with up to three kinds of classifying fields: a list of conditions, a list of keywords, and a list of MeSH terms[5]. None of these three data sources provide a reliable, lay-friendly categorization for clnicaltrials.gov data. Accordingly, we built a text classifier that ingests relevant attributes for each clinicaltrials.gov record and assigns each record a lay-friendly category.

The first step in building Engage's classifier was developing the list of categories to which trials would be assigned. To do this, three raters independently assigned categories to a random sample of 200 recruiting clinical trials. The raters then met and developed a consensus regarding the category list that would be used. We then used these categories to develop a ground-truth training set for the classifier. The same three raters categorized 636 more trials and again met to achieve a consensus rating for each trial. As we worked through the ground truth data set, we found it necessary to expand our list of categories from 47 to 58. Categories in the final list of 58 included "breast cancer", "mental health", and "skin conditions", for example. Each category in our final list of 58 maps to a MedlinePlus health topic, allowing community members to get information on each category that is intended for the general public.

After developing the category list and creating the ground-truth data set, we created a matrix of bag-of-words models using they Python SciKit Learn library[6]. We tested four different conditions for combinations of trial attributes to ingest and, two model types: Support Vector Machine and Naïve Bayes. We assessed the accuracy of these models using both percent accuracy during ten-fold cross validation and percent accuracy when classifying a validation set of 2,900 studies with a known category.

Finally, we built Engage's web application module. This tool was built using the Ruby on Rails framework. The web application's responsive design works on a variety of device form factors. We built the web application to allow for easy searching of trials by category or keyword. We also included a registry in which community members can enter their preferences for clinical trial participation and receive email notifications when matching recruiting trials appear.

Results

The clinical trials importer works and maintains a list of currently recruiting clinical trials in the Dartmouth-Hitchcock network. In this network, there are currently 212 recruiting trials across 46 categories.

The clinical trials classifier is a Bayesian Classifier that uses the trial title, summary, conditions, keywords and MeSH terms as its input. This model has an accuracy of 93.8% in 10-fold cross validation, and 74.5% against the validation data set. Due to the iterative nature of our process for developing and assigning trial categories, we were not able to compute a multiple-rater Kappa statistic to assess inter-rater reliability.

The Engage web application is currently in final development and will be available at https://engage.dartmouth.edu pending final development iterations and user acceptance testing.

Discussion

We plan to deploy Engage publicly in 2017, and we are already working on future enhancements. Most importantly, we plan to add tools for investigators in order to more directly assist them in their recruitment efforts. Investigators will be able to edit information for existing studies, manually add studies that are not in the database, and view registry members who match their trials' eligibility criteria.

Engage requires minimal maintenance. At present the only manual maintenance needed is human confirmation of the classifier's categorization, due to the classifier's 74.5% accuracy. We have streamlined this confirmation process, with application administrators getting email notification when new trials appear, prompting them to check a box indicating that a categorization is correct, or allowing them to easily select another category if it is not.

Public awareness of the Engage tool itself is key for its ability to raise awareness about clinical trials. To promote awareness about Engage, we plan to link to it from the Dartmouth-Hitchcock landing page. We also plan to install one or more kiosks within the hospital that are running Engage and have a touch screen interface. We will monitor registry participation over time, and continue to work with both professional and community-based stakeholders to enhance Engage's usefulness.

References

1. Centers for Disease Control and Prevention. Principles of community engagement. CDCATSDR Committee on Community Engagement. 1997;

2. Treweek S, Lockhart P, Pitkethly M, Cook JA, Kjeldstrøm M, Johansen M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. BMJ Open. 2013;3(2):e002360.

3. Rimel BJ, Lester J, Sabacan L, Park D, Bresee C, Dang C, et al. A novel clinical trial recruitment strategy for women's cancer. Gynecol Oncol. 2015;138(2):445–8.

4. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov Results Database — Update and Key Issues. N Engl J Med. 2011 Mar 3;364(9):852–60.

5. Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. Jama. 1994;271(14):1103–8.

6. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.