Towards Dynamic and Interactive Retrieval of Clinical Trials
Using Common Eligibility Features

Riccardo Miotto¹, Ph.D. and Chunhua Weng¹², Ph.D.
1. Department of Biomedical Informatics, Columbia University, New York, NY
2. The Irving Institute for Clinical and Translational Research, New York, NY

Abstract
Information overload is a barrier for patients identifying relevant clinical trials online. We present a framework to
dynamically refine clinical trial search results towards a manually reviewable number of documents. At this aim, a
set of frequent features are extracted automatically offline from the eligibility criteria of a sample collection of
clinical trials. These features are then presented dynamically to the users based on feature co-occurrences and user
actions for reducing the number of resulting documents. Experiments on ClinicalTrials.gov showed that using an
average of four features, it is feasible to reduce the results of a simple search from over a thousand to ten.

Introduction
Clinical trials generate highly relevant medical evidence for effective disease treatments. Nevertheless, most clinical
trial search engines return a large number of unranked results and it is often difficult for patients to identify the
clinical trials wherein they may be eligible. In addition, also the formulation of specific and more effective queries
might not be easy for average patients. Different solutions have been proposed to overcome this problem (e.g.,
AskDory). However, most of them heavily rely on manual annotation of free-text eligibility criteria, which is time-
consuming, error-prone, and not scalable. The goal of this study is to provide a scalable solution to generate
manageable clinical trial search results.

Methods
Our method consists of two steps: (1) offline feature extraction and trial semantic indexing, and (2) online feature-
based trial search. Feature extraction automatically mines a set of multi-word meaningful patterns (e.g., “breast
cancer”, “type 1 diabetes”, “active malignancy”) in the free-text eligibility criteria that frequently appear in a
collection of clinical trials. These features are eventually used to semantically index the trials. The feature-based
search aims at refining the results of a simple query (e.g., “breast cancer”), which is likely to return thousands of
trials. After this simple search, our method begins with proposing to the user a subset of the features previously
extracted, which appears frequently in the resulting set. The user can then select one of them to filter the resulting
documents. At every user selection, the system automatically retains only the documents containing the chosen
feature and suggests a new set of features. Every subsequent set of suggested features is related to the previous
selection. In fact, the algorithm looks for the features that could reduce the most number of documents given the
current status (i.e., the features already chosen). This mechanism is achieved using a set of association rules mined
offline from the trial index, which measure the co-occurrence between the features. The filtering process repeats
until a desired number of trials are returned.

Results
We carried out an evaluation using the ClinicalTrials.gov collection as of August 2012 (about 130,000 trials). First,
we proved that the frequent feature set stabilizes if extracted using a subset covering at least 40% of all trials. This
ensures scalability to the approach since the majority of features will remain even if the collection increases. We
considered as frequent a feature appearing at least in 1% of the trials in the subset. Second, 5 independent human
raters evaluated the features extracted from 50,000 randomly selected trials, showing that 87% of them represent
meaningful concepts. The 192 approved features allowed us to index 96% of the trials in the entire collection.
Lastly, we measured how the feature-based filter mechanism affects the search process. Given 50 study topics with
more that 1,000 search results on ClinicalTrials.gov and randomized user selections, we demonstrated that it is
possible to narrow down the results of a simple search to ten trials using an average of four features.

Discussions
A user-based evaluation will follow in order to measure the user satisfaction of the search experience as well as to
compare our method with other currently available clinical trial search engines.

Acknowledgments
The project was supported by grants R01LM009886 and R01LM010815 from National Library of Medicine, R01HS019853 from
Agency for Healthcare Research and Quality, and UL1 TR000040 from National Center of Advancing Translational Sciences.