



DiscoverCT

*Discover more
about clinical trials*

2015 Final Project Report
UC Berkeley School of Information

Team Members
Colin Gerber & Jason Ost

Advisor
Professor Marti Hearst

Table of Contents

PROJECT GOAL	2
BACKGROUND	2
PRIOR RESEARCH USING THE REGISTRY	3
DATA	4
CLINICAL TRIALS TRANSFORMATION INITIATIVE	4
FREEBASE	4
NLM MEDICAL THESAURUS.....	5
PUBMED	5
METHODS	6
OFFLINE METHODS	6
<i>Institution deduplication</i>	6
<i>Linking publications to trials</i>	7
<i>Data quality ratings</i>	8
<i>MeSH term suggestions</i>	9
ONLINE METHODS.....	11
<i>Active learning interface</i>	11
<i>MeSH recommendation engine</i>	12
USER RESEARCH	12
<i>Initial inquiry</i>	12
<i>Interface testing</i>	13
RESULTS	14
INTERFACE SCREENSHOTS	14
INSTITUTION DEDUPLICATION AND PUBLICATION LINKAGE.....	20
MESH TERM SUGGESTIONS	21
USER TESTING AND NAVIGATION PATHS	22
NEXT STEPS	25
IMPROVING ACTIVE LEARNING	25
INTERFACE IMPROVEMENTS	26
EXTERNAL INTEGRATION	26
REFERENCES	27
APPENDIX	30
APPENDIX A: ENTITY-RELATIONSHIP DIAGRAM OF BACKEND DATABASE.....	31
APPENDIX B: USER TESTING SURVEY.....	32

Project goal

The ClinicalTrials.gov registry is a public resource that contains valuable information about clinical research, yet most of the data lacks structure or easy accessibility. We aim to create an interface that allows researchers, trial sponsors, and patients and their physicians to easily understand the state of clinical research on particular conditions or at a specific institution. Ideally this interface would also provide tools that permit users to actively improve the quality of data in the registry so that it is more accessible to the general public.

Background

Clinical trials are an important part of medical innovation and our public health infrastructure. These trials seek to evaluate the efficacy and safety of drugs, devices, procedures, behavioral changes, and other interventions, typically in comparison to some known treatment or a placebo—i.e., no treatment or intervention (National Institutes of Health 2014).

Trials are usually sponsored by pharmaceutical companies, academic medical centers, research foundations, and government agencies, and these funders spend a significant amount of money conducting trials each year. In 2014, for instance, the National Institutes of Health (NIH) granted \$3.2B for research, and the pharmaceutical and medical device industries spent a further \$32.3B on Phase I-IV clinical trials (National Institutes of Health 2015; PhRMA 2014).

The National Library of Medicine (NLM) at NIH has maintained a publicly accessible registry of observational and interventional studies since 1997, when the Food and Drug Administration Modernization Act (FDAMA) mandated registration of all Phase II-IV clinical trials. This database, hosted at ClinicalTrials.gov, has become more widely used since 2005, when all major scientific journals instituted a requirement that a trial be registered prior to first patient being enrolled in order to publish results (ICMJE 2015).

In fact, a number of countries host clinical trials registries (World Health Organization 2015), but the NIH registry is far larger than the others due to its early acceptance by journal editors as evidence of trial registration. This makes the registry the main source of clinical trial information worldwide.

Currently there are over 150,000 interventional studies registered at ClinicalTrials.gov, which is structured to collect detailed information regarding study design, institutional

characteristics, and study results. There are an additional 35,000 observational studies in the registry, but for the purposes of this project, we have only focused on the interventional studies, better known as clinical trials.

Prior research using the registry

Although ClinicalTrials.gov is a publicly available source of important clinical trials information, research on the registry has been relatively sparse. Zarin et al. (2005) performed the first general review of the trials in the registry, and their study happened to coincide with the journal editors' announcement of their trial registration requirement. The same group followed up with another review of the registry six years later (Zarin et al. 2011), which appears to have sparked broader interest in the use and implications of a public clinical trials registry.

Recent published research on the registry can generally be divided into two categories: surveys of trials in the registry (Califf 2012, Bell and Tudur Smith 2014), sometimes with a particular focus on a medical specialty or outcome measure (Inrig et al. 2014, Vodicka et al. 2015); and investigations into the results reported in the registry (Kuehn 2012, Saito and Gill 2014), with an especial focus on the correspondence between submitted and published results (Riveros et al. 2013, Hartung et al. 2014). The intense focus on trial results submitted to the registry is driven primarily by ethical and public policy concerns about the transparency of publicly funded research (Anderson et al. 2015, Enserink 2015): despite reporting requirements covering a large share of trials, less than 10% of interventional trials in the ClinicalTrials.gov database have any results reported.

In contrast to the growing body of research on the (lack of) submission of trial results, very little research has investigated the quality of data in the registry (Guharoy 2014), and there are virtually no published statistics about how the general public is using this resource. Since trial registration requirements have imposed significant compliance burdens on trial investigators (Getz et al. 2011, O'Reilly et al. 2015), many investigators simply meet the minimum requirement of registration without providing detailed protocol description, site location data, eligibility criteria, or other important information that could be used to inform the public of their activities. Moreover, the registry lacks a high degree of standardization across trials, complicating issues of information organization and retrieval.

Data

Clinical Trials Transformation Initiative

The Clinical Trials Transformation Initiative (CTTI), hosted at Duke University, has endeavored to convert the ClinicalTrials.gov registry data into a relational database format, which they publish approximately twice each year in Oracle, SAS, and pipe-separated plain text formats (Tasneem et al. 2012, CTTI 2015). The latest version, published in September 2014, contains information for over 160,000 clinical studies, including around 132,000 interventional studies.

We did not have access to Oracle or SAS, so opted to load the plain text version of the CTTI database into a MySQL instance. Prior to loading the data, we performed some preprocessing steps:

- Removal of line breaks and excessive whitespace in multi-line fields such as trial protocol description and eligibility criteria
- Removal of suspicious trial enrollment numbers, e.g. “9999999”
- Update of erroneous Medical Subject Heading (MeSH) terms so they aligned with the official MeSH vocabulary (National Library of Medicine 2014)

All the steps of our data preprocessing and loading process exist in a set of Python and SQL scripts that can be easily updated to load future releases of the CTTI database. Appendix A provides an entity relationship diagram of our website’s backend database, indicating its connection to the CTTI database.

Freebase

Freebase, formerly known as Metaweb, is an open-source repository of people, places, and things (Freebase 2011). Each entity is described using a series of resource description framework (RDF) triples, which are also used to connect the entity with other people, places, and/or things. For instance, the major public hospital in San Francisco “is named” San Francisco General Hospital and “was founded” in 1850; it also is “part of” the UCSF Helen Diller Family Comprehensive Cancer Center and “employs” a number of personnel.

We used the Freebase API to identify canonical institution entities during our sponsor and facility deduplication process (see below). Freebase was also the source of our institution descriptions, locations, and images. It is worth noting that Google owns

Freebase and has decided to shut down the service in favor of their Knowledge Graph, which relies on much of the same underlying data (Freebase 2015).

NLM medical thesaurus

Most clinical trials in the registry are tagged with Medical Subject Heading (MeSH) terms, a controlled vocabulary hierarchy developed by the NLM to standardize descriptions of biomedical text and concepts. This is helpful for understanding the relationships between trials, researchers, and institutions, but it is often not ideal for a search implementation because the average user does not colloquially use MeSH terms. For example, trials studying cancer treatments often use the MeSH term “neoplasm” (the medical term for tumor), but most laypeople have never heard of a neoplasm.

We needed to link MeSH terms with their common names in order to improve the accessibility of the registry data. Fortunately, the NLM has developed a health topic resource called MedlinePlus, which offers commonly used names for many frequently diagnosed conditions (National Library of Medicine 2015). We downloaded the most recent version of the MedlinePlus health topic XML and parsed it in order to find useful synonyms for the MeSH terms found in the registry.

PubMed

PubMed is a search engine for biomedical literature, providing access to more than 24 million citations from MEDLINE, life science journals, and online books (National Library of Medicine 2005). This service allows you to search for publications based on a wide variety of fields, including keyword, MeSH term, publication date, and author, and it is the primary resource used by researchers and clinicians when searching for publications.

Only 21% of completed trials in the ClinicalTrials.gov registry have been linked with any publication, so it is quite likely there are missing links between trials and subsequent publications. The PubMed API provides programmatic access to the search engine and document description information (Sayers 2010), allowing us to comb through a vast number of potentially relevant documents in order to identify publications that are likely to be related to a given trial. See section below about linking publications to trials for more information.

Methods

Offline methods

Institution deduplication

The majority of information submitted to ClinicalTrials.gov is unstructured text lacking any standardization, which has resulted in a large number of mistyped and otherwise duplicated institution names. For example, a human can tell that “Johns Hopkins University”, “John’s Hopkins University” and “John Hopkins” all refer to the same institution, but a search for “Johns Hopkins University” will only match trials listed with institution names that exactly match the search term which would only be the first name.

Moreover, various departments within an institution often identified themselves as such, usually with varying acronyms, abbreviations, and punctuation marks. Thus, data related to Johns Hopkins University is associated with several different keys instead of a single canonical key: the registry contains 273 unique strings that we associated with Johns Hopkins University, and these exist at 458 different locations—i.e., combinations of city, state, and ZIP Code.

In order to make the data as accessible as possible to those interested in the research activity of a particular institution, we sought to make the data related to one institution queryable under a single canonical string representing that institution, so that when a user searches for “Johns Hopkins University”, they retrieve all the results they were seeking.

To achieve this goal we first retrieved all the facility names and locations in the database and used the Python dedupe package to identify likely duplicates based on the name and location information for each record. This package employs an active learning approach to link duplicate records using sophisticated predicate blocking techniques (Bilenko 2006, Gregg et al. 2015).

There is some judgment left to the user of this algorithm about how aggressively to merge duplicates. For example, the thresholds we set ultimately merged facilities like Johns Hopkins University and Johns Hopkins Medical School into a single entity, but the Johns Hopkins Bloomberg School of Public Health remained its own distinct entity. It is debatable whether this should in fact be three (or more) distinct entities, or perhaps all a single entity. A visual inspection of dozens of trials indicated that, in this case,

many trials at “Johns Hopkins University” were actually located at the medical school or the affiliated Bayview Medical Center, while those attributed to the Bloomberg School of Public Health were nearly always in the same location.

After linking facility names and locations that referred to the same institution, we matched these clusters to trial sponsor names, which tend to be more standardized, in order to identify all the trials associated with an institution. To aid this process we used the aforementioned Freebase API to find the matching business, hospital, university, or other institution that could serve as the institution's canonical representation.

Linking publications to trials

It is important to have publications linked with clinical trials and institutions because peer reviewed publications are one of the best indicators of successful trials, investigators, and institutions. If, for example, an institution had a very low number of publications compared to the number of trials they had run related to a certain disease, this would be an indicator that they have not had very many successful trials in that disease category.

While it is possible for investigators to associate publications to the trials they have registered with ClinicalTrials.gov, only around 29,000 studies in the database have any linked publications. In order to find additional relevant publications, we retrieved a list of all the investigators in the registry and stripped the names of any honorifics (e.g., Dr., Prof.). Once we had a cleaned list of investigator names, we used the PubMed API to query data about all the publications each investigator had published.

Because there were over 90,000 unique investigators in the database and we had to make two API calls per investigator (one to get the publication IDs and one to get data about the publications), we ended up having to making around 180,000 API calls. PubMed has several limitations on the use of their API when making this volume of queries, including only making three calls per second and only making calls between 9 pm and 5 am. As a result, it took almost a week to finish making all the API calls. In addition, because there were so many investigators and potential publications, we ended up with ~45GB of XML data to sort through.

Once we had all the data, we processed it, separating the articles that had direct links to trials and those that didn't. (NB: an author is able to include a ClinicalTrials.gov identifier in their PubMed description, which many investigators did without including the publication in the trial registry.) We then dropped any articles published prior to the investigator's earliest trial, as well as articles that had no matching MeSH terms with

any of the trials in which the investigator had been involved. With this new set of investigator publication data, we were able to link publications to trials based on a weighting system that looked at number of matching MeSH terms, whether or not the matching MeSH term was a primary term for the publication (which was a marker in the publication xml), and the amount of time between the publication date and trial date.

With these weights we were able to split up the potentially linked publications into three groups—likely, probably, possibly connected to the trial—to give the user an idea of how confident the system is that the publication is correctly matched with the trial.

Data quality ratings

One goal of our project is to improve the data quality of the clinical trials registry, and one of the main ways this can be accomplished is by improving the quality of data entered into the registry in the first place. To this end we developed a rating system for the quality of data entered for each trial, in an effort to expose the areas where investigators and institutions need to improve their submissions.

We combined a variety of measures to come up with this ranking, including the quality of the dates, MeSH terms, site description, general completeness, trial protocol description, and eligibility criteria. We aim for transparency by publishing our criteria on the project website. The positive and negative factors for each measure include:

- Research facility data
 - Positive factors: valid city and country, ZIP Code (if in U.S.), a meaningful site name, a valid investigator name
 - Negative factors: a generic site name (e.g., “Investigation Site #24”), a recruiting status that is inconsistent with the trial’s overall recruiting status
- Protocol description
 - Positive factors: approximately 500 words or more
 - Negative factors: boilerplate text (i.e., appears quite often across trials), a high number of very frequent terms like “trial”, “study”, or “intervention”
- Eligibility criteria description
 - Positive factors: sections clearly labeled “inclusion criteria” and “exclusion criteria”
 - Negative factors: more than 30 eligibility criteria
- MeSH classification
 - Positive factors: four or more condition terms, two or more intervention terms

- Negative factors: no condition or intervention terms
- Dates
 - Positive factors: presence of start and completion dates, presence of “actual” or “anticipated” completion date type
 - Negative factors: obviously incorrect dates (e.g., “December 2099”), an “anticipated” completion date that occurs in the past

Each trial is rated on these measures, and ratings are also aggregated across trials so that each institution has an overall data quality rating. Our hope is that these ratings will expose data quality problems for researchers and institutions, and this awareness will lead them to improve the quality of their trial registrations.

MeSH term suggestions

Because MeSH is a controlled vocabulary, trials can be more easily retrieved and compared when they are tagged with all relevant MeSH terms. More than 23,000 trials, or 14% of those in the CTTI database, have no MeSH condition terms associated with them, and another 51,000 trials (31% of the database) are tagged with just one condition term. This limits the usability of this data set since many searches for a particular condition will fail to retrieve all relevant results.

To improve this coverage we used machine learning algorithms to generate suggestions of MeSH terms that apply to each trial based on the portion of trials in the database that are well-described using MeSH terms. Specifically, we used three different methods—a maximum entropy classifier, a series of logistic regression classifiers, and a K-Nearest Neighbors (KNN) model—in order to generate suggestions that are displayed on an individual trial's page.

Maximum Entropy classifier

The maximum entropy classifier predicts the single most likely MeSH condition term with which a trial should be tagged. Although the CTTI database includes over 3,300 distinct MeSH condition terms, we limited this classifier to around 1,800 MeSH condition terms that were applied to ten or more trials in the database. This ultimately excluded only about 10,000 of the 140,000 trials tagged with one or more MeSH condition terms.

We constructed a tf-idf matrix, where each “document” represented a single MeSH term and consisted of all the protocol description text for each trial tagged with that term. Stopwords and punctuation were removed, and all words were lowercased prior to the

tf-idf matrix creation. The classifier was trained using this matrix and the predictions were generated for each new trial using similar transformations of its description text.

Logistic regression classifiers

The series of logistic regression classifiers predict the likelihood an individual trial is related to the hypernym concepts at the second level of the MeSH hierarchy. There are around 290 condition-related concepts at this level of the hierarchy, and this level is typically a bridge between the most generic concepts (e.g., cancer, cardiopulmonary disease) and more specific illnesses (e.g., breast cancer, high blood pressure). Few trials are actually tagged with terms residing at this concept level, so these predictions would serve mainly as a starting point for a professional reviewer.

As in the maximum entropy classifier, each hypernym was converted into a “document” containing all the description text of trials tagged with terms falling under that hypernym. Stop words, punctuation, tokens shorter than three characters, and any wholly numeric tokens were removed, and all text was lowercased; the feature vector was simply a frequency of each of the remaining tokens, normalized by vector length. Then the hypernym’s feature vector was compared to a similar feature vector for all trials that didn’t include the hypernym in a logistic regression in order to generate coefficients for the terms that most predicted a relationship with that particular hypernym.

K Nearest Neighbors classifier

The final model is a straightforward KNN model, following from work by Trieschnigg et al. (2009), who found that a KNN-based algorithm clearly outperformed other classifiers when suggesting MeSH terms for article abstracts submitted to PubMed. For this model, a single tf-idf matrix was generated for all trials tagged with at least one MeSH term. Again, the text used for each trial was its description text stripped of stop words and punctuation, and with the text lowercased. Unlabeled trials go through a similar transformation, and the algorithm identify the 10 nearest neighbors as defined by euclidean distance between the documents’ tf-idf vectors.

The (manually-assigned) MeSH terms associated with those 10 nearest neighbors are aggregated and weighted according to how many of the neighbors are tagged with the term and how far the neighbors are from the unlabeled document. The output of this algorithm is then a list of all MeSH terms associated with the 10 nearest neighbors of the unlabeled trial, ranked by projected relevance.

Online methods

Active learning interface

Eligibility criteria are a critical factor in determining whether an individual can participate in a specific trial, yet these criteria are submitted to the registry as completely unstructured blocks of text, except for basic age and gender criteria. This makes it difficult for patients or their physicians to efficiently determine which trials they may be able to join.

To ameliorate this problem, we created an interactive "active learning" process that enables users to select a term from a trial's eligibility criteria section and build out a group of terms that encompass a eligibility *concept*. For example, the concept of "birth control" might include terms like birth control, contraceptive, condom, IUD, etc. The interface for the tool allows users to accept and exclude different terms while creating the concept.

In the background, this interface is using two different algorithms to provide suggested terms that a user can accept or reject for the concept. The first relies on the word2vec Python package, which uses a multi-dimensional vector representation of words and phrases to efficiently estimate similarities across huge text corpora (Mikolov et al. 2013a & 2013b). When given a seed term, a model trained on all trials' eligibility criteria returns terms that often appear in similar contexts. For efficiency and robustness, the model only includes around 20,000 frequently appearing terms, so if a seed term is not in the model the active learning process moves on to the predictor step.

Whereas a concept is comprised of terms, predictors are words and phrases that often appear with the terms in that concept, but are not directly related to the concept itself. For example, terms related to the concept "birth control" may often appear with "method of" or "effective use"; they are unlikely to appear with "disease progression" or "unknown cause". The predictor step finds noun phrases and other word chunks that frequently appear in the same sentence as a concept terms, and suggests these to a user for acceptance or rejection.

These predictors, in turn, inform the next round of term suggestion. The system suggests noun phrases that frequently appear with the predictors, and a user can accept or reject these, which informs the next predictor step. A user can go back and forth between these term and predictor steps, accepting or rejecting terms and predictors, in order to develop a concept by identifying as many associated terms as possible.

Once a concept is created it is saved in a staging table where an administrator can review and approve the associated terms. After being approved, the concept is associated with all trials to which it applies, and then will be accessible for filtering trials on the site. This enables users to help structure the free text of the trial eligibility criteria data in order to increase retrieval.

MeSH recommendation engine

While our primary goal is to improve access to the data already in the ClinicalTrials.gov registry, another way to improve the data in the database is to help improve it before it is even entered. The controlled MeSH vocabulary improves information retrieval because trials can be discovered using a standard set of descriptors rather than the trial investigator's specific, and possibly idiosyncratic, terminology. To support researchers who seek to describe their registered trial using MeSH terms, we have an online MeSH recommendation engine that accepts any text (currently limited to 8,000 characters) and suggests a set of relevant MeSH condition terms.

This tool uses the KNN technique described for the offline process of suggesting MeSH condition terms: a user's text is compared to each trial's protocol description, and the MeSH terms associated with the 10 most similar trials are returned as suggestions. Because a brute-force approach of multiple pairwise comparisons is extremely slow, the model, in this case, uses a latent semantic indexing approach to dramatically reduce the dimensionality of the text, allowing extremely fast comparisons (Rehurek and Sojka 2010).

User research

Initial inquiry

Prior to developing our interface, we performed a preliminary set of interviews with a set of clinical studies experts, including:

- Winston Chiong, MD, PhD, an Assistant Professor in Neurology at the University of California, San Francisco. Professor Chiong performs clinical research and had recently gone through the registration process at ClinicalTrials.gov, but is otherwise unfamiliar with the database or its potential uses.
- Jennifer Ahern, PhD, MPH, an Associate Professor in Epidemiology at the University of California, Berkeley's School of Public Health. Professor Ahern is an

expert in observational studies, rather than clinical trials, and was also mostly unfamiliar with the ClinicalTrials.gov registry.

- Jack Colford, MD, PhD, MPH, at the University of California, Berkeley's School of Public Health. Professor Colford is an expert in clinical trial design and is very familiar with clinical trials registries, including ClinicalTrials.gov.

Although our experts had varying levels of experience with the ClinicalTrials.gov registry, each of them identified exciting potential new uses of the database during the course of our conversations. Professor Chiong, for instance, was interested in the possibility of using the registry's results data to conduct a systematic review, while Professor Ahern mentioned the possibility of using it to identify and study off-label uses of interventions. Professor Colford felt the registry could be used to verify that published results were produced using the protocol set forth at the beginning of the trial.

Based on these interviews, we decided to make the researcher our primary persona, since they have the technical expertise to make the trials more understandable. This informed our goals to develop tools that would assist experts in improving their data, rather than simply providing a better interface for the existing registry. The patient and their physician remained an important part of our public access mission, but became secondary personas that would benefit from the data quality improvements generated by trial investigators and other researchers.

Interface testing

We performed user testing via both in person interviews and an online survey (see Appendix B). In both settings we provided a list of tasks for the users to accomplish using our DiscoverCT.org website. We created the tasks in a way that they would cover all of the major interactions we envisioned potential users performing on the site. One thing to note is that there are two primary groups of potential users—researchers and patients—and the tasks covered all the interactions for both of the user groups.

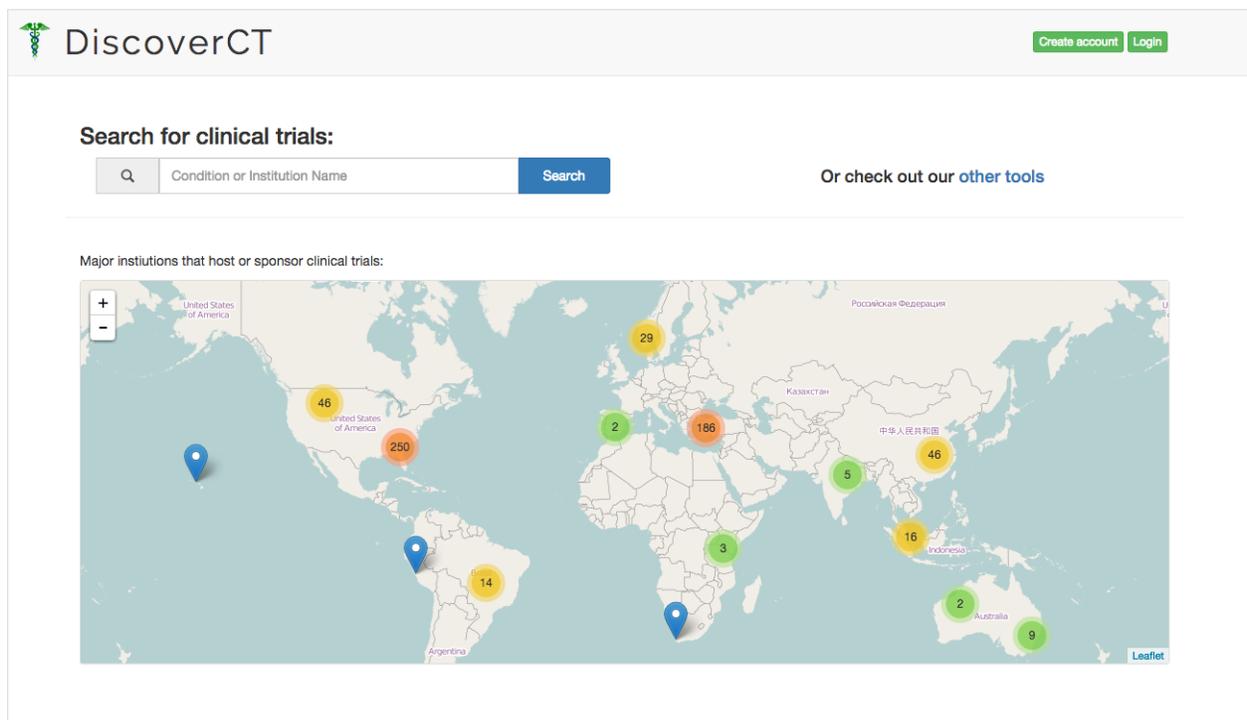
The primary tasks in the user interviews were searching for a condition or institution (the choice of which was left up to the user), filtering the clinical trial results from the search, exploring the page describing a specific trial, and using the active learning interface to create a new eligibility criteria concept. After all the tasks were complete we also asked the users to freely explore the site and provide any additional feedback they had. During the in-person interviews we also had them explore the MeSH term suggestion tool.

For each task we had additional subtasks (which can be seen in Appendix B) which were aimed at getting feedback on specific features of interest. For example we had a pop-up modal that described the active learning system, and we were very interested in seeing what how helpful the users found that information. We collected open-ended responses of their impression of each subtask, as well as having them rank how clear or effective each step was.

Results

Interface screenshots

The main product for this project is our user interface, available at DiscoverCT.org:



The homepage



Search results for 'hiv'

Conditions (9 results)

- HIV Infections (3962 trials)
- HIV (3887 trials)
- HIV/AIDS (3887 trials)
- HIV Seropositivity (140 trials)
- Hives (90 trials)
- HIV/AIDS and Infections (36 trials)
- HIV-Associated Lipodystrophy Syndrome (19 trials)
- HIV Wasting Syndrome (13 trials)
- HIV Enteropathy (4 trials)

Major institutions (2 results)

- Albert Einstein College of Medicine of Yeshiva University (104 trials)
- HIV Vaccine Trials Network (22 trials)

Other trial sponsors (39 results)

- The HIV Netherlands Australia Thailand Research Collaboration (54 trials)
- CIHR Canadian HIV Trials Network (20 trials)
- HIV Prevention Trials Network (14 trials)

Search results



HIV Infections

Synonyms

HIV Human immunodeficiency virus HIV/AIDS AIDS Acquired Immunodeficiency Syndrome

Summary

HIV, the human immunodeficiency virus, kills or damages cells of the body's immune system. The most advanced stage of infection with HIV is AIDS, which stands for acquired immunodeficiency syndrome.

HIV often spreads through unprotected sex with an infected person. It may also spread by sharing drug needles or through contact with the blood of an infected person.

Women can get HIV more easily during vaginal sex than men can. And if they do get HIV, they have unique problems, including:

- Complications such as repeated vaginal yeast infections, severe pelvic inflammatory disease (PID), and a higher risk of cervical cancer
- Different side effects from the drugs that treat HIV
- The risk of giving HIV to their baby while pregnant or during childbirth

There is no cure, but there are many medicines to fight both HIV infection and the infections and cancers that come with it. People can live with the disease for many years.

Leading Institutions



National Institute of Allergy and Infectious Diseases (NIAID)

1089 trials



Johns Hopkins University
Baltimore, Maryland

286 trials



University of Rochester
Rochester, New York

270 trials

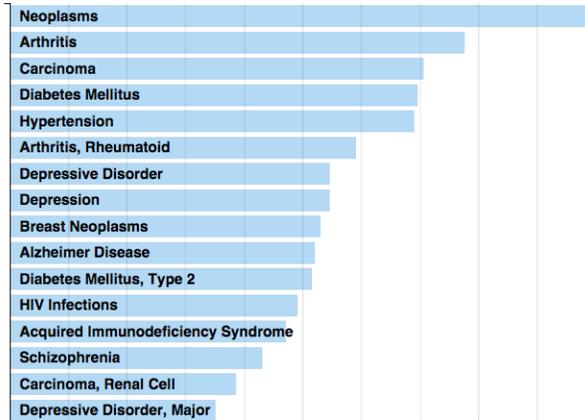
Condition page



Pfizer 3730 trials

New York City, New York

Most Frequently Studied Conditions



Institution Data Quality Ratings ?

Overall data quality	★★★★☆
MeSH classification quality	★★★★☆
Site data quality	★★★★☆
Protocol description quality	★★★★☆
Trial criteria description quality	★★★★★
Date quality	★★★★★

Summary

Pfizer, Inc. is an American multinational pharmaceutical corporation headquartered in New York City, and with its research headquarters in Groton, Connecticut, United States. It is one of the world's largest pharmaceutical companies by revenues. Pfizer develops and produces medicines and vaccines for a wide range of medical disciplines, including immunology, oncology, cardiology, diabetology/endocrinology, and neurology. Pfizer's products include the blockbuster drug Lipitor, used to lower LDL blood cholesterol; Lyrica; Diflucan, an oral antifungal medication; Zithromax, an antibiotic; Viagra; and Celebrex/Celebra, an anti-inflammatory drug. Pfizer was founded by cousins Charles Pfizer and Charles F. Erhart in

Institution page



Trials addressing HIV Infections 566 trials match query (most recent shown first)

Patient view [Go to researcher view](#)

This view only shows trials that are in the recruiting phase or earlier.

Refine these results by providing information about the trial participant, then click **Update results**

Gender: Female Male

Age:

Location

Trials within of

Eligibility Criteria

Yes birth control? No

Inflammation?

Only show trials accepting healthy volunteers

Update results

Gene Therapy After Frontline Chemotherapy in Treating Patients With AIDS-Related Non-Hodgkin Lymphoma

This trial is **recruiting**. It is an **interventional** assessment of a **biological, drug, and/or other** intervention. (No information was provided about the phase of the trial.)

Location(s): City of Hope Medical Center, Duarte, California 91010

Condition(s): Lymphoma; Lymphoma, Non-Hodgkin; Precursor Cell Lymphoblastic Leukemia-Lymphoma; Lymphoma, Large-Cell, Immunoblastic; Lymphoma, Large B-Cell, Diffuse; HIV Infections; and Acquired Immunodeficiency Syndrome

Intervention(s): busulfan; pharmacological study; laboratory biomarker analysis; and lentivirus vector rHIV7-shTAR-CCR5RZ-transduced hematopoietic progenitor cells

Three Chemo Regimens as an Adjunct to ART for Treatment of Advanced AIDS-KS

This **Phase 3** trial is **not yet recruiting**. It is an **interventional** assessment of a **drug** intervention.

Location(s): No trial sites listed

Condition(s): HIV Infections

Intervention(s): Etoposide; Bleomycin and Vincristine (BV); Coformulated EFV/FTC/TDF; and Doxorubicin HCL Liposome Injection (PLD)

Study of People With HIV Infection Who Have High Viral Loads Despite Combination Antiretroviral Therapy

This trial is **recruiting**. (No information was provided about the phase or intervention(s) of the trial.)

Location(s): National Institutes of Health Clinical Center, 9000 Rockville Pike, Bethesda, Maryland 20892

Condition(s): Immunologic Deficiency Syndromes; HIV Infections; Acquired Immunodeficiency Syndrome; and Virus Diseases

Intervention(s): None listed

List of trials, with patient filters

DiscoverCT

Trials associated with Pfizer 3730 trials match query (most recent shown first)

Make selections, then click

Intervention Type(s)

All (default)
Drug
Procedure
Behavioral
Other

Trial Status(es)

All (default)
Completed
Recruiting
Active, not recruiting
Not started

Condition(s)

Only show trials that have submitted results

Study of Safety And Efficacy Of ReFacto AF In Previously Untreated Hemophilia A Patients In The Usual Care Setting

This Phase 4 trial is **recruiting**. It is an **interventional** assessment of a **procedure** intervention.

Location(s): Pfizer Investigational Site, Lille Cedex, France, and 22 other facilities
Condition(s): Hemophilia A
Intervention(s): Laboratory Tests

Effects of Oral Sildenafil on Mortality in Adults With PAH

This Phase 4 trial is **not yet recruiting**. It is an **interventional** assessment of a **drug** intervention.

Location(s): No trial sites listed
Condition(s): Hypertension, Pulmonary and Hypertension
Intervention(s): sildenafil citrate

Extension Study Evaluating Etanercept in 3 Subtypes of Childhood Arthritis

This Phase 2/Phase 3 trial is **active, but not recruiting**. It is an **interventional** assessment of a **drug** intervention.

Location(s): The Children's Hospital Westmead, Westmead, Sydney, Australia, and 34 other facilities
Condition(s): Arthritis, Psoriatic; Arthritis; and Arthritis, Juvenile Rheumatoid
Intervention(s): etanercept

Supportive Scheduling in Metastatic Basal Cell Carcinoma (mBCC)

List of trials, with researcher filters

DiscoverCT

Assessing the Long Term Effectiveness and Safety of Biotherapies in the Treatment of Psoriasis

This trial is **recruiting**. (No information was provided about the phase or intervention(s) of the trial.)

Lead trial sponsor: Assistance Publique - Hôpitaux de Paris

Collaborating institutions: French Health Products Safety Agency; Société de Dermatologie Française; Janssen, LP; Pfizer; Abbott; and Merck Sharp & Dohme Corp.

Research Site

Henri Mondor Hospital
Creteil, France

See this trial on [ClinicalTrials.gov](https://clinicaltrials.gov)

Conditions addressed by this trial

Officially assigned:

- Psoriasis

Suggested by DiscoverCT:

No suggested terms.

Trial Data Quality Ratings ?

Overall data quality	★★★★★
MeSH classification quality	★★★☆☆
Site data quality	★★★★★
Protocol description quality	★★★★★
Trial criteria description quality	★★★★★
Date quality	★★★★★

Trial page



Identify important criteria concepts

Assessing the Long Term Effectiveness and Safety of Biotherapies in the Treatment of Psoriasis

Go to trial page: [NCT01617018](#)

Below are the eligibility criteria for the above-referenced trial. Words and phrases that may represent general criteria concepts are highlighted. Select a term on the right in order to develop it further and find trials that share a similar criteria concept. After your work is reviewed by an administrator, this concept will appear in the trial search filters.

A good criteria concept is one that applies to a nontrivial number of trials, but may be expressed using a variety of words or phrases.

Eligibility criteria

Inclusion criteria:

- Patients aged 18 years
- Having been informed of the objectives and conduct of the research and having signed a written informed consent to participate

Potential criteria concepts

objectives	Not part of any concept	Start a new concept
conduct	Not part of any concept	Start a new concept
research	Not part of any concept	Start a new concept
written informed consent	Not part of any concept	Start a new concept

Structure criteria page



Concept: smoking

A term is a word or phrase that is related to the concept.

Is the term **family history** related to this concept?

Progress: 5 / 20 terms until predictor step

Included terms

regular use
 pack year
 pack years
 cigarettes day
 smoke
 smoked
 cigarettes
 pack
 having smoked
 tobacco user
 cigarette
 current smoker

Excluded terms

medication
 no history
 alcohol
 performance status
 patients must
 chemotherapy
 opinion
 platelet count
 times
 entry
 upper limit
 participation

Included predictors

pack per
 current or
 cigarette smokers
 smokers :
 products in
 used any
 smoker of
 smokers defined
 smokers who
 number of
 smoking history
 current cigarette

Excluded predictors

greater than
 non-smoker for
 day for
 ex-smokers with
 met the
 criteria for
 non smokers
 period preceding
 preceding the
 subjects who
 have a
 reduce the

Active learning interface



Medical Subject Heading (MeSH) suggestions

Paste some text:

the second leading risk factor associated with death and disability-adjusted life-years (DALYs), accounting for over 450,000 deaths and nearly 10,000 DALYs[a1] [ao2] [LH3] . [Murray 2013]. Tobacco use is especially relevant to the urologic community; recent data have shown that nearly half of all bladder cancer cases may be attributable to tobacco use. [Strope 2008] Smoking has also been linked to renal cell carcinoma, upper tract urothelial carcinoma, and erectile dysfunction. [Hunt 2005, Hagiwara 2013] These tobacco-related diseases, particularly bladder cancer, represent significant preventable patient morbidity and high costs to the U.S. health care system. [Murray 2013]

Smoking has been shown to have a significant impact on surgical recovery. Smoking is associated with an increased risk of perioperative cardiovascular, pulmonary, and wound healing complications, including re-intubation, respiratory failure, wound infections, anastomotic dehiscence, re-intubation, and inferior long-term surgical outcomes. [Sorensen 2012, Khullar 2012] In addition, tobacco use is associated with a higher rate of perioperative complications, including longer hospital stays, higher rates of ICU admission, greater need for repeat surgery, decreased patient satisfaction, and higher overall costs of care. [Khullar 2012]

Despite the strong link to urologic cancer, multiple studies have demonstrated that urologist do not address smoking cessation with their patients. Only 22% of patients with bladder cancer were aware that tobacco use was a risk factor for bladder cancer and only 7% of patients with bladder cancer reported that their urologist recommend they stop smoking. [Dearing 2005] A national survey of 1,800 American urologists reported that 56% of them never discussed smoking cessation with their patients. In that cohort, 38% of respondents believed they were unqualified to give proper smoking cessation counseling and 41% believed that cessation would not alter their patients' disease course. [Bjurlin 2010]

Submit

Suggested terms

- Smoking
- Tobacco Use Disorder

MeSH suggestion page



Administrator tools

Criteria concept review ([Go to MeSH term assignment review](#))

You are currently reviewing fragility fracture ▾

Terms added by test from test

- osteoporotic fracture
- fracture
- vertebral fracture
- fractures
- fragility fractures
- vertebral fractures
- Select all

Approve terms

Reject this entire concept

Administrator approval interface for criteria concepts



Administrator tools

MeSH term assignment ([Go to criteria concept review](#))

Please review the following assignments

Title	New term	User name	User institution
<input checked="" type="checkbox"/> Pharmacogenetic Determinants Of Treatment Response In Children	Leukemia	test	test
<input checked="" type="checkbox"/> Pharmacogenetic Determinants Of Treatment Response In Children	Leukemia, Lymphoid	test	test
<input checked="" type="checkbox"/> Temsrolimus, Carboplatin, and Paclitaxel as First-Line Therapy in Treating Patients With Newly Diagnosed Stage III-IV Clear Cell Ovarian Cancer	Disease Progression	test	test
<input checked="" type="checkbox"/> Immunological Mechanisms of Oralair® in Patients With Seasonal Allergic Rhinitis	Hypersensitivity	test	test
<input checked="" type="checkbox"/> Select all			

Approve selected assignments

Administrator approval interface for MeSH term assignments

Institution deduplication and publication linkage

The trial registry database has approximately 271,000 unique facility names at 528,000 locations. There are a further 30,000 unique sponsor names, many of which overlap with the facility names.

We were able to combine 121,000 unique facilities (out of 528,000 name/location combinations) into 1,075 major institutions that are also linked to their canonical representation in Freebase. 2,000 sponsors could also be linked to these same institutions. Because many of these facilities and sponsors are associated with multiple trials, 78% of studies in the registry are associated with one of these 1,075 major institutions.

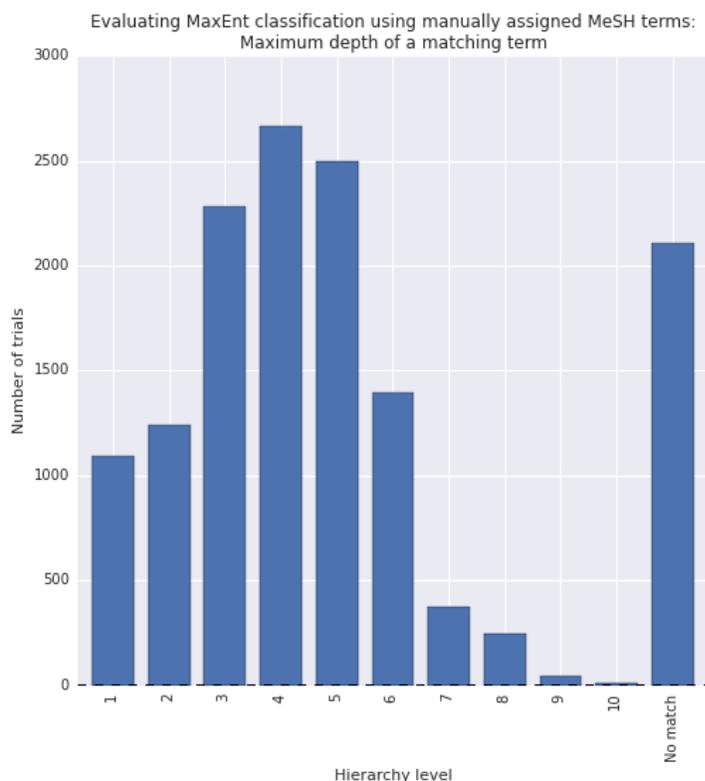
Following deduplication, we disregarded around 190,000 facilities because they had meaningless names like “Local Institution” or “Site #4” with no specific location information. The remaining facilities were linked, if possible, to sponsors using simple name matching, and are also retrievable via the search interface.

We also had some success in linking potential publications to trials in the registry. The existing registry has links between approximately 140,000 publications and 28,000 trials. Using PubMed to link publications using author (investigator) name, MeSH

term(s), and publication (trial) date, we associated a further 28,000 publications with 6,000 trials.

MeSH term suggestions

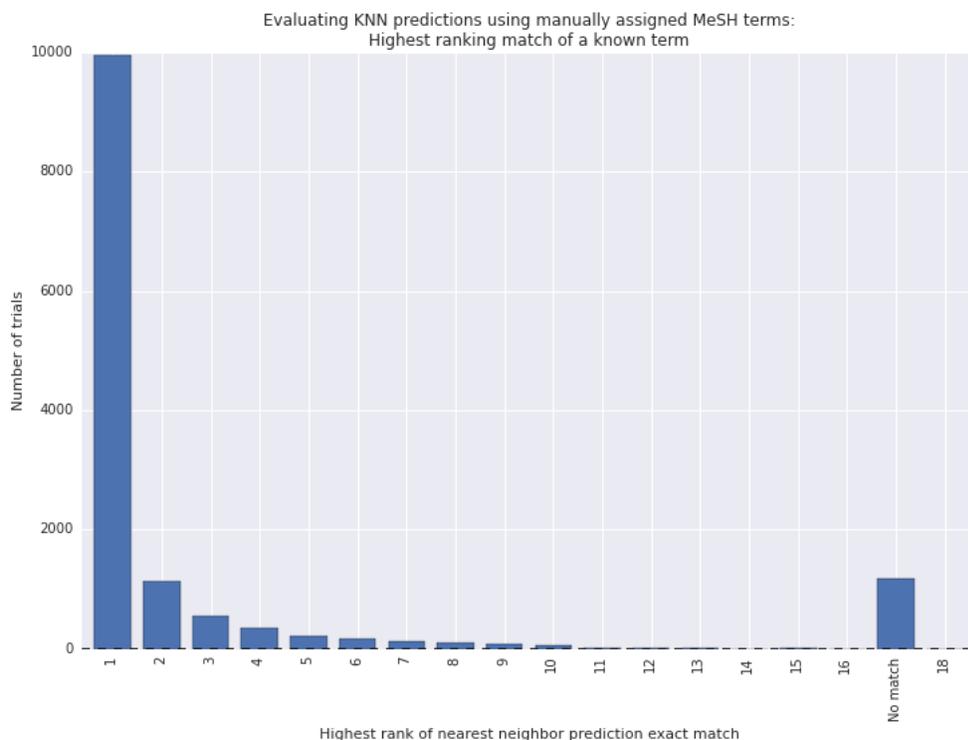
We evaluated the maximum entropy model’s single-class prediction by looking at how close this prediction was to manually assigned MeSH terms on a holdout sample. (For this and the other models, we held back 10% of labeled trials, or roughly 14,000 randomly selected trials, in order to test the models we trained on the remaining 90%.) The vast majority (85%) of trials in the holdout sample had a predicted term that matched a manually assigned MeSH term at some level of the hierarchy. In fact, 67% of the holdout sample predictions matched known terms at the fourth level or deeper in the hierarchy, indicating excellent specificity in identifying relevant MeSH terms:



Evaluating the maximum entropy classifier: depth of the closest matching term in the MeSH hierarchy

The KNN model produces a ranked list of MeSH condition terms from “neighboring” trials, so to evaluate this model we calculated the highest ranking suggestion that matched a manually assigned term; just 8% of trials in the holdout sample failed to have any match among the KNN list. Moreover, 71% of trials in the holdout sample had a

manually assigned MeSH term that was the top-ranked KNN recommendation. Spot-checking indicates that this model often has a high rate of overlap with other manually assigned MeSH terms as well, although these results are difficult to statistically summarize. In short, we had the same findings as Trieschnigg et al. that KNN model is superior to other approaches in identifying relevant MeSH terms for unlabeled text.



Evaluating the KNN classifier: highest output rank of term that exactly matches an existing term

The series of logistic regression models were difficult to evaluate, but did not appear to perform well on a manual inspection of the results. We made some efforts to more quantitatively understand the quality of these suggestions, but ultimately left them out of our interface due to lack of confidence that they would provide meaningful information about a trial.

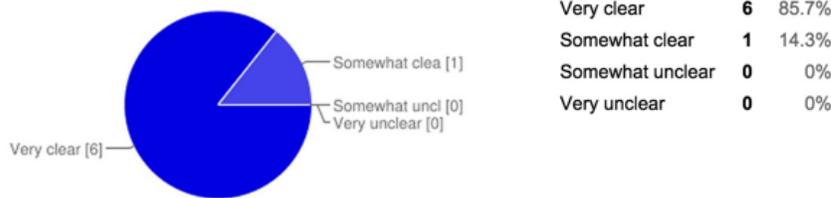
User testing and navigation paths

The results of our user testing were helpful in improving our user interface. These tests identified users' pain points when navigating the site, while also giving us the assurance

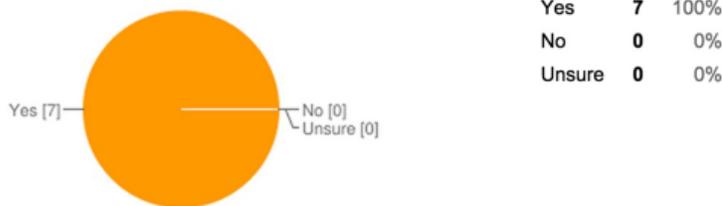
that other areas of the site did not need further attention. The summary of these tests can be found at http://bit.ly/discoverct_usertest

For example, we found that the landing page (initial impression), search interface, trial results filters, and the trial description pages were all in good shape according to the users interviewed and surveyed. (See the survey in Appendix B for the exact question wording.)

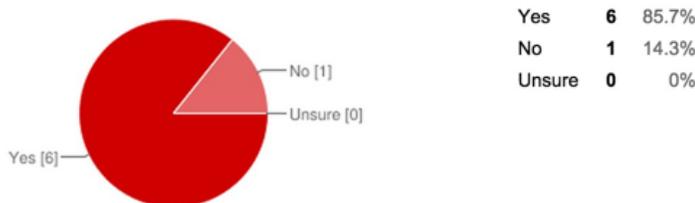
Initial impression



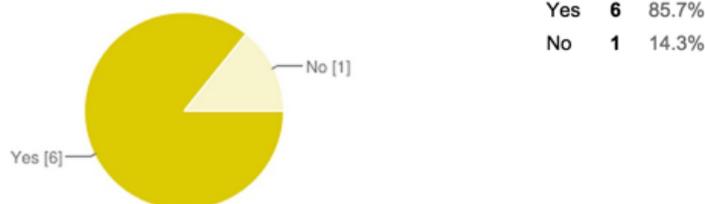
Searching



Filtering trial results

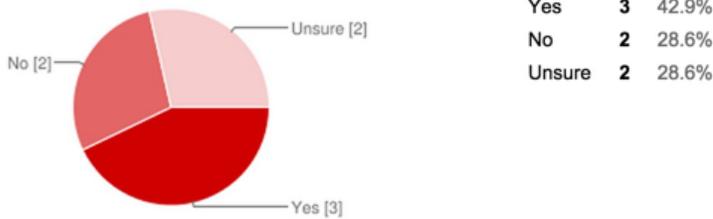


Trial description page



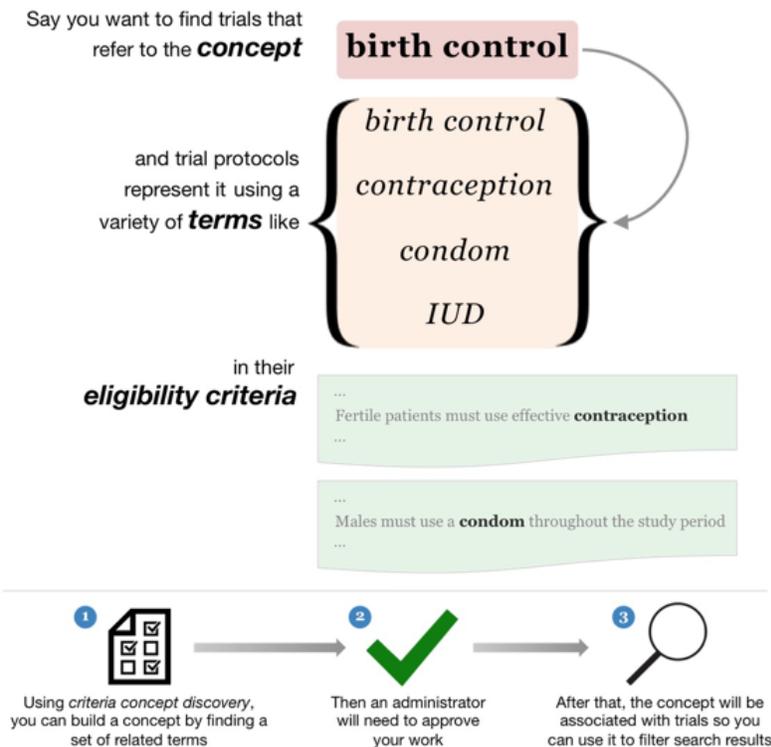
On the other hand, it became clear that we still needed to improve the user interface surrounding the active learning tool. Test results indicated that the descriptions and interface were not very clear to many of the users.

Creating a concept



Exploring responses to open-ended questions from the in person interviews and surveys, we found that people were having a difficult time grasping the motivation behind the active learning tool. We had not adequately explained the system in our pop-up modal that preceded the use of the system, so we came up with a diagram that visually explains the concepts behind the system as well as what happens after you finish using the tool:

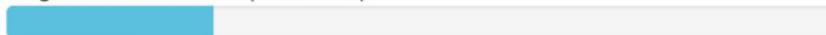
What is criteria concept discovery?



We also added a progress bar to the interface to let users know where they are in the process, as we found they did not have a clear understanding of how long the process would take.

Is the term **allergy** related to this concept?

Progress: 5 / 20 terms until predictor step



There were several other minor design and wording changes we made throughout the site in response to the feedback to reduce confusion and improve the flow of the site.

Next steps

Improving active learning

Going forward there are two major aspects we would like to add to the active learning interface to improve its accuracy and recall: we would like to integrate the word2vec package more deeply into the active learning process, and we would like to add an inverse document frequency (idf) element to improve predictor weighting.

We have found that word2vec is often very effective at finding synonyms of terms within the corpus (although it can occasionally fail on terms you would expect it to catch). We tested incorporating this algorithm further into the term collection process by creating an extended system where every time a user accepted a term generated by the active learning system it would pass it to word2vec and would return the top 3 terms that had not been seen by the user before for approval or denial. This resulted in the user being able to flesh out their concept much more quickly. Unfortunately, in the time frame of the project we were not able to translate this new word2vec interaction into the site's active learning framework.

Further, the addition of idf in the predictor process would enable us to more effectively weight the predictors, resulting in the system showing more relevant suggestions to the user. Currently, predictors are based solely on a "term frequency" model, so the most frequently co-occurring predictors tend to look the same for many concepts. If we could efficiently calculate and store the inverse document frequency of every possible predictor phrase in the criteria eligibility text, we could weight predictors more appropriately to suggest more relevant words and phrases to the user.

Interface improvements

Although criteria eligibility concepts make it easier to discover relevant trials, it was more challenging for users to understand how these concepts were built and developed in the first place. We created visual diagrams and additional description text to make this process more clear, but also feel that a concept summary interface may make it easier to understand exactly which terms are included in a particular concept.

We also had some interest in directly contacting a trial investigator or administrator, and in the future would like to identify the best source of contact information for a trial in order to provide it to interested users.

External integration

Finally, we believe the data improvements enabled by DiscoverCT.org could be useful beyond our website, perhaps including electronic medical record (EMR) integration or the provision of additional structured data to ClinicalTrials.gov.

EMRs are widely used by hospitals and other health care organizations to manage a patient's medical information. These electronic files typically include an array of historical and diagnostic information about a patient, such as known allergic reactions, medications they have taken, and past surgeries and diagnoses.

The trial eligibility criteria concepts created using our active learning interface could be developed in order to align with EMRs' descriptions of patient characteristics. Ideally, a health care provider could use a crosswalk between these systems to identify all patients who may be eligible for a particular trial, or all trials for which a particular patient may be eligible. This may require significant effort from an EMR vendor and/or health care institution, but could drastically reduce the friction of trial enrollment, which is currently conducted in a largely opportunistic fashion.

In addition, the ClinicalTrials.gov registry could benefit from our institution deduplication efforts to address the lack of standardization in research facility and trial sponsor names, as well as the additional MeSH terms suggested by our algorithms and accepted by users. By providing a feedback loop to the primary data repository at NIH, the benefits conferred by improved data would not be limited to our site but could be shared by all users of the ClinicalTrials.gov registry.

References

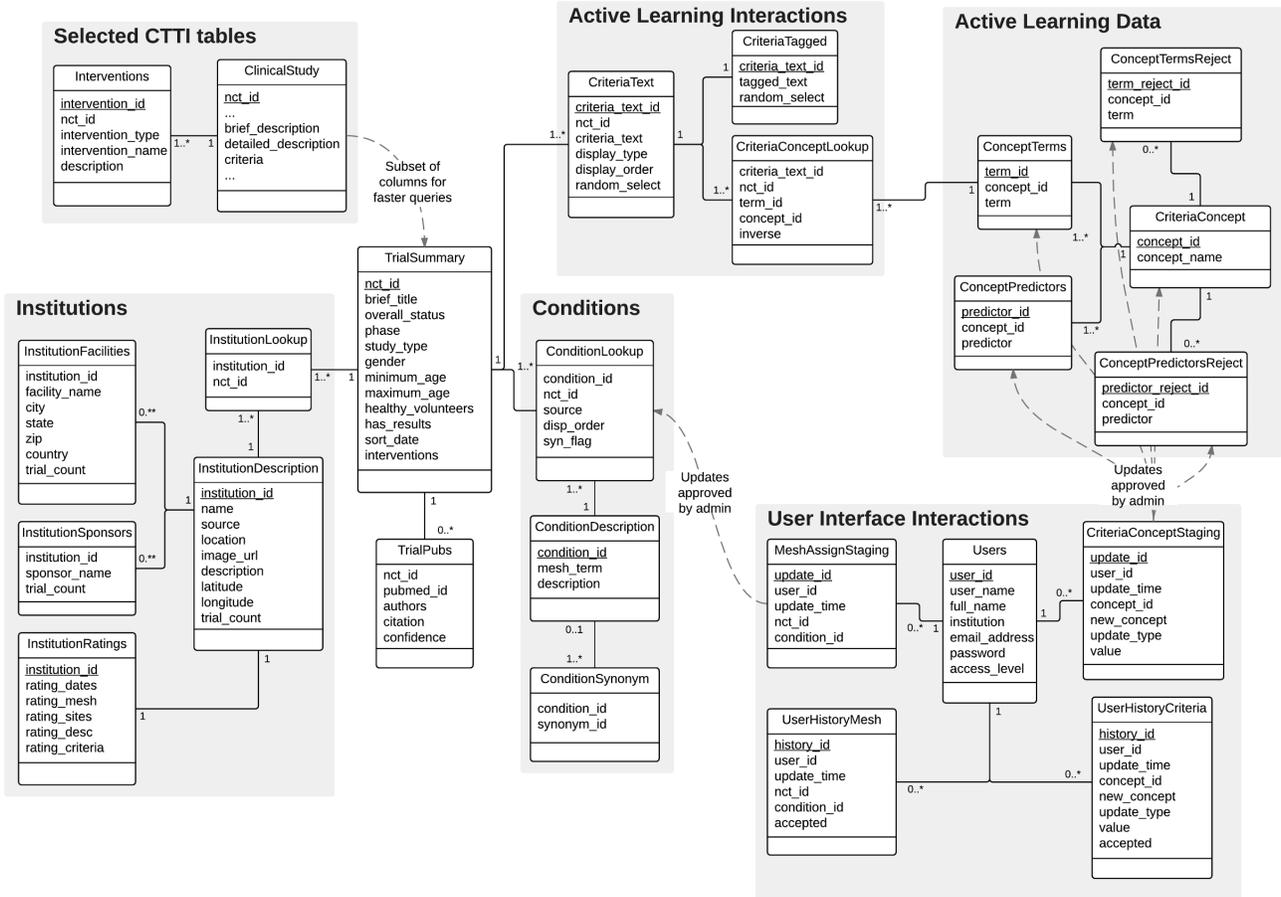
- Anderson, M. L., Chiswell, K., Peterson, E. D., Tasneem, A., Topping, J., & Califf, R. M. (2015). Compliance with Results Reporting at ClinicalTrials.gov. *New England Journal of Medicine*, 372(11), 1031–1039. <http://doi.org/10.1056/NEJMsa1409364>
- Bell, S. A., & Tudur Smith, C. (2014). A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of ClinicalTrials.gov. *Orphanet Journal of Rare Diseases*, 9(1), 170. <http://doi.org/10.1186/s13023-014-0170-0>
- Bilenko, M. (2006, August). *Learnable Similarity Functions and Their Application to Record Linkage and Clustering*. University of Texas, Austin, Texas. Retrieved from <http://www.cs.utexas.edu/~ml/papers/marlin-dissertation-06.pdf>
- Califf, R. M. (2012). Characteristics of Clinical Trials Registered in ClinicalTrials.gov, 2007-2010. *JAMA*, 307(17), 1838. <http://doi.org/10.1001/jama.2012.3424>
- CTTI. (2015). State of Clinical Trials. Retrieved May 1, 2015, from <http://www.ctti-clinicaltrials.org/what-we-do/analysis-dissemination/state-clinical-trials/aact-database>
- Enserink, M. (2015). Withholding results from clinical trials is unethical, says WHO. *Science*. <http://doi.org/10.1126/science.aab2484>
- Freebase. (2011). What is Freebase? Retrieved May 7, 2015, from http://wiki.freebase.com/wiki/What_is_Freebase%3F
- Freebase. (2015, March 26). As previously announced. Retrieved from <https://plus.google.com/109936836907132434202/posts/3aYFVNf92A1>
- Getz, K. A., Campo, R. A., & Kaitin, K. I. (2011). Variability in Protocol Design Complexity by Phase and Therapeutic Area. *Drug Information Journal*, 45(4), 413–420. <http://doi.org/10.1177/009286151104500403>
- Gregg, F., Eder, D., & other contributors. (2015). Making Smart Comparisons. Retrieved May 1, 2015, from <http://dedupe.readthedocs.org/en/latest/Making-smart-comparisons.html>
- Guharoy, V. (2014). ClinicalTrials.Gov: Is the Glass Half Full? *Hospital Pharmacy*, 49(10), 893–895. <http://doi.org/10.1310/hpj4910-893>
- Hartung, D. M., Zarin, D. A., Guise, J.-M., McDonagh, M., Paynter, R., & Helfand, M. (2014). Reporting Discrepancies Between the ClinicalTrials.gov Results Database and Peer-Reviewed Publications. *Annals of Internal Medicine*, 160(7), 477. <http://doi.org/10.7326/M13-0480>
- ICMJE. (2015). Clinical Trials Registration. Retrieved May 1, 2015, from <http://www.icmje.org/about-icmje/faqs/clinical-trials-registration/>

- Inrig, J. K., Califf, R. M., Tasneem, A., Vegunta, R. K., Molina, C., Stanifer, J. W., ... Patel, U. D. (2014). The Landscape of Clinical Trials in Nephrology: A Systematic Review of ClinicalTrials.gov. *American Journal of Kidney Diseases*, 63(5), 771–780. <http://doi.org/10.1053/j.ajkd.2013.10.043>
- Kuehn, B. M. (2012). Few Studies Reporting Results at US Government Clinical Trials Site. *JAMA: The Journal of the American Medical Association*, 307(7), 651–653. <http://doi.org/10.1001/jama.2012.127>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv Preprint arXiv:1301.3781*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).
- National Institutes of Health. (2014). Learn About Clinical Studies. Retrieved May 7, 2015, from <https://www.clinicaltrials.gov/ct2/about-studies/learn>
- National Institutes of Health. (2015). Funding Facts. Retrieved May 1, 2015, from <http://report.nih.gov/fundingfacts/fundingfacts.aspx>
- National Library of Medicine. (2005). PubMed Help. Bethesda, MD: National Center for Biotechnology Information (US). Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK3827/>
- National Library of Medicine. (2008). Clinical Trial Phases. Retrieved May 7, 2015, from <http://www.nlm.nih.gov/services/ctphases.html>
- National Library of Medicine. (2014). MeSH Browser. Retrieved May 7, 2015, from http://www.nlm.nih.gov/mesh/2015/mesh_browser/MBrowser.html
- National Library of Medicine. (2015). MedlinePlus XML Files. Retrieved May 7, 2015, from <http://www.nlm.nih.gov/medlineplus/xml.html>
- O'Reilly, E. K., Hassell, N. J., Snyder, D. C., Natoli, S., Liu, I., Rimmler, J., ... Stacy, M. (2015). ClinicalTrials.gov Reporting: Strategies for Success at an Academic Health Center. *Clinical and Translational Science*, 8(1), 48–51. <http://doi.org/10.1111/cts.12235>
- PhRMA. (2014). *2014 Profile, Biopharmaceutical Research Industry*. Washington, DC: Pharmaceutical Researchers and Manufacturers of America.
- Riveros, C., Dechartres, A., Perrodeau, E., Haneef, R., Boutron, I., & Ravaud, P. (2013). Timing and Completeness of Trial Results Posted at ClinicalTrials.gov and Published in Journals. *PLoS Medicine*, 10(12), e1001566. <http://doi.org/10.1371/journal.pmed.1001566>
- Saito, H., & Gill, C. J. (2014). How Frequently Do the Results from Completed US Clinical Trials Enter the Public Domain? - A Statistical Analysis of the

- ClinicalTrials.gov Database. *PLoS ONE*, 9(7), e101826.
<http://doi.org/10.1371/journal.pone.0101826>
- Tasneem, A., Aberle, L., Ananth, H., Chakraborty, S., Chiswell, K., McCourt, B. J., & Pietrobon, R. (2012). The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty. *PLoS ONE*, 7(3), e33677.
<http://doi.org/10.1371/journal.pone.0033677>
- Trieschnigg, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W., & Rebholz-Schuhmann, D. (2009). MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11), 1412–1418.
<http://doi.org/10.1093/bioinformatics/btp249>
- Rehurek, R. , & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.
- Sayers, E. (2010). Entrez Programming Utilities Help. Bethesda, MD: National Center for Biotechnology Information (US). Retrieved from
<https://www.ncbi.nlm.nih.gov/books/NBK25497/>
- Vodicka, E., Kim, K., Devine, E. B., Gnanasakthy, A., Scoggins, J. F., & Patrick, D. L. (2015). Inclusion of patient-reported outcome measures in registered clinical trials: Evidence from ClinicalTrials.gov (2007–2013). *Contemporary Clinical Trials*.
<http://doi.org/10.1016/j.cct.2015.04.004>
- World Health Organization. (2015). Data Providers. Retrieved May 1, 2015, from
http://www.who.int/ictrp/search/data_providers/en/
- Zarin, D. A., Tse, T., & Ide, N. C. (2005). Trial Registration at ClinicalTrials.gov between May and October 2005. *New England Journal of Medicine*, 353(26), 2779–2787.
<http://doi.org/10.1056/NEJMsa053234>
- Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., & Ide, N. C. (2011). The ClinicalTrials.gov Results Database — Update and Key Issues. *New England Journal of Medicine*, 364(9), 852–860. <http://doi.org/10.1056/NEJMsa1012065>

Appendix

Appendix A: Entity-relationship diagram of backend database



Appendix B: User testing survey

DiscoverCT general usability

Thank you for taking the time to give us feedback on DiscoverCT. These 10 questions should take around 10 minutes to answer. Your responses will directly influence the design and content of this project.

Please visit our site in order to answer the questions below:

<http://www.discoverct.org>

Please note: the site works best on the Chrome browser.

Initial impression

Based on the home page, how clear do you think the purpose of this site is?

- Very clear
- Somewhat clear
- Somewhat unclear
- Very unclear

Please provide any comments you have about this

Searching

Try searching for a condition (e.g. "Lung Cancer") or institution (e.g. "UCSF"). Does this work as you expect?

- Yes
- No
- Unsure

Please provide any comments you have about this

Filtering trial results

Go to the bottom of a condition or institution page and look at the list of associated trials. (If you have any trouble, use this link: <http://discoverct.org/condition?cond=2192#trial-pane>) Use the selectors to filter the trials. Does this work as you expect?

- Yes
- No
- Unsure

Please provide any comments you have about this

Trial description page

Select a trial and look at the trial description page. (If you have any trouble, use this link: http://discoverct.org/trial?nct_id=NCT00881712) Did you notice the wrench icon next to the "Eligibility Criteria" heading?

- Yes
- No

Do you understand that you can click this wrench icon?

- Yes
- No

Please provide any comments you have about these icons or the trial page generally

Logging in

Click on the "Login" button in the upper right, and log in with the username "test" and password "test" (without quotes). Are you able to log in?

- Yes
- No

Continue »

 33% completed

DiscoverCT general usability

DiscoverCT criteria concept discovery

Opening dialog

After logging in, click on the wrench icon next to "Eligibility Criteria" on the trial description page. Read the dialog box that comes up. Do you understand what this tool does?

- Yes
- No
- Unsure

Please provide any comments you have about this

Identifying a criteria concept

Click OK, and a new page should open. (If you have any trouble, use this link: http://discoverct.org/structure_trial_criteria?nct_id=NCT00881712) Take a look at the page. Do you generally understand what you are supposed to do?

- Yes
- No
- Unsure

Please provide any comments you have about this

Creating a concept

Click on a "Start a new concept" link, preferably for a term that might appear in the eligibility criteria for a number of trials. (If you have any trouble, use this link: http://discoverct.org/active_learning?term=chemotherapy) Go through several rounds of the yes/no questions. Do you understand what is happening?

- Yes
- No
- Unsure

If any dialogs pop up as you are going through this process, are they clear to you?

- Yes
- No
- Unsure

Please provide any comments you have about this page

« Back

Continue »



66% completed

DiscoverCT general usability

DiscoverCT exploration

[Free explore](#)

Please explore the site as much as you'd like, and let us know any additional feedback you have

« Back

Submit



100%: You made it.

Never submit passwords through Google Forms.