Preparing the study dataset from ClinicalTrials.Gov for analysis using disease conditions to re-group clinical trials by clinical specialty

Asba Tasneem, PhD[1], Hari Ananth[1], Jean Bolte[2], Swati Chakraborty[1], Karen Chiswell[1], PhD, Brian J. McCourt[1], Robert M. Califf, MD[2]

[1]Duke Clinical Research Institute, Durham, NC; [2]Duke Translational Medicine Institute, Durham, NC

INTRODUCTION

Clinicaltrials.gov is the largest repository of information on clinical research studies. It provides a wealth of information to patients, clinicians, and researchers on individual clinical trials. This registry contains over 100,000 studies conducted in more than 170 countries worldwide. Although registry data are available for bulk download to facilitate aggregate analysis, issues related to data structure, nomenclature, and changes in data collection over time limit the quantitative interpretation of these data. As data have accumulated in the registry, there have been increasing demands for capabilities that would allow descriptive characterization of the overall trials portfolio of the clinical research enterprise. In response, the Clinical Trials Transformation Initiative (CTTI) launched a project to create a database and documentation that would support such aggregate analysis.

Through a CTTI initiative, we built a relational database comprising information on all studies in ClinicalTrials.gov. This database includes the following features: 1) data parsed into discrete fields and 2) complete metadata. Additionally, we developed a comprehensive change history document from the changes in ClinicalTrials.gov's protocol data element definitions. To facilitate analysis of studies within and across specialties, we present additional methodology for annotation of studies by clinical specialty using a custom Medical Subject Heading (MeSH) taxonomy with subsequent validation.

MATERIALS AND METHODS

A dataset comprising 96,346 clinical trials was downloaded from ClinicalTrials.gov on Sep 27, 2010 in XML format. We first designed and implemented a relational database to analyze these aggregate data. An example of the data transformations performed includes parsing the "Study Design" data element, which contains concatenated data from several other data elements (Figure 1).

In order to regroup studies in clinical specialties, we have developed a methodology using the National Library of Medicine's (NLM) MeSH thesaurus (2010 version) combined with information from other fields of the database. Disease conditions provided by submitters (CONDITIONS) and MeSH condition terms (CONDITION_BROWSE) generated by the NLM algorithm, both of which were available in the downloaded study dataset, were used to classify specialties (Figure 2a). We also developed a process for annotating, validating, and implementing disease conditions (MeSH and non-MeSH terms) to create specialty datasets (Figure 2b). Non-MeSH condition terms were selected from interventional studies registered after September 2007 that appeared in five or more studies. Selected disease condition terms (MeSH and non-MeSH) were reviewed and annotated by faculty and clinicians within each clinical discipline at Duke University Medical Center. A quantitative profile of data quality was also created.

Figure 1: Parsing "Study Design" into its components

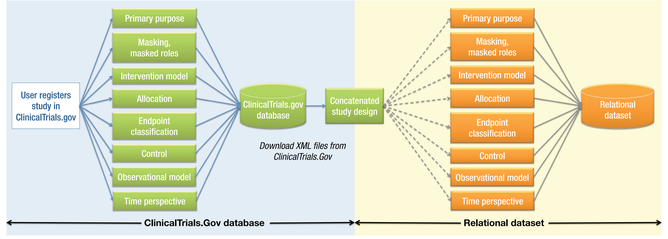
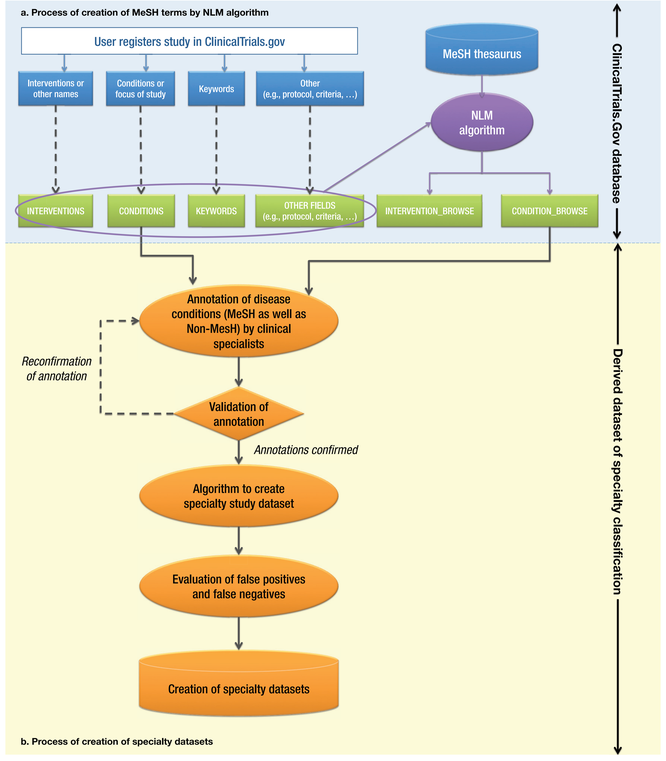


Figure 2: An overview of methodology and process of developing clinical specialty datasets.
The INTERVENTIONS, CONDITIONS, and KEYWORDS tables comprise disease condition terms provided by submitters and include both MeSH and non-MeSH terms. INTERVENTION_BROWSE and CONDITION_BROWSE tables are populated by MeSH terms generated by the NLM algorithm.



RESULTS

Relational dataset

1. A relational database was developed to support aggregate analysis of clinical trials registered with ClinicalTrials.gov.
2. The "Study Design" data element was parsed into its original data elements, which can now be analyzed individually.
3. The data dictionary includes comprehensive metadata from the analysis database.
4. Changes in individual data element definitions from 24 releases of ClinicalTrials.gov Protocol Data Element Definitions have been captured and documented.

Specialty classification

1. A methodology was developed to create specialty datasets (Figure 3).
2. An algorithm was created to classify studies into clinical specialties based on both MeSH and Non-MeSH annotations.
3. False positives (e.g., non-cardiology studies classified as cardiology) and false negatives (e.g., cardiology studies classified as non-cardiology) were evaluated by comparing algorithmic classification with manual classification using three specialties.

Figure 3: An aggregate analysis dataset of clinical trials downloaded from ClinicalTrials.gov and reclassified by clinical specialty

CONCLUSION

This project seeks to improve the public availability of aggregate data from the ClinicalTrials.gov registry. Products that will soon be publicly available include a relational database of all data in ClinicalTrials.gov as well as supporting documentation, including a detailed data dictionary that incorporates the history of changes to data element definitions. Using expert classification of MeSH and other terms describing disease conditions, we are classifying studies into 13 selected disease specialty areas and evaluated the performance of the classification method within three of the specialties. This derived dataset with specialty classification will be used to develop manuscripts describing the state of clinical trials in the U.S. and in each of 13 specialty areas. A paper describing the methodology and algorithm for disease specialty classification will be published.

ACKNOWLEDGMENTS

Deborah Zarin, MD
Director, ClinicalTrials.gov
National Library of Medicine, National Institutes of Health

Nick Ide, MS
National Library of Medicine, National Institutes of Health

Principal Investigator:
Robert M. Califf, MD
Director, Duke Translational Medicine Institute

Financial support provided by grant U19 FD003800 from the U.S. Food and Drug Administration awarded to Duke University for the Clinical Trials Transformation Initiative (https://www.trialstransformation.org/).

CONTACT INFORMATION

Asba Tasneem, PhD
asba.tasneem@duke.edu
919-668-7188

DCRI COMMUNICATIONS • APRIL 2011