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

Asba Tasneem, PhD1, Hari Ananth1, Jean Bolte2, Swati Chakraborty1, Karen Chiswell1, PhD, Brian J. McCourt1, Robert M. Califf, MD2

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 150,000 studies conducted in more than 175 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 portfolio of the clinical research enterprises. In response to this, the Clinical Trials Transformation Initiative (CTTI) launched a project to create a database and documentation that would facilitate analysis of studies within and across the clinical trials. This registry contains over 100,000 studies conducted as a registry, there have been increasing demands for capabilities that would allow descriptive characterization of the overall portfolio of the clinical research enterprises. In response to this, the Clinical Trials Transformation Initiative (CTTI) launched a project to create a database and documentation that would facilitate analysis of studies within and across the clinical trials.

A relational database was developed to support aggregate analysis of clinical trials registered with ClinicalTrials.gov. A “Study Design” data element, which contains concatenated data from several other data elements (Figure 1).

In order to migrate 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 and INTERVENTION_BROWSE) were populated by MeSH terms generated by the NLM algorithm.

As a result of the Clinical Trials Transformation Initiative, a relational database containing specialty datasets (Figure 2a). The data dictionary includes comprehensive metadata as cardiology) and false positives (e.g., cardiology studies classified as non-cardiology) were evaluated using three specialties.

RESULTS

Conception

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 areas and evaluating the performance of the current 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.

MATERIALS AND METHODS

A dataset comprising 64,246 clinical trials was downloaded from ClinicalTrials.gov on Sep 27, 2011 in XML format. We first designed and implemented a relational database to aggregate 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 migrate studies in clinical specialties, we have developed a methodology using the National Library of Medicine’s (NLM) MeSH thesaurus (2010 version) combined with MeSH condition terms (CONDITION_BROWSE and INTERVENTION_BROWSE) populated by MeSH terms generated by the NLM algorithm.

We also developed a process for annotating, validating, and implementing disease conditions (MeSH and non-MeSH) to create specialty datasets (Figure 2b). Non-MeSH condition terms were selected from international 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.

ACKNOWLEDGMENTS

Asba Tasneem, PhD

CONTACT INFORMATION

Director, Duke Translational Medicine Institute

Robert M. Califf, MD

Deborah Zarin, MD

Nick Ide, MS

Asba Tasneem, PhD

asba.tasneem@duke.edu

Asba Tasneem, PhD

4/19/11 4:28 PM