Asba Tasneem, PhD1, Hari Ananth1, Jean Bolte2, Swati Chakraborty1, Karen Chiswell1, PhD, Brian J. McCourt1, Robert M. Califf, MD2
1Duke Clinical Research Institute, Durham, NC; 2Duke 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 183,000 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 has been increasing demand for capabilities that would enable descriptive characterization of the overall portfolio of the clinical research enterprises. In response to the Clinical Trials Transformation Initiative (CTTI) launched a project to create a database and documentation that would support 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) it is derived from discrete fields and 2) it is complete metadata. Additionally, we developed a comprehensive change history document from the analysis database. The data dictionary includes comprehensive metadata definitions have been captured and documented.

RESULTS

Materials and Methods
A dataset containing 66,245 clinical trials was downloaded from ClinicalTrials.gov on Sep 27, 2010 in XML format. We first designed and implemented a relational database to store 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 (2011 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 studies (Figure 2). We also developed a process for annotating, validating, and implementing disease conditions (MeSH and non-MeSH terms) to create specialty datasets (Figure 3). Non-MeSH condition terms were selected from interventional studies revealed 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: A process of methodology and process of developing clinical specialty datasets.

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

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 dashboards describing the state of clinical trials in the U.S. and in each of the 13 specialty areas. A paper describing the methodology and algorithm for disease specialty classification will be published.

ACKNOWLEDGMENTS
The authors thank the following members of the work group for their contributions:

Deborah Zarin, MD
Director, ClinicalTrials.gov
National Library of Medicine, National Institutes of Health
Robert M. Califf, MD
Principal Investigator:
Translational Sciences, Duke University
National Library of Medicine, National Institutes of Health

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

Duke Translational Medicine Institute
Durham, NC

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.

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 as specialty classification.