AACT-Results: The Results dataset extensions for the AACT database

{AACT: Aggregate Analysis database of ClinicalTrials.gov}

Asba Tasneem, PhD
Duke Clinical Research Institute
Disclosures and Learning Objectives

- No Disclosures

After participating in this talk the learner should be better able to:

- Describe the rationale for clinical trials registration and results reporting
- Describe the informatics approach to the construction of an analyzable data set from ClinicalTrials.gov
- Describe strengths and weaknesses of ClinicalTrials.Gov dataset
Background

ClinicalTrials.gov
- National registry hosted by the NLM/NIH
- Studies conducted in the United States and around the world; sponsored by the NIH, other federal agencies, and private industry
- Currently stores >164,000 studies
- Web interface for patients and patient advocates

The Clinical Trials Transformation Initiative (CTTI):
- a public-private partnership between FDA and Duke to modernize the way clinical trials are conducted
- CTTI was established by the FDA and Duke University in 2008, and now comprises more than 60 member organizations.
- Mission: to identify and promote practices that will increase the quality and efficiency of clinical trials
- Project: To improve the public interface for use of aggregate data in ClinicalTrials.gov
Phase I: AACT-Registry and AACT-Specialty

- **Available Data**
  - Data Element Definitions document of study data elements
  - Study dataset XMLs
  - MeSH Thesaurus
  - Public XSD

- **Created AACT-Registry**
  - Aggregate Analysis database of ClinicalTrials.gov (AACT) – Oracle dataset
  - Oracle extracts in three formats (Dmp, Text, SAS)
  - Integrated metadata
  - Change History document for data element definitions
  - Integrated MeSH thesaurus
  - Parsed study design
  - Date converted to Date datatype

- **Created AACT-Specialty**
  - Grouped dataset into specialty groups
  - Annotated disease condition terms
  - 13 clinical specialties and 5 subspecialties
Phase II: AACT Results and ETL Updates

- **Available Data**
  - Public XSD
  - Results and Registry XMLs (results publicly available since 2012)
  - Data Element definitions documents – study and results data elements

- **Created AACT- Results**
  - Integrated results dataset into AACT database
  - Oracle extracts available in three formats (DMP, Text, SAS)
  - Integrated metadata

- **Built semi-automated update process**
  - Automated system to update data
  - Currently this process runs semi-annually
AACT database overview with its key enhancements

- ClinicalTrials.gov Protocol Data Element Definitions
- Metadata Tables: CURRENT_VARIABLES, ENUMERATIONS, VARIABLE_HISTORY_DATES
- Designs Table with parsed Study Design (Primary Purpose, Masking, Intervention Model, Allocation, Endpoint Classification, Control, Observational Model, Time Perspective)
- ClinicalTrials.gov (Registry + Results)

- MeSH Thesaurus
- MeSH Disease Conditions Annotated by Clinicians
- MESH_REPORTING Table for MeSH Annotation Validation
- MESH_SPECIALITY Table with Annotated MeSH conditions for each specialty (e.g. Cardiology, Oncology, Mental Health, ..., etc.)
- NON_MESH_SPECIALITY Table with Annotated free-text disease conditions for each specialty (e.g. Cardiology, Oncology, Mental Health, ..., etc.)

- Aggregate Analysis
- Customized Queries
- Comparative Data Analysis
- Direct Import into Oracle, SAS etc. (excluding Specialty data sets)

Initial Specialty Data Sets
Final Specialty Data Sets
Manual review
AACT Schema: Bird’s eye view

Registry database

Results database
AACT Metadata Summary

- Number of Tables: 42
  - Registry dataset: 24
  - Results dataset: 16
  - Registry and Results: 1
  - MeSH Thesaurus: 1

- Number of Data Elements: 270
  - Registry dataset: 153
  - Results dataset: 108
  - Registry and Results: 6
  - MeSH Thesaurus: 3

- Number of Enumerated Fields: 50

Requirements
- NLM required fields: 85
- FDAAA required fields: 36
ClinicalTrials.gov process flow diagram

Use of Informatica

ClinicalTrials.gov

Tidal Enterprise Scheduler
Perl script downloads study data in the form zip file from clinicaltrial.gov

Informatica Power Center

CLINTRIALSGOV Landing, Staging area

CLINTRIALSGOV Database

Pipe delimited files for statisticians

DCRI K Drive

DCRI Network Drive Location

Studies in XML format

Archive
The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty

Asba Tasneem1*, Laura Aberle1, Hari Ananth1, Swati Chakraborty1, Karen Chiswell1, Brian J. McCourt1, Ricardo Pietrobon1,2

1 Duke Clinical Research Institute, Durham, North Carolina, United States of America, 2 Department of Surgery, Duke University School of Medicine, Durham, North Carolina, United States of America

Abstract

Background: The ClinicalTrials.gov registry provides information regarding characteristics of past, current, and planned clinical studies to patients, clinicians, and researchers; in addition, registry data are available for bulk download. However, issues related to data structure, nomenclature, and changes in data collection over time present challenges to the aggregate analysis and interpretation of these data in general and to the analysis of trials according to clinical specialty in particular. Improving usability of these data could enhance the utility of ClinicalTrials.gov as a research resource.

Methods/Principal Results: The purpose of our project was twofold. First, we sought to extend the usability of ClinicalTrials.gov for research purposes by developing a database for aggregate analysis of ClinicalTrials.gov (AACT) that contains data from the 96,346 clinical trials registered as of September 27, 2010. Second, we developed and validated a methodology for annotating studies by clinical specialty, using a custom taxonomy employing Medical Subject Heading (MeSH) terms applied by an NLM algorithm, as well as MeSH terms and other disease condition terms provided by study sponsors. Clinical specialists reviewed and annotated MeSH and non-MeSH disease condition terms, and an algorithm was created to classify studies into clinical specialties based on both MeSH and non-MeSH annotations. False positives and false negatives were evaluated by comparing algorithmic classification with manual classification for three specialties.

Conclusions/Significance: The resulting AACT database features study design attributes parsed into discrete fields, integrated metadata, and an integrated MeSH thesaurus, and is available for download as Oracle extracts (.dmp file and text format). This publicly-accessible dataset will facilitate analysis of studies and permit detailed characterization and analysis of the U.S. clinical trials enterprise as a whole. In addition, the methodology we present for creating specialty datasets may facilitate other efforts to analyze studies by specialty groups.
Aggregate Analysis (JAMA, May 2012)

Characteristics of Clinical Trials Registered in ClinicalTrials.gov, 2007-2010

Robert M. Califf, MD
Deborah A. Zarin, MD
Judith M. Kramer, MD, MS
Rachel E. Sherman, MD, MPH
Laura H. Aberle, BS/PH
Asma Tasneem, PhD

Clinical trials are the central means by which preventive, diagnostic, and therapeutic strategies are evaluated, but the US clinical trials enterprise has been marked by debate regarding funding priorities for clinical research, the design and interpretation of studies, and protections for research participants. Until recently, however, we have lacked tools for comprehensively assessing trials across the broader US clinical trial enterprise.

In 1997, Congress mandated the creation of the ClinicalTrials.gov registry to assist people with serious illnesses in gaining access to trials. In September 2004, the International Committee of Medical Journal Editors (ICMJE) announced a policy, which took effect in 2005, of requiring registration of clinical trials as a prerequisite for publication. The Food and Drug Administration Amendment Act (FDAAA) expanded the mandate of ClinicalTrials.gov to include most non-

Context  Recent reports highlight gaps between guidelines-based treatment recommendations and evidence from clinical trials that supports those recommendations. Strengthened reporting requirements for studies registered with ClinicalTrials.gov enable a comprehensive evaluation of the national trials portfolio.

Objective  To examine fundamental characteristics of interventional clinical trials registered in the ClinicalTrials.gov database.

Methods  A data set comprising 96,346 clinical studies from ClinicalTrials.gov was downloaded on September 27, 2010, and entered into a relational database to analyze aggregate data. Interventional trials were identified and analyses were focused on 3 clinical specialities—cardiovascular, mental health, and oncology—that together encompass the largest number of disability-adjusted life-years lost in the United States.

Main Outcome Measures  Characteristics of registered clinical trials as reported data elements in the trial registry; how these characteristics have changed over time; differences in characteristics as a function of clinical specialty; and factors associated with use of randomization, blinding, and data monitoring committees (DMCs).

Results  The number of registered interventional clinical trials increased from 28,881 (October 2004–September 2007) to 40,970 (October 2007–September 2010), and the number of missing data elements has generally declined. Most interventional trials registered between 2007 and 2010 were small, with 62% enrolling 100 or fewer participants. Many clinical trials were single-center (66%; 24,788/37,520) and funded by organizations other than industry or the National Institutes of Health (NIH) (47%; 17,592/37,520). Heterogeneity in the reported methods by clinical specialty; sponsor type; and the reported use of DMCs, randomization, and blinding was evident. For example, reported use of DMCs was less common in industry-sponsored vs NIH-sponsored trials (adjusted odds ratio [OR], 0.11; 95% CI, 0.09-0.14), earlier-phase vs phase 3 trials (adjusted OR, 0.83; 95% CI, 0.76-0.91), and mental health trials vs those in the other 2 specialties. In similar comparisons, randomization and blinding were less frequently reported in earlier-phase, oncology, and device trials.

Conclusion  Clinical trials registered in ClinicalTrials.gov are dominated by small trials and contain significant heterogeneity in methodological approaches, including reported use of randomization, blinding, and DMCs.

JAMA. 2012;307(17):1838-1847  www.jama.com
Completeness for Selected Study Data Elements For Interventional Trials

Gender and Lead Sponsor required by FDAAA and ClinicalTrials.gov. Enrollment required by FDAAA. At least one of interventional model, allocation, and masking required by FDAAA. Number of arms may be required by FDAAA. DMC and number of arms introduced in 4/2007.
Studies of drugs, biologics and devices in phases 2-4 are required to be registered by FDAAA. + Includes behavioral, radiation, dietary supplement, in addition to other interventions

N=1 study did not report intervention type information
Studies Reported Results by Study Type

* The ClinicalTrials.gov “basic results” database was launched on September 23, 2008
** Includes studies with results released through 27 September 2013
Completeness for Selected Results Data Elements
For Interventional Trials with Results

Participant, Age, Gender, at least one primary outcome, all SAEs, and other non-serious AEs with ≥5% incidence required for studies reporting results.
SAE and AE reporting optional prior to September 28, 2009.
Trials Providing a Citation to Published Results by Funding Among interventional trials with results posted at ClinicalTrials.gov

% of studies

Industry | NIH | Other

Funding source derived from information provided in lead sponsor and collaborator fields. Trials with NIH involvement, (e.g., as a collaborator) but no industry lead sponsor are classified as funded by NIH. Providing reference citations, including references with published results, is optional in ClinicalTrials.gov.
Trials Providing a Citation to Published Results by Phase
Among interventional trials with results posted at ClinicalTrials.gov

Providing reference citations, including references with published results, is optional in ClinicalTrials.gov
Number of Participants Enrolled
Comparison of results vs. study data for interventional trials reporting results
Number of Participants Enrolled
Comparison of results vs. study data for interventional trials reporting results

Axes limited to \( \leq 1000 \) participants
Number of Primary Outcomes

Comparison of results vs. study data for interventional trials reporting results

![Graph showing the comparison of the number of primary outcomes reported in results vs. study data. The graph is a scatter plot with the x-axis labeled as 'Number of primary outcomes - from study data' and the y-axis labeled as 'Number of primary outcomes - from results data.' The data points show a positive correlation between the two measures.]
Number of Secondary Outcomes
Comparison of results vs. study data for interventional trials reporting results
Results Data Reporting in ClinicalTrials.gov: Baseline Measures – variations in reporting

- Age
- BMI
- Body Weight
- Race
- Ethnicity
- Gender
- Region of Enrollment
- Smoking Status
Baseline Measure: Age

Baseline Measure: Gender

Reporting Age

Categories of reporting Age

Reporting Gender

Categories of reporting Gender

Reporting Age, Customized

Categories of reporting Customized Age

Reporting Gender, Customized

Categories of reporting Customized Gender
Baseline Measure: Race/Ethnicity

Reporting Race (NIH/OMB)

Categories of Race reporting per NIH/OMB

Reporting Race/Ethnicity, Customized

Categories of reporting customized Race/Ethnicity
How can I download AACT?

- **Oracle Extracts (Registry + Results)**
  - Oracle dmp
  - Pipe delimited text output
  - SAS CPORT transport

- **Supporting Documents**
  - Comprehensive Data Dictionary
  - High Level Data Dictionary
  - Readmes

- **Points to Consider When Using AACT**

Download Database: CTTI website

ClinicalTrials.gov & AACT: key milestones

ClinicalTrials.gov

- 2000 – Study database launched
- 2005 – ICMJE requirement policy
- 2007 – FDA Amendment Act (FDAAA) enacted
- 2008 – Results reporting included in Protocol Registration System (PRS)
- 2012 – Results database made publicly available

AACT: Aggregate Analysis database of ClinicalTrials.gov

- 2011 – AACT database launched (dataset download: Sep 27, 2010)
- 2012 – AACT Specialty Classification (dataset download: Sep 27, 2010)
- 2013 – AACT-Results launched (dataset download: Sep 27, 2012)
- 2014 – Semi annual updates (dataset download: Mar 27, 2014)
The top 3 most visited pages on the CTTI website are (in descending order):
- The Homepage
- The AACT Database
- The State of Clinical Trials Project (Project that includes AACT)

The average site visitor spends 6 min and 43 sec on the AACT page, which is nearly 5 times more than the other top 10 pages on the CTTI website.

During this timeframe, there were 1,811 page views (1496 unique) on the AACT page.

Number of clicks through to the AACT database zip files – 887
49% of AACT page visitors go the extra step and download the files.
Acknowledgments

Duke University

- Robert Califf (PI)
- Karen Chiswell, Statistician
- Philip D'Almada, Statistician
- Surendra Gonigunta, Programmer
- Skip Maza, Programmer
- Prathima Chintala, Tester
- James Topping, Informaticist
- Sara Calvert (Project Manager)

NLM (ClinicalTrials.gov)

- Deborah Zarin (Director, ClinicalTrials.gov)
- Nick Ide