OPEN ACCESS Freely available online

PLoS one

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

**Asba Tasneem[1]*, Laura Aberle[1], Hari Ananth[1], Swati Chakraborty[1], Karen Chiswell[1], Brian J. McCourt[1], Ricardo Pietrobon[1,2]**

**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.

**Citation:** Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, et al. (2012) The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty. PLoS ONE 7(3): e33677. doi:10.1371/journal.pone.0033677

**Editor:** Joel Joseph Gagnier, University of Michigan, United States of America

**Received** October 14, 2011; **Accepted** February 14, 2012; **Published** March 16, 2012



**Funding:** Financial support for this work was provided by cooperative agreement U19 FD003800 awarded by the U.S. Food and Drug Administration to Duke University in support of the Clinical Trials Transformation Initiative. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: asba.tasneem@duke.edu

## Introduction

ClinicalTrials.gov (www.ClinicalTrials.gov) is a registry of human clinical research studies. It is hosted by the National Library of Medicine (NLM) at the National Institutes of Health (NIH) in collaboration with the U.S. Food and Drug Administration (FDA). As mandated by federal law [1], ClinicalTrials.gov provides a central resource for information about clinical trials; in addition, it increases the public visibility of such research. The registry currently contains over 100,000 research studies conducted in more than 170 countries and is widely used both by medical professionals and the public. New research studies are being submitted to the registry by their respective sponsors (or sponsors' designees) at a rate of approximately 350 per week [2]. Due to legislative [1] and institutional [3] requirements enacted in the latter half of the previous decade, compliance with registry obligations is assumed to be high for U.S. drug and device trials, and the consistency, quality, and maintenance of registry data have improved with increased use [4]. However, the registry has not been optimized for the analysis of aggregate data, and a systematic effort to create and maintain a database for this purpose has not previously been undertaken.

In November 2007, the FDA and Duke University announced the formation of a public-private partnership to improve the quality and efficiency of clinical trials. This collaboration of more than 60 organizations and government agencies was convened by Duke University under a memorandum of understanding with FDA, and is now known as the Clinical Trials Transformation Initiative (CTTI) [5]. CTTI leaders recognized that ClinicalTrials.gov represented a promising source for benchmarking the state of the clinical trials enterprise, as the registry contains studies from the full range of sponsoring organizations. Increasing the usability of ClinicalTrials.gov data may therefore facilitate systematic evaluation of clinical studies aimed at building the

knowledge base needed to inform medical practice and prevention.

As data have accumulated in ClinicalTrials.gov, users have increasingly sought capabilities that would allow aggregated descriptive characterization of the national research portfolio; however, access and data usability issues, including data format and design, present obstacles. A number of related initiatives, including the Ontology of Clinical Research (OCRe) [6], Human Studies Database (HSDB) [7], CDISC Protocol Representation Model [8], and LinkedCT [9] projects, are addressing ontological annotations, large-scale data mining, data representation format, and external association of these data, respectively. The results of this project are complementary to these initiatives and are expected to collectively advance this area of study as a whole.

In this article, we report on CTTI's efforts to prepare and maintain a publicly accessible analysis dataset derived from ClinicalTrials.gov content—the database for aggregate analysis of ClinicalTrials.gov (AACT). We also discuss efforts to extend the utility of the analysis dataset by means of an associated clinical specialty taxonomy designed to support research policy analyses.

## Methods

### 1. Creation of the AACT

Key design features of AACT include 1) the capacity to extend the dataset by parsing existing data; 2) linking to additional data resources, such as the Medical Subject Headings (MeSH) thesaurus; and 3) integrated metadata. A framework for extensions allows entire studies or individual fields to be associated with new data resources while preserving provenance. In addition, the integrated data dictionary developed for this project facilitates browsing and analysis of ClinicalTrials.gov and AACT metadata. Finally, the database incorporates a flexible design that can accommodate future developments, such as coding biospecimen type, sponsors, and OCRe annotations. Figure 1 shows key enhancements achieved by building the AACT.

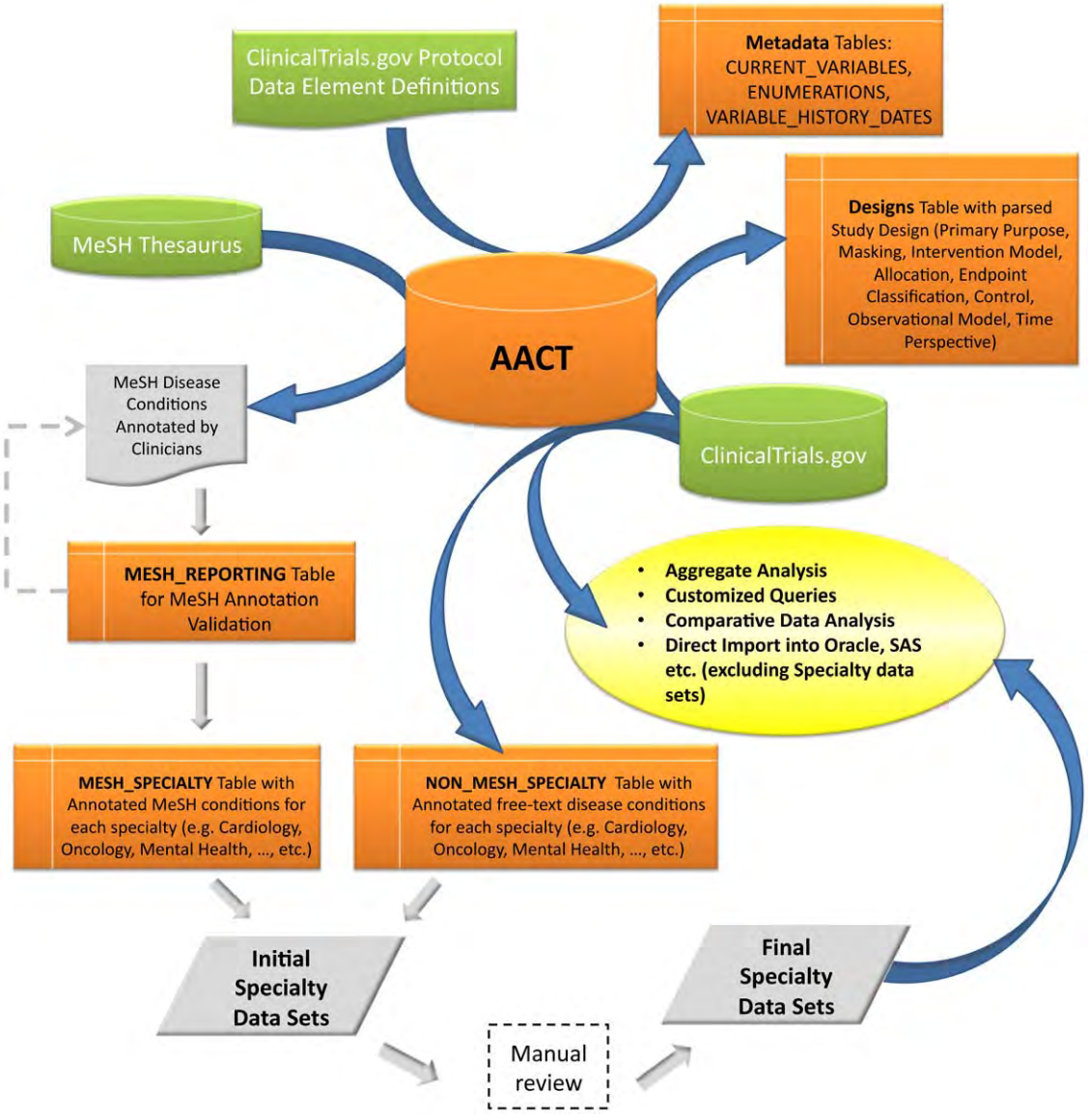


**Figure 1. A schematic representation of the database for Aggregate Analysis of ClinicalTrials.Gov (AACT) with its key enhancements.**
doi:10.1371/journal.pone.0033677.g001

**1.1. Data Sources.** A dataset comprising 96,346 clinical studies was downloaded from ClinicalTrials.gov in XML format on September 27, 2010. We chose ClinicalTrials.gov for our study because it is the largest database of its kind and because it covers the full range of clinical conditions, includes a broad group of trial sponsors [10], and has a regulatory mandate [1]. The date of download was chosen to coincide with the anniversary of the enactment of the FDA Amendments Act (FDAAA) 3 years earlier, which mandated the registration of certain trials of FDA-regulated drugs, biologics, and devices [1].

We downloaded the 2010 MeSH thesaurus (http://www.nlm.nih.gov/mesh/2010/download/termscon.html) and merged it with the AACT database, where it was used as a lookup table to locate corresponding tree numbers, referred to as *MeSH IDs*, for all MeSH terms associated with each clinical trial in ClinicalTrials.gov. Persons or organizations who submit studies to the registry are requested to provide the *condition* and *keyword* data elements as MeSH terms.

**1.2. Data Model.** ClinicalTrials.gov data element definitions, xsd specifications for registry data submission, and downloaded study XML files were used to represent data specifications for the downloaded data. A physical data model was designed using Enterprise Architect (Sparx Systems Pty Ltd, Creswick, Victoria, Australia); this model depicted data tables and their data columns, as well as relationships between and among tables. An optimal structure was achieved through normalization, which was used to organize data efficiently, eliminate redundancy, and ensure logical data dependencies by storing only related data within a given table [11]. The database (Figure 2) was normalized to the Second Normal Form (2NF), a set of criteria designed to prevent logical inconsistencies while reducing data redundancy [12].

We assigned data type and length of data elements based on patterns observed for each data element in the downloaded XML files. Whenever possible, we followed guidelines provided in ClinicalTrials.gov's draft Protocol Data Element Definitions [13] when assigning lengths to given data elements. Data were housed in Oracle RDBMS, version 11.1 g (Oracle Corporation, Redwood Shores, California, USA). Enterprise Architect 7.1 was used for database design and additional transformation rules were documented as extract-transform-load (ETL) specifications. PL/

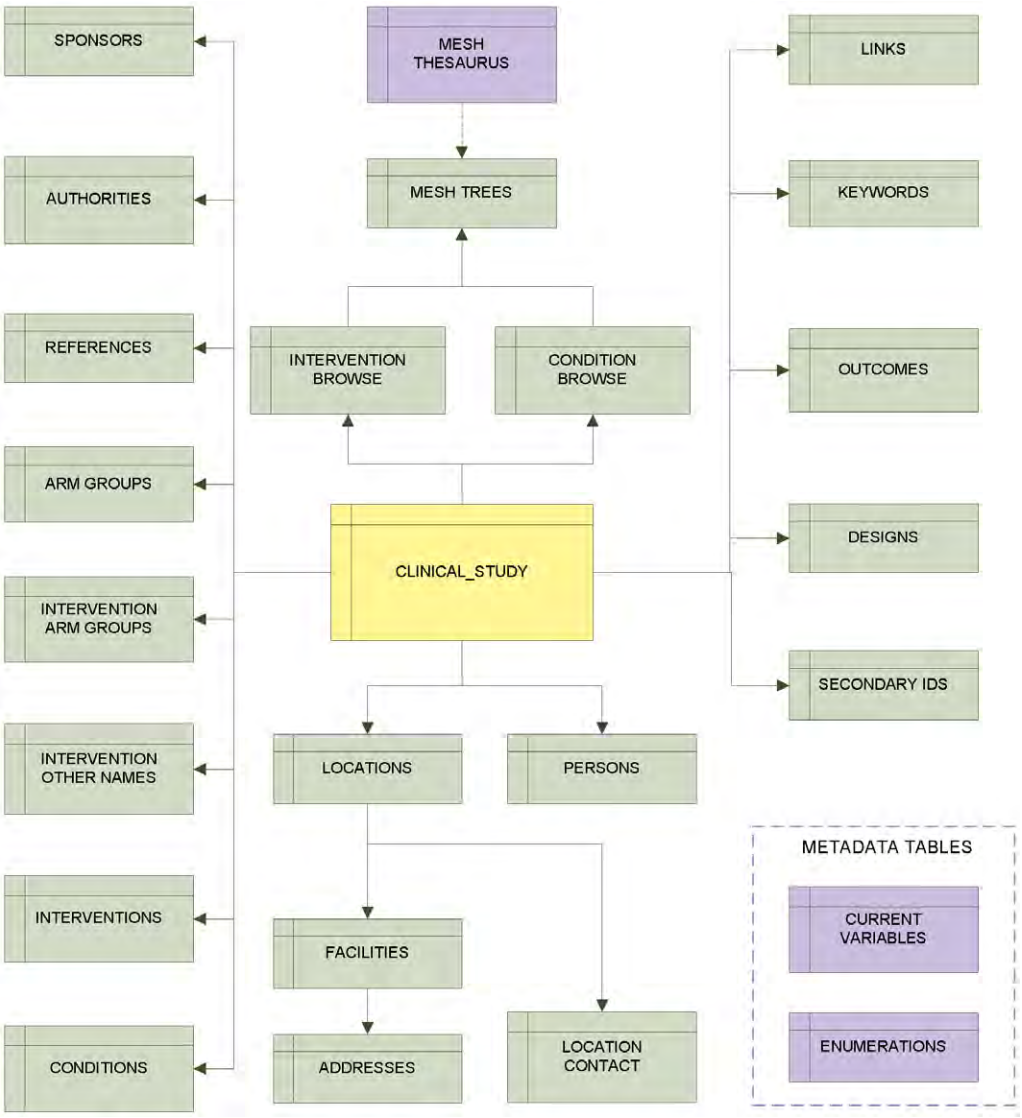


**Figure 2. High-level Entity-Relationship Diagram (ERD) for AACT.**
doi:10.1371/journal.pone.0033677.g002

**Table 1.** Escape characters and replacements.

| Escape character | Replacement |
|---|---|
| ' | ' |
| " | " |
| & | & |
| " | > |
| < | < |

doi:10.1371/journal.pone.0033677.t001

SQL packages that used Oracle's inbuilt DBMS_LOB package to read the input XML files and load the data into the designed tables appropriately were developed. Quality control and operational support processes were developed using standard SQL queries through Toad for Data Analysts (Quest Software, Aliso Viejo, CA, USA) and Cognos ReportNet (CRN) (IBM Corporation, Armonk, NY, USA). We extended the core data model to accommodate both data management and data curation purposes. Error log tables and indexes were created for testing, debugging, and performance enhancement. Manual user acceptance testing was performed by randomly selecting five studies per data element (from a total of 109 data elements) from the AACT database. The values associated with each data element were tested for correctness and completeness by comparing them with the original source data from downloaded XML files. We also created integrated data dictionary tables as reference tables holding explicit data element definitions and system metadata (Tables S1 and S2).

During the course of database development, the NLM made several new data elements available for public download, some of which included information about the FDA (e.g., Section 801 clinical trials, studies with FDA-regulated interventions, and expanded-access studies). In addition to these, MeSH condition and intervention terms generated by the NLM algorithm were also made available for public download.

In XML files downloaded from ClinicalTrials.gov, the single data element *Study Design* contains a string of concatenated values for various different components of a study design, such as primary purpose, interventional model, observational model, allocation, endpoint classification, time perspective, and masking. While this format is well-suited for supporting information retrieval, it does not readily accommodate aggregate data analysis of the components within the *Study Design* data element. For this reason, data from *Study Design* was parsed into its components and stored in a separate table called DESIGNS. Additional data elements (*Design Name* and *Design Value*) were created to store all components of study design and their respective enumerated values. Values related to masking/blinding (e.g., *Single; Double-Blind*) were further parsed into their components, along with the list of corresponding masking subjects (*Participant*, *Investigator*, *Outcome Assessor*, and *Caregiver*).

Several challenges were encountered while loading the database, including foreign characters embedded in XML files

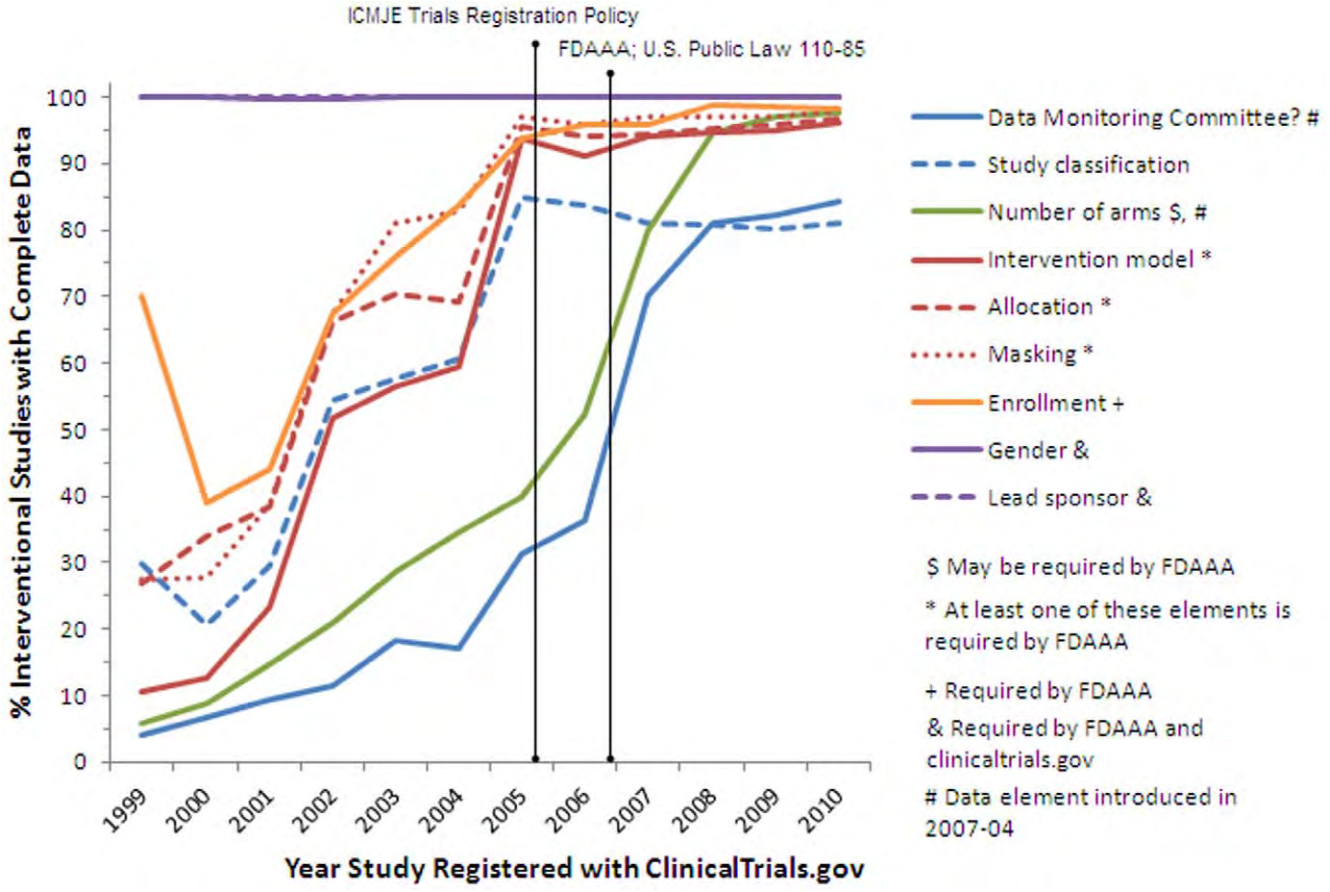


**Figure 3. Percentage of interventional studies with complete data by registration year for selected data elements.**
doi:10.1371/journal.pone.0033677.g003

with most of the data elements; these had to be replaced with character references (see Table 1 for examples).

Other circumstances that prompted several database design iterations included the facts that the maximum length for each data element noted by ClinicalTrials.gov's May 2010 Protocol Data Element Definitions document was not always consistent with the complete dataset, and one-to-one or one-to-many relationships between or among data elements were not obvious in the XML data type definition from ClinicalTrials.gov.

**1.3. Quality Assessment.** Of the 96,346 studies downloaded from ClinicalTrials.gov in September 2010, a total of 79,413 (82.4%) were interventional (i.e., a study in which an investigator following a protocol assigns research participants to receive specific interventions, as opposed to an observational study),

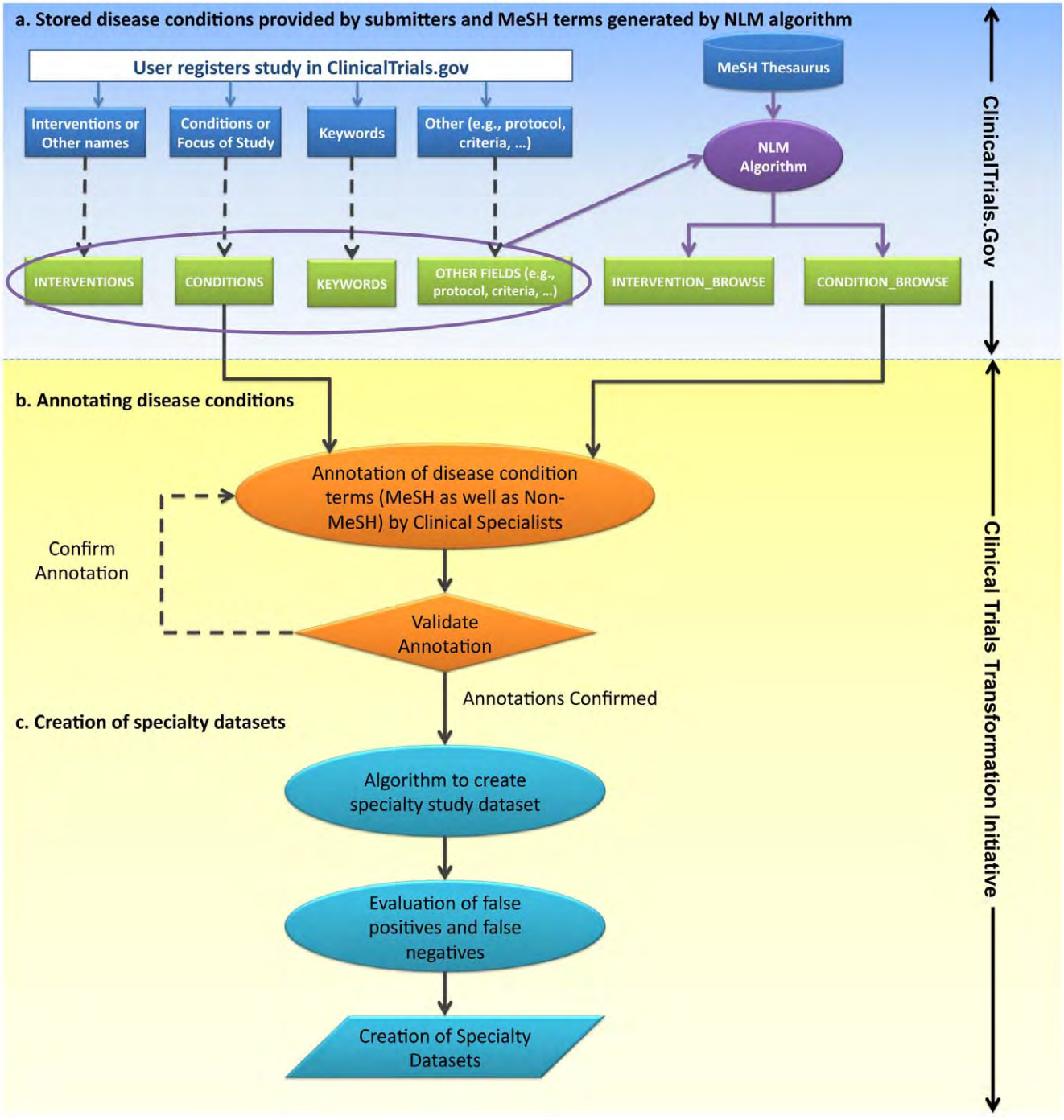


**Figure 4. An overview of methodology and process of developing clinical specialty datasets.** The INTERVENTIONS, CONDITIONS, and KEYWORDS tables consist of disease condition terms provided by data submitters that include both MeSH and non-MeSH terms. The INTERVENTION_BROWSE and CONDITION_BROWSE tables are populated by MeSH terms generated by NLM algorithm (a) Process illustrating how MeSH terms are created in ClinicalTrials.gov. Tables and data shown here does not represent entire ClinicalTrials.gov database (b) Process illustrating the annotation and validation of disease conditions (c) Process illustrating the creation of specialty datasets.
doi:10.1371/journal.pone.0033677.g004

DOCKET ALARM

# Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

## Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

## Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

## Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

## API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

**LAW FIRMS**
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

**FINANCIAL INSTITUTIONS**
Litigation and bankruptcy checks for companies and debtors.

**E-DISCOVERY AND LEGAL VENDORS**
Sync your system to PACER to automate legal marketing.