Points to Consider When Using AACT

This document suggests points for investigators and analysts to consider when planning a statistical analysis of the ClinicalTrials.gov database. It is not intended to be a comprehensive guide.

Population: which studies are likely to be represented in the ClinicalTrials.gov registry?

Virtually any clinical study may be registered at ClinicalTrials.gov. However, the registry is more likely to include certain types of study relative to others. These biases are summarized by Zarin et al. [1]:

> *"… [T]here are undoubtedly trials that are not registered in ClinicalTrials.gov or any other publicly accessible registry. Coverage in ClinicalTrials.gov is likely to be most complete for trials of drugs or devices that are sponsored by U.S.-based or multinational organizations (e.g., major pharmaceutical companies)."*

The ClinicalTrials.gov trial registry was released for the registration of studies on February 29, 2000. The database downloaded by the Clinical Trials Transformation Initiative (CTTI) and the Duke Clinical Research Institute (DCRI) on September 27, 2010 includes 96,346 studies. Of these, 79,413 are interventional studies in which participants are assigned according to a research protocol to receive specific interventions. The registration of studies has been mandated to a large extent by requirements (both legal and institutional) implemented as part of the Food and Drug Administration Amendments Act (FDAAA),as well as by requirements introduced by the International Committee of Medical Journal Editors (ICMJE) and the European Medicines Agency (EMA) regarding registration of clinical studies. Table 1 describes the scope of these requirements.

Table 1: Scope of Interventional Studies Covered by Major Reporting Policies*

Policy Requirements	Registration requirements	Effective date
FDAAA[2]	The following must be registered in ClinicalTrials.gov: • Interventional studies of drugs, biologics, or devices (whether or not approved for marketing) • Studies phases 2 through 4 • Studies with at least 1 U.S. site • Studies conducted under IND/IDE	**September 27, 2007.** Studies initiated after this date, or with a completion date later than 12/25/2007 are subject to FDAAA requirements. Registration is required no later than 21 days after first patient is enrolled.
ICMJE[3]	The following must be registered in ClinicalTrials.gov or other approved registry: • Interventional studies of any intervention type, phase, or geographical location	**July 1, 2005.** Studies initiated after this date must be registered before first patient enrolled; studies initiated before this date must be retrospectively registered to be considered for publication.
EMA[4,5]	The following must be registered in ClinicalTrials.gov or other approved registry: • Interventional studies of drugs or biologics (whether or not approved for marketing) • Phase 1 studies (pediatrics only); • Studies in phases 2 through 4 • Studies taking place in at least 1 European Union site	**May 1, 2004.** EMA launched EudraCT https://eudract.ema.europa.eu/ **March 22, 2011.** The EU Clinical Trials Register was launched by the European Medicines Agency (EMA). (https://www.clinicaltrialsregister.eu/)

** Adapted from [1]. For complete descriptions of policy requirements, see the references cited.*
EMA denotes European Medicines Agency; FDAAA, Food and Drug Administration Amendments Act; ICMJE, International Committee of Medical Journal Editors; IDE, investigational device exemption; IND, investigational new drug application.

Based on these requirements, the following are examples of characteristics that may influence the likelihood that a study is included in the ClinicalTrials.gov registry:

- Interventional studies are more likely to be registered than observational studies.
- Studies that began before the ICMJE requirement in July 2005 are less likely to be registered, especially if their results are unpublished (e.g., negative studies).
- Studies with drug, biological, or device interventions are more likely to be registered than studies of other interventions.
- Studies with at least one site in the United States or European Union are more likely to be registered than studies with no such sites.
- Studies involving a drug or device that is manufactured in the United States are more likely to be registered than studies involving a drug or device manufactured outside of the United States.
- Studies subject to an IND or IDE are more likely to be registered (i.e., if the study is intended to support approval for marketing in the United States).
- Phase 1 adult drug studies or small feasibility studies of devices are less likely to be registered.
- Studies in pediatric populations may be more likely to be registered.

Duplicate records

Because of the quality assurance measures applied by ClinicalTrials.gov staff on registration entries, we can be reasonably certain that each study entered in ClinialTrials.gov refers to a unique clinical study. However there may be a small number of duplicate records within the database [6].

What type of questions can be investigated using the ClinicalTrials.gov data?

The version of the ClinicalTrials.gov database that has been made publicly available through the CTTI and the DCRI contains only study registration records. These describe the study characteristics, including sponsor, disease condition, type of intervention, participant eligibility, anticipated enrollment, study design, locations, and outcome measures. Although ClinicalTrials.gov now also collects summary results and adverse events for studies, these data are not included in the current version of the CTTI/DCRI database.

We anticipate that investigators will use the current database to explore the characteristics of selected subsets of clinical studies (e.g., typical enrollment for a phase 3 study in breast cancer patients), and to compare and contrast these characteristics across different subgroups of studies (e.g., sponsor; device versus drug intervention; or prevention versus treatment). However, researchers will not yet be able to use the data to perform meta-analyses of results or adverse events from clinical studies registered at ClinicalTrials.gov (e.g., to compare the efficacy and safety of different diabetes therapies).

Interpretation of variables

When interpreting the study characteristics collected for a study registered with ClinicalTrials.gov, investigators are encouraged to refer to the data element definitions available at: http://prsinfo.clinicaltrials.gov/definitions.html. Interpretation of a variable may depend on:

- **How the question was phrased.** For example, the definition of "Sponsor" does not necessarily imply that the sponsor is the agency paying for the clinical study, as might be expected from the common use of the term.
- **Whether the respondent can enter a free-text answer to a specific question, or is restricted to a fixed set of possible responses.**
    - Note that the definition of a data element and the available responses may have changed over time. Refer to the change history of variables in the comprehensive data dictionary for details.
- **Whether there is dependence between fields.** Certain data elements need to be interpreted together with a second data element. For example, data elements such as enrollment date and completion date have a companion data element that indicates whether the value in the first field is an anticipated or actual value.
    - Note that the study record may be updated by the owner of the record at any time. Fields such as enrollment type may be changed from anticipated to actual, indicating that the value entered now reflects the actual rather than the planned enrollment. When data are downloaded, the result is a snapshot of the database at that particular time point, and the history of changes made to the field is lost.

Data completeness and accuracy

Some data elements are more likely than others to have missing information, depending on several known factors. For example:

- **The data element being required by the FDAAA and/or the ClinicalTrials.gov website.** Refer to data element definitions and the comprehensive data dictionary for specifics regarding these requirements, as well as for information on when the requirements went into effect for particular data elements.
- **The date when the data element was introduced.** Not all data elements were included in the database at the time of its launch in 2000, but were added later. Studies registered after FDAAA when into effect must meet more requirements than studies registered earlier in the life of ClinicalTrials.gov.
- **The branching structure of questions.** The availability of certain questions to the person registering depends on answers to previous questions. For example, questions about bio-specimen retention are only available for observational studies. Therefore, interventional studies should be excluded when analyzing data elements pertaining to bio-specimens.
- **The list of possible answers for data elements with a fixed set of responses.** For example, questions that include "N/A" as a possible response are likely to have fewer missing values than questions that do not provide a "N/A" response.

"Missingness" of data may also depend on other unknown factors. However, regardless of the cause of missing data, users of ClinicalTrials.gov datasets are encouraged to specify clearly how missing values and "N/A" values are handled in their analysis. For example, are studies with missing values excluded from statistics summarizing that data element, or are they included? In some cases, missing values may be imputed based on other fields (e.g., if a study has a single arm, it cannot employ a randomized design). In other cases, a sensitivity analysis may be appropriate for exploring the effect of different assumptions about the missing values on analysis results.

Although the FDAAA and other requirements do not apply to all fields in the database, users might consider including only studies registered post-FDAAA (September 2007). This will help to limit the number of missing values across many data elements. Users could also consider annotating data elements used in analysis according to whether or not they are FDAAA-required fields, if the user believes this might affect the extent of missing data.

Even when the data elements for a particular study are complete, users are cautioned to have modest expectations about the accuracy of the data. As described by Zarin and colleagues [1], ClinicalTrials.gov has implemented several measures to assure data quality. For example, staff applies automated business rules that

alert providers when required data are missing or are internally inconsistent. In addition, some manual review is also performed, and a record may be returned to the data provider if revision is required. However, ClinicalTrials.gov staff cannot always validate the accuracy of submitted data (e.g., against an independent source). As Zarin et al. note, "… individual record review has inherent limitations, and posting does not guarantee that the record is fully compliant with either ClinialTrials.gov or legal requirements" [1].

During our own analysis of the ClinicalTrials.gov database, several unrealistic values for numeric data elements were encountered, such as an anticipated enrollment of several million subjects. When aggregate summaries of numeric data are provided, analysts are encouraged to use measures that are robust to outliers, such as medians and interquartile ranges, rather than measures such as means ± SD, which could be strongly influenced by unusually large or small values. Users may also wish to run their own consistency checks (e.g., to compare whether the number of arm descriptions provided for the study matches the data element that quantifies the number of arms in the study design).

***Missing information in LOCATIONS, FACILITIES, and ADDRESSES data sets**:*

In general, the presence of a record in a dataset indicates that information was submitted to ClinicalTrials.gov for at least one variable in that data set. However, in Version 2.0 of the posted data, the LOCATIONS, FACILITIES, and ADDRESSES datasets contain at least one record for each study (each nct_id), even for studies that did not submit any information on facilities or locations to ClinicalTrials.gov. To identify these studies:

- Merge the LOCATIONS data set with the FACILITIES data set by the nct_id and location_id keys.
- Merge this merged set with the ADDRESSES data set by the nct_id and facility_id keys.

Records that are missing values for the variables listed below did not submit locations or facilities information to ClinicalTrials.gov and should therefore be deleted from these datasets:

- status
- name
- city
- zip
- state
- country

Data elements in AACT that were not part of the original ClinicalTrials.Gov dataset download and how to differentiate them from the source data elements

There are additional variables in the AACT database that were not part of the original download. The source of each variable in the AACT database is indicated in the Comprehensive Data Dictionary (Current_Variables tab, 'Variable_Source' column). All those variables that indicate 'NLM' were found in the NLM's DTD file; others that indicate 'CTTI' were created in the AACT database. Examples of some of these new data elements include primary keys, parsed data fields (e.g. 'study design'), formatted download date field, and MeSH IDs. A brief description for all those variables is provided under 'CTTI Notes' column of the Data Dictionary.

Integrated MeSH thesaurus in the AACT Database

The 2010 version of MeSH thesaurus has been integrated in the AACT database. There are 25,586 descriptors in 2010 MeSH; each with one or more tree numbers or MeSH IDs. In the AACT database, MeSH thesaurus is available in the MESH_TREES_RAW table. For information on using and downloading MeSH thesaurus, please visit http://www.nlm.nih.gov/mesh/.

MeSH terms from ClinicalTrials.Gov dataset

The data submitters (to ClinicalTrials.gov database) can provide MeSH as well as free-text terms in the disease conditions, interventions, and keywords fields. Data from these fields are stored in the CONDITIONS, INTERVENTIONS and KEYWORDS tables respectively in the AACT database. Additionally, an NLM algorithm also evaluates studies and applies MeSH terms studies. These annotated MeSH terms are populated in the condition_browse and intervention_browse fields in the ClinicalTrials.gov database. In the AACT database, these MeSH terms are available in the CONDITION_BROWSE and INTERVENTION_BROWSE tables. The MeSH terms in the dataset are not specified with any MeSH ID or hierarchy. This association with the MeSH thesaurus can only be done by matching the term itself to any entries in the thesaurus. Such a list of MeSH terms with all possible MeSH ID matches extracted from MeSH thesaurus (MESH_TREES_RAW table) is stored in the MESH_TREES table. There are some orphan MeSH terms in the CONDITION_BROWSE (20 terms) or INTERVENTION_BROWSE (918 terms) that were not found in the 2010 MeSH Thesaurus were excluded from the MESH_TREES table. These orphan MeSH terms existed in the older versions of the MeSH thesaurus and were previously assigned by NLM algorithm to ClinicalTrials.Gov studies. Additionally, MeSH terms in CONDITION_BROWSE (16 terms) or INTERVENTION_BROWSE (12 terms) with apostrophe character as well as those with variance in capitalization were unintentionally excluded from the MESH_TREES table.

References

1. Zarin, D. A., Tse, T. T., Williams, R. J., Califf, R. M., and Ide, N. C. (2011). The ClinicalTrials.gov results database – update and key issues. *N Engl J Med* 364: 852–60.
2. Food and Drug Administration Amendments Act of 2007. Public Law 110-95.
3. Laine, C., Horton R., DeAngelis C.D., et al. Clinical trial registration – looking back and moving ahead. *N Engl J Med* 356: 2734–6.
4. Communication from the Commission regarding the guideline on the data fields contained in the clinical trials database provided for in Article 11 of Directive 2001/20/EC to be included in the database on medicinal products provided for in Article 57 or Regulation (EC) No 726/2004. In: European Commission, ed. Official Journal of the European Union, 2008. (2008/C 168/02.)
5. Guidance on the information concerning paediatric clinical trials to be entered into the EU Database on Clinical Trials (EudraCT) and on the information to be made public by the European Medicines Agency (EMEA), in accordance with Article 41 of Regulation (EC) No 1901/2006. In: European Commission, ed. Official Journal of the European Union, 2009. (2009/C 28/01.)
6. Zarin, D. A., Ide, N. C., Tse, T. et al. (2007). Issues in the registration of clinical trials. *JAMA* 297: 2112—2120.