Identification of Common Data Elements from Pivotal FDA Trials

Craig S. Mayer, MS[1], Nick Williams, Ph.D[1], Vojtech Huser MD, Ph.D[1]
[1]Lister Hill National Center for Biomedical Communication, National Library of Medicine, NIH Bethesda, MD

Abstract

It is difficult to arrive at an efficient and widely acceptable set of common data elements (CDEs). Trial outcomes, as defined in a clinical trial registry, offer a large set of elements to analyze. However, all clinical trial outcomes is an overwhelming amount of information. One way to reduce this amount of data to a usable volume is to only use a subset of trials. Our method uses a subset of trials by considering trials that support drug approval (pivotal trials) by Food and Drug Administration. We identified a set of pivotal trials from FDA drug approval documents and used primary outcomes data for these trials to identify a set of important CDEs. We identified 76 CDEs out of a set of 172 data elements from 192 pivotal trials for 100 drugs. This set of CDEs, grouped by medical condition, can be considered as containing the most significant data elements.

Introduction

In recent decades, requirements to share data from completed human clinical trials has been increasingly adopted by many sponsors of clinical research.[1,2] When data is shared, all data elements used by a trial are typically included in the bundle of shared data files and other metadata artifacts for a trial.[3] A data dictionary file that lists all data elements (DEs) used by a given trial is often one of the required pieces of metadata for a trial. If the same DE is used in multiple trials, it can be regarded as a Common Data Element (CDE).[4] DEs used in prior completed trials can inform greatly the development and use of CDEs. Identification of a useful set of CDEs is an important goal (and an ongoing challenge) for informatics advancement of research enterprise.[5] They can aid in the efficient analysis and effective comparison or harmonization of multiple clinical datasets and are imperative in maximizing the reusability of clinical trial data. Several ongoing CDE initiatives (such as PhenX, PROMIS, CDEs from National Institute of Neurological Disorders and Stroke) define and promote a list of CDEs. CDEs are crucial for increased harmonization of data collected across many trials.

Several clinical research informatics analyses of DEs from individual interventional trials or observational studies were published in the past. When DEs are aggregated across many trials, the number of elements can be overwhelmingly high. One study analyzed a subset of 1 414 DEs from all 24 938 PhenX DEs found in 426 studies.[4] Another study looked at 75 CDEs across a sample of 17 European clinical studies in acute or chronic diseases in oncology, neurology, diabetes, cardiovascular and inflammatory diseases.[6] Due to this volume of available CDEs, prioritization and identification of the most significant data elements is a challenging task. Our motivation was to use computational methods to identify important DEs and if they are used in multiple trials, generate a list of significant CDEs.

One source of data elements for a large number of trials is the ClinicalTrials.gov (CTG) registry. CTG collects primary and secondary outcomes for included studies. As of February 12, 2020, CTG contained 303 267 outcomes (86 177 primary, 217 090 secondary). This staggeringly high number of outcomes is too rich for efficient and accurate analysis and can be reduced by only looking at a subset of trials. Our assumption was to focus on DEs of trials that drive regulatory approval of new drugs and devices (pivotal trials). Drug approval in most countries (e.g., US, EU member countries, Japan or Australia) rely on showing drug efficacy (among other important considerations). Primary outcomes of Phase 3 trials thus contain crucial DE information. For example, for approval of the drug Seysara (sarecycline, for acne), the pivotal trial NCT02320149 had the outcome ‘Absolute Change in Facial Inflammatory Lesion Counts’. In the US, endpoints important for regulatory context are also listed in FDA guidance documents. For example, FDA guidance titled ‘Uncomplicated Gonorrhea: Developing Drugs for Treatment: Guidance for Industry’ in section ‘Efficacy Endpoints’ recommends the following endpoint: ‘establishment of a negative culture at the infection site or sites approximately 3 to 7 days after receipt of antibacterial drug therapy (microbiological cure)’.[7]

In this study, we analyzed Food and Drug Administration (FDA) approval documents (namely Drug Trial Snapshots) and linked approved drugs to their pivotal trials, whose outcomes we used to identify a set of CDEs. These CDEs we view as a set of the most important CDEs to be considered by various ongoing or future CDE consensus initiatives.

Methods

This project consisted of two components. In Part I, we operate on the drug and trial level and establish a set of pivotal trials by extracting pivotal trial identifiers in the Drug Trial Snapshots of FDA approved drugs. In Part II, we use this set of pivotal trials as input to extract data elements used in these trials (using the CTG registry). Specifically, we map the textual description of a primary outcome into a formal data element using a manual annotation approach. For example, this is seen for the condition of cystic fibrosis, where there are two drugs, five pivotal trials identified and all five trials share an outcome data element (or CDE) of percent predicted Forced Expiratory Volume in 1 Second (ppFEV1). Using this process, we arrive at important CDEs (naturally organized by condition) for tens of medical conditions covered by our set of pivotal trials.

Part I: Pivotal Trials

Set of pivotal trials

Because pivotal trials are determined by a drug regulatory agency (in the US, the US Food and Drug Administration; FDA), we used FDA resources to identify pivotal trials. A pivotal trial is defined as a study to collect definitive evidence of the safety and efficacy of a drug for an intended use.[8] Pivotal trials are typically Phase 3 trials for drugs that have already shown safety and effectiveness in data collected in prior studies and contribute to regulatory approval for that drug for the defined use.

Since 2015, FDA started publishing snapshots of trials for each approved drug (technically, a New Drug Application [NDA]). The Drug Trial Snapshots provide concise information to the public about the drug, as well as about who participated in the clinical trials leading to the drug's approval. The FDA website provides a listing of 234 Drug Trial Snapshots as of January 31, 2020. Key columns included: drug, date of FDA approval and link to Drug Trial Snapshot.

We generated a list of approved drugs via the FDA Drug Trial Snapshots found at https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots.[9] For example, in 2019, a trial snapshot for a new drug application for Accrufer (ferric maltol; treatment of low iron stores) was created and contains a section 'Who participated in the clinical trials?' that lists the three pivotal trials and their CTG trial identifiers (NCT IDs): NCT01252221, NCT01340872, and NCT02968368.

To computationally arrive at a list of pivotal trials, we thus developed an R script to extract CTG trial identifiers from the Drug Trial Snapshots. The script then scraped the page and returned all NCT IDs present in each Drug Trial Snapshot and the drug the trial came from (list of pivotal trials). We excluded all drugs that did not return any results, as identifying a trial from anything other than the NCT ID would require manual review of all snapshots and would be difficult to property identify and link to CTG data possibly leading to inaccurate identification of the stated trial. A preliminary analysis showed that for drugs approved prior to 2017, there were none with NCT IDs, for drugs approved in 2017, only 15 of 47, 31.9%, of FDA drug snapshots contained NCT IDs, while in 2018 the percentage of Drug Trial Snapshots with NCT IDs was much higher at 95.1%, 58 of 61 snapshots. Considering this milestone, we arrived at the final set of analyzed pivotal trials by limiting our search to drugs approved on or after January 1, 2018. The parsing was done on January 31, 2020.

Number of trials

We analyzed how many pivotal trials per drug we had in the final pivotal trial sample. Having multiple trials per drug and multiple trials per disease is important in our effort to arrive at CDEs. For a data element to qualify as a CDE, it must be present in at least two trials (otherwise it is sometimes referred to as a *unique* data element; a logical opposite of *common* data element).

Matching trials from Drug Trial Snapshots to ClinicalTrials.gov records

We took the list of NCT IDs identified from the Drug Trial Snapshot and linked them to the CTG records associated with them. To do this we used the Aggregate Analysis of ClincalTrials.gov (AACT) database published and maintained by Duke University. It is a relational version of CTG data that is created by parsing the XML representation of each study record found on CTG.[10]

From pilot experimentation of our planned methodology, we noticed that in few rare cases the Drug Trial Snapshot extracted NCT IDs did not link to a valid CTG identifier. In other words, the FDA Drug Trial Snapshots may (in rare instances) contain a typo in the identifiers.

Next, we wanted to group trials not only by drug (by being mentioned in the same FDA snapshot document for a single approved drug), but also by medical condition. To achieve this, we used the CTG condition field by taking the Medical Subject Heading (MeSH) term associated with the trials CTG record and grouped the trials by condition. MeSH terminology is a controlled vocabulary of medical conditions used primarily by PubMed. It is, however, also used by CTG to link trials to medical conditions and to possibly aggregate similar condition entries into higher-level groupings. In CTG, the MeSH term is the primary condition or main focus of the study.[11] Each trial can have one or multiple MeSH terms associated with it. For example, NCT02475629 has only one MeSH term, 'HIV infections', which is the main focus of the trial, while NCT01800045 has three MeSH terms of focus, 'sleepiness', 'narcolepsy', and 'cataplexy'. We did not do any hierarchical grouping of related MeSH terms. We excluded any condition that only had one trial. This exclusion was done because for conditions with just a single trial, the DEs from such trial could never qualify as CDEs (since for a CDE at least two trials must use the DE). We also excluded general MeSH terms because the granularity of such terms is too high for our purpose of identifying significant CDEs by disease. The excluded general MeSH terms were 'disease', 'infection' and 'syndrome'. In addition, if two or more closely related MeSH terms were all present in the exact same set of trials and no others, we only considered the more granular term. For example, for arthritis and rheumatoid arthritis, we further considered only rheumatoid arthritis, as these two terms both appeared in the same manner across all nine relevant trials.

Part II: Data Elements

Trial Outcomes

The CTG registry collects primary and secondary outcomes in trial registration data. Since we were interested in the most important data elements, we only analyzed primary outcomes. Using the previously mentioned AACT database, we retrieved all primary outcomes for our set of pivotal trials.[10] Each trial can list one or multiple primary outcomes for the trial. For example, pivotal trial NCT01866111 for drug Xcopri (cenobamate, for treatment of partial-onset seizures), listed 'Percent reduction in seizure frequency of complex partial and/or secondarily generalized and/or simple partial motor seizures relative to the baseline' as the only primary outcome. Alternatively, pivotal trial NCT02348619 for the drug Sunosi (solriamfetol, for treatment of narcolepsy or obstructive sleep apnea) has two primary outcomes, 'Change in the Epworth Sleepiness Scale (ESS)' and 'Change in the Maintenance of Wakefulness Test (MWT)'. Primary outcomes (as specified in CTG), however, do not yet fully correspond to data elements, and therefore require further processing of the outcome text.

Annotation of primary outcomes

To identify DEs we annotated each outcome by looking at the title field in the CTG outcomes table. If not enough information for annotation was present in this field, we expanded our review to include the description field as well. For example, pivotal trial NCT01868997 for the drug Tepezza (teprotumumab-trbw; for the treatment of thyroid eye) had outcome 'Responder Status at Week 24', which is not detailed enough to find a useful concept, so we expanded our review to include the outcome description. The description was 'Responders were defined as participants with a reduction in clinical activity score (CAS) of > 2 points, and a reduction in proptosis of >2 mm in the study eye, and no deterioration (increase in CAS or increase in proptosis) in the non-study eye'.

We annotated each outcome, by dividing it into two categories: concept and modifier. The concept is the principal endpoint or main focus of the primary outcome and states what is being measured or observed. For example, the concept for outcome 'Change in Hb concentration from baseline to Week 16' in trial NCT02968368 is 'Hemoglobin'. In some cases, there were multiple concepts within one outcome; in these cases the outcome was annotated twice, once for each concept. An example of a two concept annotation is the outcome 'Percentage of Participants With an Elevated Aspartate Aminotransferase (AST) or Alanine Aminotransferase (ALT) Laboratory Value That Was > 5 Times the Upper Limit of Normal (ULN)' from NCT01505634 which has the concepts Aspartate Aminotransferase and Alanine Aminotransferase. Table 1 shows examples of complete annotations.

Table 1. Example outcome annotation.

Outcomes	Concept	Mod: Study-Level	Mod: Patient	Mod: Time	Mod: Value
Percentage of Participants With HIV-1 RNA<50 Copies/mL at Week 48	HIV-1 RNA	Percentage/Proportion		Week 48	<50 copies/mL
Number of Participants With Stable Platelet Response (Count of >50,000/µL Between Weeks 14 and 24)	Platelet	Number	Stable Response	Week 14 to 24	50,000/µL
Percentage of Participants Achieving 90% Improvement i (PASI90) at Week 16	PASI Score	Percentage/Proportion	Improvement	Week 44	0.9

The modifier category was created based on multiple flavors of semantically similar main concepts that were observed during a pilot annotation done by our team. The modifier has four types: 1) study-level modifier, 2) patient-level modifier, 3) time modifier, and 4) value modifier. A *study-level modifier* is a statement of how the outcome is evaluated for the overall study. An example of this is in trial NCT01225731 that has the primary outcome of 'Percentage of Participants With a Psoriasis Area and Severity Index (PASI) 75 Response' and therefore has a study-level modifier of 'percentage of participants'. A *patient-level modifier* states how the outcome is evaluated pertaining to the individual patient. A *patient-level modifier* is similar to the *study-level modifier* but operates on the patient level instead. For example, in the trial NCT02614183 the primary outcome 'Change From Baseline in the Number of Monthly Migraine Headache Days' is analyzed based on the 'change from baseline' for each participant. A *time modifier* states when the outcome will be evaluated. For NCT02706873 and primary outcome 'Percentage of Participants With an American College of Rheumatology 50% (ACR50) Response at Week 12', the timeframe is '12 weeks'. Finally, the *value modifier* states a specific value (or threshold) that is provided for a given concept within the defined primary outcome. For example, in trial NCT02526160 and outcome 'Percentage of Participants Achieving Mean Serum Phosphorus Levels Above 2.5 mg/dL [0.81 mmol/L]', the value modifier is 'greater than 2.5 mg/dL' for serum phosphorus.

Identification of CDEs

We took each concept identified from the pivotal trial outcomes and determined whether it was a unique DE or a CDE. We defined a CDE as a DE that appears in more than one pivotal trial for a given condition, while a unique DE is a concept that only appears in one pivotal trial for a condition. To identify whether two concepts were the same, we semantically mapped all of the concepts manually to show equivalency between synonymous terms, such as HIV-1 RNA and Viral Load which both mapped to HIV-1 RNA. This semantic mapping was solely done manually using expert assessment and mapping by a single reviewer. For DE identification, we excluded concepts for adverse events and other similar concepts, since our focus was on efficacy data elements that drive the drug regulatory approval (and not safety data elements).

Once all DEs were identified, to asses various types, we categorized them into the following categories: lab measurement (e.g., mean corpuscular hemoglobin concentration), non-lab measurement (e.g., weight), clinician observation, symptom, questionnaire, condition-related event or mortality-related event.

Modifiers

We also analyzed the modifiers from the annotated outcomes and identified the most commonly used modifiers for pivotal trial outcomes. The modifiers provide information on how the concepts are evaluated and show not just commonness on a concept level, but also commonness in context. In CDE evaluation, context is valuable in assessing not just whether a concept is common, but also if the way the CDE is being used and evaluated is comparable between multiple trials. For example, the fact that all relevant trials use identical value modifiers or time modifiers (in addition to an identical concept).

Results

Part I: Pivotal Trials

Set of analyzed drugs

Using the spreadsheet list at the FDA website, we identified Drug Trial Snapshots for 107 drugs that were approved on or after January 1, 2018. FDA publishes one snapshot document per approved drug. For seven approved drugs, the corresponding snapshot did not contain any trial NCT IDs and we removed those drugs from our analysis. Thus, our final sample consisted of 100 analyzed drugs for which we had at least one listed pivotal trial NCT ID.

Set of Pivotal Trials

When counting the number of distinct pivotal trials, we identified 192 pivotal trials with NCT IDs from the 100 drugs. Most snapshots correctly listed NCT IDs for all pivotal trials for that drug. However, there were also some Drug Trial Snapshots that included both trials that had an NCT ID and trials that were listed using a trial acronym, but omitted any NCT ID. In the set of pivotal trials, we included the trials with NCT IDs, but did not attempt to identify the trials listed in ways other than by NCT ID (trial listed by acronym but missing trial id phenomenon). The reason for this was that matching a trial to a correct NCT ID using only an acronym (and the investigative drug and possibly the sponsor) can be ambiguous.

Number of trials

Table 2 lists how many pivotal trials were found in our set of 100 analyzed drugs with NCT IDs. For 55 drugs in our sample, we found two or more pivotal trials. The average number of trials per drug was 1.92. While the FDA rule typically requires at least two pivotal trials for each drug, the regulation has provisions for situations where approval can be based on one adequate pivotal trial. A single pivotal trial was listed for 45 drugs. The counts provided in Table 2 of trials per drug are based solely on the presence of NCT IDs in the Drug Trial Snapshot and not the actual counts of trials for each drug (see missing trial id phenomenon explanation above). The highest number of trials for a drug was six (for drug Motegrity).

Table 2. Count of drugs based on trial count for each drug.

Number of trials (with NCT ID listed in snapshot)	Count of Drugs
1	45
2	31
3	16
4	4
5	3
6	1

Matching trials from Drug Trial Snapshots to ClinicalTrials.gov records

In the initial extraction, we started with 192 NCT IDs. In five instances, an NCT ID with no discernable errors was the only trial identifier provided in the Drug Trial Snapshot, but it did not link to any valid CTG record and the trial was unable to be properly identified. This reduced the number of analyzed pivotal trials to 187.

After full linkage of pivotal trials to CTG registry, we were able to take advantage of the rich set of CTG metadata about each trial. Grouping of pivotal trials by disease MeSH term revealed 90 conditions that had multiple trials. The conditions with the most trials were ‘neoplasms’ (15 trials), ‘migraine disorders’ (12 trials), and ‘arthritis, rheumatoid’ (9 trials). Using grouping by condition, we were able to identify key CDEs significant for the study of a given condition.

The analysis of phase of pivotal trials is shown in Table 3. One expectation we had was that a vast majority of pivotal trials would be Phase 3 that per definition focus on proving drug effectiveness. This hypothesis was confirmed as 135 out of 187, 72.2%, of the trials were Phase 3. On the other hand, we observed that 9 pivotal trials (4.8%) were Phase

1. In some cases, this could be explained as the Phase 1 trial was just one of multiple trials associated with the drug approval. In two cases, the Phase 1 trial listed was the only trial for that drug. Those cases are: (1) for drug Ayvakit (avapritinib) where the only trial listed was NCT02508532, '(NAVIGATOR) Study of BLU-285 in Patients with Gastrointestinal Stromal Tumors (GIST) and Other Relapsed and Refractory Solid Tumors'; and (2) for drug Tibsovo (ivosidenib) the only trial was NCT02074839, 'Study of Orally Administered AG-120 in Subjects With Advanced Hematologic Malignancies With an IDH1 Mutation'. Table 3 shows the count and percentage of trials for each phase.

Table 3. Trial count by phase.

Phase	Trial Count	% of total trials
Phase 3	135	72.2
Phase 2	33	17.7
Phase 1	9	4.80
Phase 1/Phase 2	5	2.70
Phase 2/Phase 3	3	1.60.
Phase 4	1	0.50
N/A	1	0.5

Part II: Data Elements

Trial Outcomes

Our exclusion criteria removed any MeSH term with fewer than two trials, any vaguely formulated outcome with not enough detail, or any outcome related to adverse events. These criteria removed 88 outcomes, reducing the number of annotated outcomes from 303 to 215.

Annotation of primary outcomes

Manual annotation of outcomes (and descriptions) merged similar outcomes into outcome concepts (or outcome DEs). We found 97 distinct outcome concepts (regardless of condition) from the 215 total outcomes. For multiple reasons, we chose to analyze condition-DE pairs rather than isolated DEs. When paired with a condition, we analyzed 172 condition-DE pairs. 75 DEs appeared in multiple conditions. For example, 'Progression-Free survival' was a DE for 'breast neoplasms', 'leukemia', and 'lymphoma'.

Identification of CDEs

Next, we took the 172 DE and condition pairs and identified which DEs appear in multiple trials. We found that 76, 44.2%, of the DEs were CDEs within a condition (present in 2+ trials). This promotion of a DE into CDE status is a crucial goal and result of our method. We further labeled 38 of them (50.0% of all 76 CDEs) as unanimous CDEs. *Unanimous CDEs* are CDEs that are present in all pivotal trials for a given condition. For example, HIV-1 RNA is present in all 7 trials for HIV infections. For the majority of conditions analyzed (49 out of 90 conditions; 54.4%) we identified at least one CDE. Table 4 shows DEs for a subset of the conditions studied. The subset shown in the table was purposely designed to show a range of examples: highly common as well as a few unique DEs (marked with a star). The complete list of all 172 DEs by condition is available at the project repository at https://github.com/lhncbc/CDE/tree/master/pivotal.

Table 4. Subset of conditions and their CDEs or unique DEs(*).

Condition	DE from Outcome	Drug Count	Count of Trials Using DE	Total Count of Trials for Condition	Percentage of Trials Using DE
Migraine Disorders	Headache Pain	4	4	12	33.3%
Migraine Disorders	Monthly Migraine Days	5	8	12	66.7%
Migraine Disorders	Most Bothersome Symptom	4	4	12	33.3%
Constipation	Overall Response Rate	2	2	8	25.0%
Constipation	Spontaneous Complete Bowel Movements (SCBM)	2	6	8	75.0%
Psoriasis	Physician's Global Assessment (PGA)	2	6	8	75.0%
Psoriasis	Psoriasis Area and Severity Index (PASI) Score	2	8	8	100.0%
HIV Infections	HIV-1 RNA	3	7	7	100.0%
HIV Infections	Virologic Failure*	1	1*	7	14.3%
Cystic Fibrosis	Percent Predicted Forced Expiratory Volume in 1 Second (ppFEV1)	2	5	5	100.0%
Amyloidosis	All-Cause Mortality*	1	1*	3	33.3%
Amyloidosis	Cardiovascular-Related Hospitalizations*	1	1*	3	33.3%
Amyloidosis	Modified Neuropathy Impairment Score +7	2	2	3	66.7%
Amyloidosis	Norfolk Quality Of Life Diabetic Neuropathy (QoL-DN) Questionnaire*	1	1*	3	33.3%

The most unanimous CDE (with the most trials) was the 'American College of Rheumatology Response Criteria', which was an outcome for all 9 rheumatoid arthritis trials. While most conditions only had one CDE, 31 of 49 conditions (63.3%), other conditions had multiple CDEs. This shows that a condition can be evaluated using multiple measures. This is seen in psoriasis with two highly common CDEs: 'Psoriasis Area and Severity Index (PASI) Score' (8 trials) and 'Physician's Global Assessment (PGA)' (6 trials) were highly common within the 8 total psoriasis trials. We also observed that existence of CDEs for a condition was not necessarily impacted by the number of drugs for that condition. For the condition with the highest number of drugs, neoplasm (8 drugs), 4 of 10 DEs (40.0%) were CDEs, while for the condition with the second highest count of drugs, migraine disorders (5 drugs), all three DEs were CDEs. Some identified CDEs were too general and lacked enough descriptiveness to be of real value. Examples of these general CDEs were 'Overall Response Rate' and 'Clinical Cure'. The definition of many general CDEs is condition specific.

Analysis of data elements

We classified all 172 DEs into types and found that more than half were lab measurements (88, 51.2%). Table 5 lists the counts of DEs, CDEs and the CDE percentage for each type. When it comes to CDEs, the highest number of CDEs is from lab tests with 31, but that is only 35.2% of all lab test DEs making it the lowest proportion of all types. Questionnaire had the highest proportion of CDEs to total DEs within the type as there were 5 CDEs to 7 total DEs (71.4%).

Table 5. Count of DEs and CDEs by type.

Type	DE-Count	CDE-Count	% CDE
Condition-Related Event	16	8	50.0
Lab Test	88	31	35.2
Mortality	4	2	50.0
non-Lab Measurement	32	13	40.6
Clinician Observation	13	9	69.2
Questionnaire	7	5	71.4
Symptom	12	8	66.7

Modifiers

During annotation, modifiers proved to be a useful construct in semantically merging similar outcomes. Modifiers were used in 85.1% of all outcomes. For some outcome concepts, such as 'Overall Survival', no modifiers were needed. The results specific to each of the four types of modifiers we defined are described below.

Patient-Level: Patient level modifiers were found for 117 of the 215 distinct annotated outcomes (54.4%). We found 19 different patient-level modifiers that appeared in at least two of the 215 outcomes. The most commonly used ones were 'Change From Baseline' (found in 46 outcomes), 'Change' (appearing in 9), and Absence (appearing in 8). Table 6 shows the complete list of patient level modifiers found and the count of outcomes in which they appeared.

Table 6. Patient-level modifiers by outcomes count.

Modifier: Patient-Level	Outcome Count
Change From Baseline	46
Change	9
Absence	8
Improvement	8
Reduction	7
Clear or Almost Clear	6
Elevated	4
Reduction from Baseline	4
Did Not Require	3
Duration	3
Early	3
Favorable	3
Proportion	2
Rate of Change	2
Stable Response	2
Time to Last	2
Time-to-response	2

Study-Level: Of the 215 distinct outcomes analyzed, 90 (41.9%) had a study-level modifier. These 90 outcomes consisted of only two different study-level modifiers, 'number of participants' (5 outcomes, 5.6% of outcomes with study-level modifiers), and 'percentage/proportion of participants' (85 outcomes, 94.4%). The study-level modifier is useful since a DE that would include count or percentage in its definition, would be difficult to assess at a single

patient level. For example, for status evaluation of a single patient (in an Electronic Health Record [EHR] setting), the data element 'percentage of participants with virologic failure' makes no sense.

Value: There were 45 (20.9% of 215) outcomes that included a value modifier and declared a set measurement value to target. In some cases, the targeted value was common across multiple trials. For example, 6 of 7 HIV trials with HIV-1 RNA as a DE targeted 50 copies/mL for the outcome endpoint.

Time: 107 (49.8%) outcomes denoted a time frame for evaluation.

Discussion

Pivotal Trials

To our knowledge, our study is the first clinical research informatics effort to analyze data elements using a subset of pivotal trials (trials used in drug regulatory approval) from ClinicalTrials.gov. The key factor enabling the use of computerized methods to list all pivotal trials is the FDA initiative of publishing Drug Trial Snapshots. Outside the domain of trial data elements, Moore et al. analyzed the cost of pivotal trials for 52 drugs using annual summary reports from FDA Center for Drug Evaluation and Research.[12]

Our results show the importance of consistent use of NCT IDs (not just trial acronyms) to unambiguously identify each trial in FDA data. Another consideration is identification of pivotal trials for approvals prior to 2018. In addition to using Drug Trial Snapshots, we have experimented with FDA's Application Programming Interface (API) at https://open.fda.gov/apis/drug/label. We investigated the consistency of the presence of NCT IDs in section 14 (Clinical Studies) of drug labels. Even for drugs approved in the past, it is common to update drug labels. Such updates are initiated by a New Drug Application holder (sponsor) and can be used to add NCT IDs to section 14 of labels where previously studies are identified only by acronyms (and sometimes sponsor-specific study IDs). Either thanks to Drug Trial Snapshots or labels, the body of pivotal trials extracted using our methodology will likely only grow in the future. With a larger set of pivotal trials, we would be able to increase the number of conditions we can study using this approach (our study currently analyzed 90 conditions). Even though our primary motivation for this work was to arrive at a set of CDEs by disease (or medical condition), we consider the set of pivotal trials (and the methodology to generate it computationally) an interesting byproduct of this effort that can be used for other informatics research.

Data Elements

Common Data Element initiatives have defined thousands of standardized data elements (PhenX has 24 385 DEs, NINDS has 78 916 CDEs).[13,14] It can be overwhelming to identify CDEs, which have the most significance. Primary outcomes of pivotal trials provide a much more streamlined and targeted list of key CDEs. This smaller volume of DEs is also seen when comparing pivotal trials to all trials, as the average amount of primary outcomes for pivotal trials compared to all trials is much less. For pivotal trials, the average amount of primary outcomes per trial is 1.6, compared to 2.1 primary outcomes per trial for all trials. This method of efficiently identifying CDEs through pivotal trial primary outcomes is also not too restrictive, as the CDEs would be expected to be present in a vast amount of trials for a given condition. This is borne out as taking one of the identified CDEs, 'HIV-1 RNA', and doing a simple string search of all primary outcomes revealed 154 studies that included this CDE.

Limitations

Our work has several limitations. First, our analysis only analyzed the US drug approved list and the results may not apply to other countries. An alternative analysis could consider drug regulatory agencies from other countries (European Medicine Agency, Japan or China). Second, we only analyzed Drug Trial Snapshots and did not consider other documents, such as labels for approved drugs. Moreover, we only included trials that clearly listed an NCT ID which led to a smaller pivotal trial final sample. While we used a limited sources for generating the set of trials, our research does not aim to find a comprehensive list of CDEs used in clinical trials, but rather to identify a small subset of DEs representing important CDEs in clinical trials that can be used in trial development and CDE initiatives in an effort to improve their usability. Third, we used a single reviewer to transform trial primary outcomes into DEs. We see the main scientific contribution of our work in identification of a subset of pivotal trials. The choice of simple annotation methods was to demonstrate that a sensible set CDEs (organized by disease) can be generated. We acknowledge that much better annotation methods (and computerized pipelines) can be employed. Fourth, we did not roll up MeSH terms to combine similar conditions and did not assign one principal disease term to each trial. For a larger set of pivotal trials, this additional higher-level disease aggregation may produce superior disease groupings.

Relevant prior literature: Yuan et al. analyzed a variety of data-driven criteria for cohort identification through CTG and developed a tool called Criteria2Query.[15] Their methodology includes natural language processing of eligibility criteria. In contrast to their work, we focused on arriving at a significant subset of trials and for accuracy reasons, used human-mediated conversion. Before making this decision, we used their tool at http://www.ohdsi.org/web/criteria2query and provided feedback to their development team on changes that would be needed for us to be able to utilize their methodology on our input set of trials and data elements.

Conclusion

Pivotal clinical trials help identify the most significant CDEs, which can facilitate data harmonization across trials. Of the 172 DEs, 76 (44.2%) were CDEs within a disease. We observed conditions with multiple CDEs as well as identical CDEs significant for multiple conditions. More than half of the 172 identified DEs, 88 (51.2%) were lab tests, while the second highest number of DEs were non-lab measurements (32 DEs, 18.6%). This was true of CDEs as well, as there were 31 (40.8%) Lab test CDEs and 13 (17.1%) non-Lab measurement CDEs. We also identified several common patient-level modifiers (such as 'change from baseline', 'change', and 'absence') and two study-level modifiers (number and percentage of participants). Our approach can help CDE standardization initiatives identify significant CDEs in past completed trials.

Acknowledgement: This research was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and NIH Office of AIDS Research. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

References

1. NOT-OD-16-149: NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information [Internet]. [cited 2019 Mar 12]. Available from: https://grants.nih.gov/grants/guide/notice-files/not-od-16-149.html
2. Policy on clinical trials | Wellcome [Internet]. [cited 2019 Mar 12]. Available from: https://wellcome.ac.uk/node/1939
3. Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open. 2017 Dec 14;7(12):e018647.
4. Huser V, Amos L. Analyzing Real-World Use of Research Common Data Elements. AMIA Annu Symp Proc AMIA Symp. 2018;2018:602–8.
5. Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, et al. Improving the value of clinical research through the use of Common Data Elements. Clin Trials J Soc Clin Trials. 2016 Dec;13(6):671–6.
6. Doods J, Botteri F, Dugas M, Fritz F. A European inventory of common electronic health record data elements for clinical trial feasibility. Trials. 2014;15:18.
7. 2015 - Uncomplicated Gonorrhea Developing Drugs for Trea.pdf [Internet]. [cited 2020 Mar 2]. Available from: https://www.fda.gov/media/88904/download
8. Development & Approval Process | Drugs | FDA [Internet]. [cited 2020 Mar 5]. Available from: https://www.fda.gov/drugs/development-approval-process-drugs
9. Drug Trials Snapshots | FDA [Internet]. [cited 2020 Feb 28]. Available from: https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots
10. AACT database download [Internet]. Available from: https://aact.ctti-clinicaltrials.org/download
11. ClinicalTrials.gov Protocol Registration Data Element Definitions for Interventional and Observational Studies [Internet]. [cited 2020 Mar 9]. Available from: https://prsinfo.clinicaltrials.gov/definitions.html
12. Moore TJ, Zhang H, Anderson G, Alexander GC. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. JAMA Intern Med. 2018 Nov 1;178(11):1451.
13. PhenX Toolkit: Resources [Internet]. [cited 2020 Mar 2]. Available from: https://www.phenxtoolkit.org/resources/download
14. CDE Catalog | NINDS Common Data Elements [Internet]. [cited 2020 Mar 2]. Available from: https://www.commondataelements.ninds.nih.gov/cde-catalog
15. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc. 2019 Apr 1;26(4):294–305.