bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Obstacles to the Reuse of Study Metadata in ClinicalTrials.gov

Laura Miron, Rafael S. Gonçalves, Mark A. Musen
Stanford Center for Biomedical Informatics Research
Stanford University School of Medicine
1265 Welch Rd Stanford, CA 94305
lmiron@stanford.edu
650-725-3279

Keywords: clinical trial, metadata, eligibility criteria, ontology, data quality

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## ABSTRACT

**Objective:** ClinicalTrials.gov is a registry of clinical-trial metadata whose use is required by many funding agencies and scientific publishers. Metadata are essential to the reuse of data, but issues such as heterogenous metadata schemas, inconsistent values, and usage of free text instead of controlled terms pervade many metadata repositories. Our objective is to evaluate the quality of metadata about clinical studies in ClinicalTrials.gov and to document strategies to improve metadata accuracy.

**Methods:** Using 302,091 metadata records, we evaluated whether values adhere to type expectations for Boolean, integer, date, age, and value-set fields, and whether records contain fields required by the Food and Drug Administration. We tested whether values for *condition* and *intervention* use terms from biomedical ontologies, and whether values for *eligibility criteria* follow the recommended format.

**Results:** For simple fields, records contain correctly typed values, but there are anomalies in value-set fields. Contact information, outcome measures, and study design are frequently missing or underspecified. Important fields for search, such as c*ondition* and *intervention,* are not restricted to ontology terms, and almost half of the values for *condition* are not from MeSH, as recommended. Eligibility criteria are stored as unstructured free text.

**Conclusions:** *ClinicalTrials.gov*'s data-entry system enforces a schema with type restrictions, freeing records from common issues in metadata repositories. However, lack of ontology restrictions or structure for the condition, intervention, and eligibility criteria elements significantly impairs reusability. Searchability of the database depends on infrastructure that maps free-text values to terms from UMLS ontologies.

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## INTRODUCTION

Metadata are the lifeblood of biomedical data. At the simplest level, metadata are data that describe other data. In practice, we expect metadata to be structured and standardized, and to be useful in making the underlying data *findable* and *reusable*. High-quality metadata enhance scientific reproducibility and transparency, allow researchers to pool studies to increase the statistical power of inferences,[1] and enable the use of "big data" machine learning techniques. International metadata repositories such as the National Center for Biotechnology Information's (NCBI) BioSample and the European Bioinformatics Institute's (EBI) BioSamples repositories encourage data reuse through the availability of comprehensive metadata. They each gather metadata from several different repositories of biological data into a centralized, searchable database. Ideally, they also ensure that metadata follow unified standards and schema regardless of the author, source, and format of the original data.

Unfortunately, biomedical metadata are plagued by numerous quality issues. Hu et al. examined the quality of the metadata that accompany data records in the Gene Expression Omnibus (GEO) and found that they suffered from type inconsistency (e.g., numerical fields populated with non-numerical values), incompleteness (required fields not filled in), and the use of many syntactic variants for the same field (e.g., "age, Age, Age years, age year").[2] GEO predates the recent push for stricter metadata standards, and GEO metadata are created through an outdated spreadsheet system that allows users to submit unconstrained key–value pairs. GEO does not provide a structured vocabulary of field names to guide the author. GEO also does not provide controlled terminologies for field values, resulting in different versions of the same concept with no semantic link between them. There is no automated validation of submitted metadata, and manual curation is both time-consuming and error-prone.[2]

In past work, we documented similar problems in the NCBI BioSample and EBI BioSamples repositories. Unlike GEO, they were designed with the express purpose of standardizing and organizing the metadata from several different databases.[3] Both repositories provide data dictionaries of preferred field names and expected types, and recommend usage of specific ontologies for some fields, but they still allow arbitrary user-defined field names and provide very limited validation for values. Thus, a significant proportion of values do not adhere to their expected types, and the majority of values that ideally should use ontology terms do not.[4,5]

Use of terms from well-known domain-specific ontologies is one of the fundamental guidelines enumerated by the FAIR principles for making scientific data and metadata Findable, Accessible, Interoperable, and Reusable.[6] Where appropriate, values should be defined as globally unique and persistent identifiers, such as the URIs of terms in (a particular version of) an ontology[1]. A globally unique identifier denotes a term unambiguously, regardless of homonyms, and a persistent identifier gives researchers who consume metadata a reliable pointer to information about the term, such as labels, synonyms, and definitions.

*ClinicalTrials.gov* [7] is a Web resource created and maintained by the National Library of Medicine (NLM). It includes a repository of clinical-trial registrations, which are structured records of key–value pairs summarizing a trial's start and end dates, eligibility criteria, interventions prescribed, study design, names of sponsors and investigators, prespecified outcome measures, among other details. Clinical-trial registrations are data entities in their own right. The US Food and Drug Administration (FDA) requires registration for trials of certain

[1] e.g., http://purl.bioontology.org/ontology/MESH/D003920

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

drugs and devices as a means to protect human subjects, and the International Committee of Medical Journal Editors (ICMJE) publishes results only from trials registered before their start dates, to prevent selective reporting of positive results. Most published evaluations of *ClinicalTrials.gov* focus on missing, incomplete, or incorrect contact information,[8,9] lack of specificity in reported outcomes,[10,11] timeliness of registrations and results reporting,[12–14] and discrepancies between information in *ClinicalTrials.gov* and peer-reviewed publications.[14,15]

| Type | Explanation |
|---|---|
| ERROR | Problems that must be addressed (e.g., missing required* content, internal inconsistency) |
| WARNING | Items that are FDAAA required or (FDAAA) may be required (e.g., Study Start Date data element) |
| ALERT | Problems that need to be addressed |
| NOTE | Potential problems that should be reviewed and corrected, if possible (if not possible, then may ignore) |

***Figure 1- Warning levels in the PRS system***
*The PRS metadata-entry system displays four levels of error to the user: 'Error', 'Warning', 'Alert', and 'Note'. A record may not be submitted for review while any 'Error' messages remain, but may accepted with outstanding warnings of the other levels at the discretion of the ClinicalTrials.gov reviewer. Source: "PRS Overview and Resource Orientation"* [36]

Clinical-trial registrations are metadata records that summarize clinical studies around the world. In most cases, they are the only globally findable and accessible metadata that exist for a trial. Our objective is to evaluate whether clinical-trial registrations follow well-established standards for metadata quality—specifically whether they follow FAIR principles, and to document improvements to the metadata that could enhance their potential for reuse by programmatic systems. We have not seen other published research that describe analyses of *ClinicalTrials.gov* metadata according to these criteria, or that propose improvements to substantially enhance metadata quality.

## BACKGROUND

*ClinicalTrials.gov* is the largest of several international clinical trial registries. In 1997, 2005, and 2006, respectively, the FDA, the ICMJE, and the World Health Organization (WHO) released statements mandating the registration of clinical trials "study plans" or "protocols" in an approved registry.[16–18] The purpose of such registries is to make information available to patients about studies for which they are eligible or enrolled, and to prevent selective reporting bias, wherein researchers report positive results more than negative ones.[17]

These policies were intended only to govern true "interventional" clinical trials, in which subjects are prospectively assigned interventions. Today, however, *ClincalTrials.gov* contains three kinds of records: *interventional* trials, *observational* trials (which may additionally be designated *patient registries*), and *expanded access* records. Observational trials are those in which outcomes are retrospectively or prospectively observed, but interventions are not prescribed, including case–control studies, cohort studies, and cross-sectional studies. Expanded access records exist in conjunction with existing interventional records, in cases where study sponsors also administer the studied interventions to patients ineligible for the main cohort.

*ClinicalTrials.gov* was first released by the NLM in 2000, and it consists of a clinical-trial registration repository, a results database released in 2008, the protocol registration system (PRS) for submitting records, and a patient-facing website. The website search portal allows search based on conditions, interventions, eligibility requirements, sponsors, locations, and other fields within the metadata.

*ClinicalTrials.gov*'s data-entry system is the PRS, a Web tool that provides form-based entry pages and that includes a quality-review pipeline with both automated validation rules and

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

manual review by a member of the NLM staff. The PRS form-based entry system employs several methods to improve data quality. Markers by each field name indicate whether the element is required. Radio buttons are used for Boolean values and drop-down menus for value-set fields (**Figure *2***). Automated validation messages of four possible warning levels appear when errors or likely-wrong values are detected (**Figure 1**). Only when all errors are resolved may the author submit the record for manual review, where it will be either accepted or sent back with requested edits.

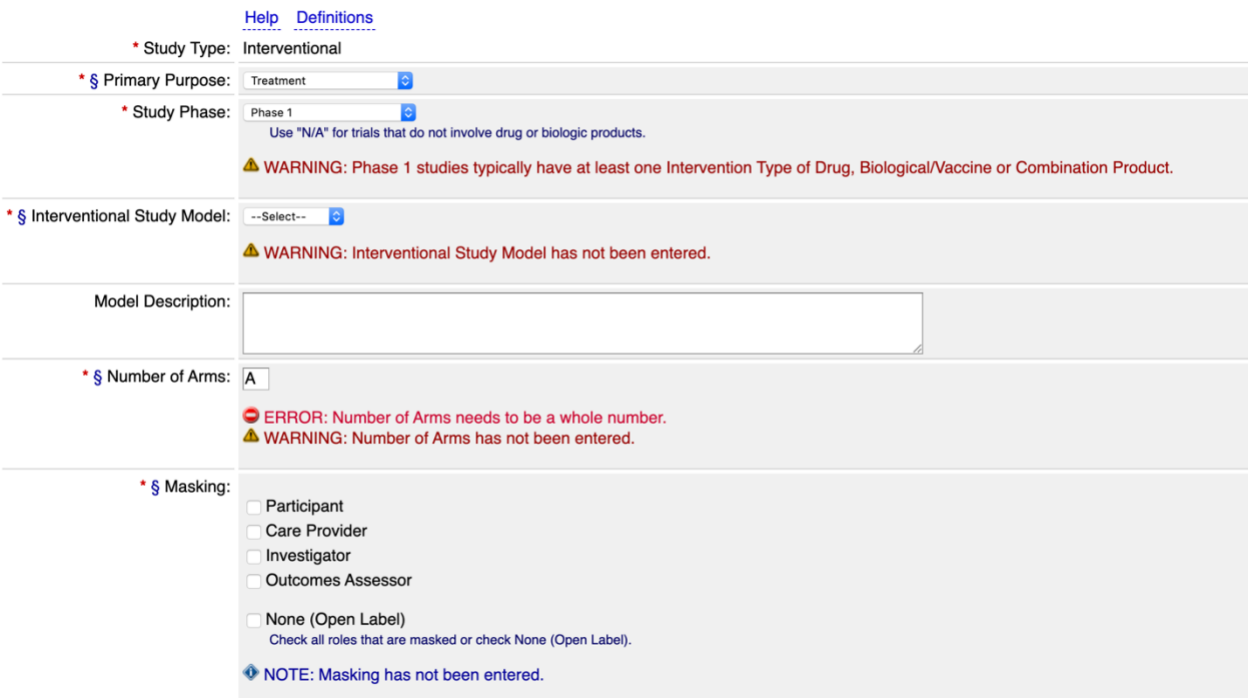


**Figure 2 - Data entry form in the PRS system**
One of several form pages for entering data in the PRS. Red asterisks (*) indicate required fields; red asterisks with a section sign (*§) indicate fields required since January 18, 2017. Additional instructions are provided for 'study phase' and 'masking' fields, and automated validation messages of levels 'Note', 'Warning' and 'Error' can be seen. A validation rule ensures that the value for 'number of arms' is an integer. Another rule checks both the chosen value for 'study phase' (Phase 1), and the (lack of) interventions that are enumerated on a separate page of the entry system. However, there is unexplained inconsistency in the warning levels for missing required elements (missing 'masking' generates a 'note', while missing 'interventional study model' and 'number of arms' both generate 'warnings').

## METHODS AND MATERIALS

*ClinicalTrials.gov* records are available as Web pages—accessible through the system's search portal—and as XML files.[2] We downloaded all public XML trial records (n=302,091) on April 3, 2019. We conducted our analysis of the PRS system using a test environment, which allows records to be created but never submitted, maintained by Stanford University.

### Data Element Definitions and Schema

[2] https://clinicaltrials.gov/AllPublicXML.zip

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ClinicalTrials.gov data elements are defined in a free-text "data dictionary"[3] and in an XML schema declaration (XSD). We examined the type definition of each field according to both specifications, and we documented fields where discrepancies exist.

[3] https://prsinfo.clinicaltrials.gov/definitions.html

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

***Table 1 – Important Field Definitions***

| | |
|---|---|
| *Condition* | The name(s) of the disease(s) or condition(s) studied in the clinical study |
| *Intervention* | The intervention(s) associated with each arm or group, most commonly a drug, device, or procedure |
| *Eligibility Criteria* | A limited list of criteria for selection of participants in the clinical study, provided in terms of inclusion and exclusion criteria |
| *Outcome Measure* | A prespecified measurement used to determine the effect of experimental variables on human subjects in a clinical study |

## Adherence to Type Expectations

We assigned each metadata field in the data dictionary a category based on the type of validation we would perform:

- Simple type (date, integer, age, Boolean) – Validate records against the XSD.
- Value-set field (data dictionary provides enumerated list of acceptable values) – Programmatically check values against expected values from data dictionary.
- Ontology-controlled field – Validate values against the expected ontologies.
- Free text – No validation.

## BioPortal as a Tool to Evaluate Use of Ontology Terms

We used the National Center for Biomedical Ontology (NCBO) BioPortal API to match values for the *condition* and *intervention* fields to ontology terms. Currently, only *condition* is ontology-restricted in ClinicalTrials.gov. The data dictionary says to "use, if available, appropriate descriptors from NLM's Medical Subject Headings (MeSH)-controlled vocabulary or terms from another vocabulary, such as the Systemized Nomenclature of Medicine–Clinical Terms (SNOMED-CT), that has been mapped to MeSH within the Unified Medical Language System (UMLS) Metathesaurus."[19] To test adherence to this restriction, we used BioPortal to search for exact matches for each term, restricted to the 72 ontologies in the 2019 version of UMLS. To evaluate the degree to which *intervention* field values were ontology-restricted, we queried all ontologies in BioPortal for exact matches for each intervention.

## Defining Completeness

Completeness is a fundamental quality characteristic of metadata,[20] and "minimum required information" standards exist for many biomedical subdomains, such as the "minimum information about a microarray experiment" (MIAME) standard.[21] FAIRsharing.org [22] provides a registry of such standards for metadata in various biomedical domains. For clinical-trial data, two main policies govern minimum information standards: the ICMJE/WHO trial registration dataset [23], and section 801 of the FDA Amendments Act of 2007 (FDAAA801) [24]. Although FDAAA801 only *legally* applies to trials of FDA-regulated drugs and devices within the US, *ClinicalTrials.gov* uses it to define the minimum required fields for all trials, and the WHO and ICMJE both recognize FDAAA801 as a standard equivalent to their own. We

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

therefore checked completeness of interventional trials against the data-element definitions in the FDAAA801 final rule.

The final rule lists 41 required fields. We ignored five fields that are conditionally required based on information unavailable to us (e.g., collaborating sponsors must be listed, only if they exist). We also ignored 3 fields stored internally by *ClinicalTrials.gov* but not made public, and 3 fields which were not added to ClinicalTrials.gov until November 2017 concerning FDA regulations. For all interventional trials, we determined the percentage of trials that are missing each required data element. We obtained the same statistics for the set of interventional trials with study start dates on or after January 18, 2017, the effective date of the FDAAA801 final rule, when data-element definitions were finalized.

# RESULTS

## Field Names

BioSample, BioSamples, GEO, and many other repositories allow metadata authors to submit field *names* as well as values, and as a result they often contain multiple syntactically different representations of the same attribute (e.g., "age, Age, Age years, age year").[5,25] Since the form-based PRS metadata-entry system does not allow user-defined fields, *ClinicalTrials.gov* does not suffer from this problem.

## Simple Type Expectations

The ClinicalTrials.gov XSD schema contained type definitions for all Boolean, integer, date, and age fields, and all records validated against this XSD (**Table 2**). Therefore, all records contain correctly typed values for all occurrences of these elements.

***Table 2 - Adherence to type expectations for Boolean, integer, date, and age fields.***
*All Boolean, integer, date, and age fields are typed in the XSD, and all values for these fields in all public records are correctly typed. 'Date' XML elements may optionally have an attibute designating them 'Actual', 'Anticipated', or 'Estimate'. For 'age' fields, records may represent equivalent ages with different units, e.g. '2 Years' and '24 Months'.*

| Type | Num. fields | Field Names | Format |
|---|---|---|---|
| Boolean | 11 | *has_expanded_access, has_dmc, is_fda_regulated_drug, is_fda_regulated_device, is_unapproved_device, is_ppsd, is_us_export, expanded_access_type_individual, expanded_access_type_intermediate, expanded_access_type_treatment, gender_based* | 'Yes\|No' |
| Integer | 3 | *number_of_arms, number_of_groups, enrollment* | xs:integer |
| Date | 4 | *start_date, completion_date, primary_completion_date, verification_date* | '(Unknown\|((January\|February\|March\|April\|May\|June\|July\|August\|September\|October\|November\|December)(([12]?[0-9]\|30\|31), )?[12][0-9]{3}))', *plus optionally*: 'Actual\|Anticipated\|Estimate' |
| Age | 3 | *minimum_age, maximum_age* | 'N/A\|([1-9][0-9]*(Year\|Years\|Month\|Months\|Week\|Weeks\|Day\|Days\|Hour\|Hours\|Minute\|Minutes))' |

## Value-Set Type Expectations

The trial metadata contained very few "rogue" values (not drawn from the data dictionary) for value-set fields (**Table 3**). Only nine of fifteen fields are typed within the XSD,

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

however, and the untyped fields appear as free text to programs ingesting the raw XML files. For two fields, the value-set in the data dictionary uses different syntax than the values that appear in the XML records (**Table 3**). The dictionary lists the acceptable values for *allocation* as “Single Group”, “Parallel”, “Crossover”, “Factorial”, and “Sequential”, but values appear in the records as “Single Group *Assignment*”, “Parallel Group *Assignment*”, etc. For the *masking* field, the data dictionary instructs the user to select from “Participant”, “Care Provider”, “Investigator”, “Outcomes Assessor”, or “No Masking”, but values appear in the actual metadata with the additional text “Single”, “Double”, “Triple”, or “Quadruple” to indicate the number of individuals providing masking.

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

| **Field** | **Valid Value Set** (data dictionary) | **Records With Rogue Values** | **Observed Rogue Values** | **Value Set Defined in XSD?** |
|---|---|---|---|---|
| Study Type | Interventional, Observational, Observational [Patient Registry], Expanded Access | 0% | -- | ✓ |
| Overall Recruitment Status | Not yet recruiting, Recruiting, Enrolling by invitation, "Active, not recruiting", Completed, Suspended, Terminated, Withdrawn | 0% | -- | ✓ |
| Responsible Party, by Official Title | Sponsor, Principal Investigator, Sponsor-Investigator | 0% | -- | ✓ |
| Study Phase | N/A, Early Phase 1, Phase 1, Phase 1/Phase 2, Phase 2/Phase 3, Phase 3, Phase 4 | 0% | -- | ✓ |
| Intervention Type | Drug, Device, Biologic/Vaccine, Procedure/Surgery, Radiation, Behavioral, Genetic, Dietary Supplement, Combination Product, Diagnostic Test, Other | 0% | -- | ✓ |
| Sex | All, Male, Female | 0% | -- | ✓ |
| Sampling Method | Probability Sample, Non-Probability Sample | 0% | -- | ✓ |
| Overall Study Official's Role | Study Chair, Study Director, Study Principal Investigator | 0% | -- | ✓ |
| Individual Site Status | Not yet recruiting, Recruiting, Enrolling by invitation, "Active, not recruiting", Completed, Suspended, Terminated, Withdrawn | 0% | -- | ✓ |
| Interventional Study Model | Single Group, Parallel, Crossover, Factorial, Sequential | 0%* | Notes: Values appear in XML as "Single Group Assignment", "Parallel Group Assignment", etc. | ✗ |
| Masking | *Select all that apply:* Participant, Care Provider, Investigator, Outcomes Assessor, No Masking | 0%* | Notes: Values appear in XML as "Double(Participant, Care Provider)", "Single(Investigator)", etc. | ✗ |
| Allocation | N/A, Randomized, Nonrandomized | .041% | Random Sample | ✗ |
| Arm Type | Experimental, Active Comparator, Placebo Comparator, Sham Comparator, No Intervention, Other | .0073% | Case, Control, Treatment Comparison | ✗ |
| Observational Study Model | Cohort, Case-Control, Case-Only, Case-Crossover, Ecologic or Community Studies, Family-Based, Other | 9.7% | Case Control, Defined Population, Natural History | ✗ |
| Time Perspective | Retrospective, Prospective, Cross-sectional, Other | 1.1% | Longitudinal, Retrospective/Prospective | ✗ |

***Table 3 – Value-Set Fields Validation Results***

*Fifteen fields are defined as value-sets (enumerated types) in the ClinicalTrials.gov data dictionary, but 6 are not typed within the XSD. Four of these 6 contain rogue values. For the "interventional study model" and "masking" fields, all values in public records are valid, but the data dictionary does not correctly describe the format of values. The actual format of "interventional study model" values include the word 'assignment' (e.g., 'Parallel Group Assignment' rather than 'Parallel Group'). Values for masking include the word 'single/double/triple/quadruple' in addition to the types of individuals providing masking.*

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## Completeness

We evaluated the completeness of 239,274 interventional records by measuring the percentage of trials missing values for each of the 29 fields required by FDAAA801 (**Table 4**). Sixteen fields are missing in a negligible number (<.05%) of records. FDAAA801 fields with no sub-elements are less likely to be missing than sub-elements of a top-level field, such as the 5 required sub-fields of *study design*. We refer to fields that are missing any required sub-fields as "underspecified".

Study design is underspecified in 74,175 records (31%), most commonly missing method of allocation to study arms or basic arm information (name, type, and interventions for each arm), although these sub-elements are required according to ICMJE/WHO specifications. At least one listed outcome measure lacks a description in 80,300 (34%) of trials, even though "name of the measure", "description or metric", and "time frame" for each outcome are also required ICMJE/WHO elements. The most prevalent type of missing information is contact information—163,113 records (68%) do not contain a full name, phone number, and email, either for the main contact person or for each listed location, as required by FDAAA801.

Trials with start dates within the scope of the FDAAA801 final rule have substantially more complete metadata. The number of trials that are missing several elements drops to zero, because newly added PRS validation rules will not allow a record to be submitted for review without them. The percentage of trials with underspecified outcome measures and facility information decreases, but outcome measures still lack detailed description in 7,800 records (17%), and 13,494 records (29%) lack fully-specified contact information.

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

***Table 4 - Missing required fields, before and after passage of FDA final rule***

*For fields required by the FDAAA801 final rule, table lists the percentage of all interventional records (n=239,274) missing the field, and the percentage of all interventional records with start dates after the effective date of the final rule (n=46,289) missing the field.*

***m** indicates multiple instances of field are permitted; a multiple field is considered 'missing' if there are no listed occurrences of field. **c** indicates a conditionally required element, such as 'Why Study Stopped', which is required only if the study terminated before its expected completion date. Conditionally required elements are considered missing if they are both missing and conditionally required for the given record. The 'Study Design', 'Primary Outcome Measure', and 'Facility Information' fields contain several required sub-fields. Table gives the percentage of records missing each sub-field, and the percentage of "underspecified" records where any required sub-field is missing.*

*Fields i-H (pediatric postmarket surveillance of a trial), i-K (other names for interventions), i-Q (post prior to U.S. FDA approval or clearance), i-R (product manufactured in or exported from the U.S.), i-X (secondary outcome measure information) have been omitted because they are only conditionally required, based on information unavailable to us. Fields i-N (studies a U.S. FDA-regulated device product), i-O (studies a U.S. FDA-regulated drug product), and i-P (device or product not approved or cleared by U.S. FDA) have been omitted because they were not added to ClinicalTrials.gov until 1/11/2017 and pertain only to whether the trial is subject to FDAAA801 regulations. Fields iv-C (U.S. Food and Drug Administration IND or IDE number), iv-D (human subject protection board review status), and iv-F (responsible party contact information) have been omitted because they are not made public.*

| **Required Field Name** | | **Percentage of Records Missing Field** | |
|---|---|---|---|
| | | All Interventional Records (n=239274) | Interventional Records starting on or after 01/18/17, effective date of Final Rule (n=46289) |
| **(i) Descriptive Information** | | | |
| (A) Brief Title | | 0% | 0% |
| (B) Official Title | | 2.9% | 0.022% |
| (C) Brief Summary | | 0.00042% | 0% |
| (D) Primary Purpose | | 3.5% | 0.013% |
| (E) Study Design | interventional study model | 2.9% | 0% |
| | number of arms | 10% | 0.66% |
| | allocation | 24% | 26% |
| | masking | 2.3% | 0.0022% |
| | arm information **m** | 10% | 0.64% |
| | *Total records with underspecified study design* | 31% | 26% |
| (F) Study Phase, for an applicable drug trial | | 0% | 0% |
| (G) Study Type | | 0% | 0% |
| (I) Primary Disease or Condition Being Studied in the Trial, or the Focus of the Study **c**,**m** | | 0.00042% | 0% |
| (J) Intervention Name(s), for each intervention studied **m** | | 0% | 0% |
| (L) Intervention Description, for each intervention studied **m** | | 19% | 0.063% |
| (M) Intervention Type, for each intervention studied **m** | | 0% | 0% |

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

| | | | |
|---|---|---|---|
| (S) Study Start Date | | 1.3% | 0.0065% |
| (T) Primary Completion Date | | 6.3% | 0% |
| (U) Study Completion Date | | 5.7% | 0.045% |
| (V) Enrollment | | 1.7% | 0% |
| (W) Primary Outcome Measure Information, for each primary outcome measure **m** | outcome measures *missing* | 3.3% | 0% |
| | time frame | 3.9% | 0% |
| | description | 34% | 17% |
| | *Total records with outcome measure missing or underspecified* | 48% | 17% |
| **(ii) Recruitment Information** | | | |
| (A) Eligibility Criteria | | 0.022% | 0.0043% |
| (B) Sex/Gender | | 0.0054% | 0% |
| (C) Age Limits | | 0.0054% | 0% |
| (D) Accepts Healthy Volunteers | | 55% | 0% |
| (E) Overall Recruitment Status | | 0% | 0% |
| (F) Why Study Stopped **c** | | 20% | 0.14% |
| (G) Individual Site Status **c** | | 35% | 27% |
| (H) Availability of Expanded Access | | 1.7% | 1.8% |
| **(iii) Location and contact information** | | | |
| (A) Name of the Sponsor **m** | | 0% | 0% |
| (B) Responsible Party, by Official Title | | 0.015% | 0.0048% |
| (C) Facility Information **m** | facility name | 11% | 0.01% |
| | facility city | 0.0057% | 0% |
| | facility country | 0.0057% | 0% |
| | *Missing location-specific contact info and overall contact info* | 68% | 29% |
| | *Total records with underspecified facility information* | 68% | 29% |
| **(iv) Administrative Data** | | | |
| (A) Unique Protocol Identification Number | | 0% | 0% |
| (B) Secondary ID | | 0.011% | 0.021% |
| (E) Record Verification Date | | 0% | 0% |

## Ontology-Restricted Fields

The only field currently restricted to terms from an ontology is the *condition* field. Rather than being restricted to a single ontology, authors are encouraged to use either MeSH terms, or terms than can be mapped to MeSH through the UMLS metathesaurus. Within the metadata records, values for *condition* appear as simple strings (e.g., “diabetes mellitus”) rather than as

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

globally unique, persistent identifiers. During metadata creation, PRS attempts to map user-submitted *condition* strings to UMLS concepts. If the mapping is successful, PRS accepts the user string as-is, without including the UMLS concept identifier in the metadata or replacing the user string with a standard syntactic representation of the concept. Alternative spellings ("tumor" vs "tumour") and synonyms ("breast cancer" vs "malignant neoplasm of the breast") are not harmonized. *ClinicalTrials.gov* addresses searchability issues that would normally arise in a database containing unharmonized synonyms by building a computation engine into its search portal that parses queries for UMLS concepts, and includes synonyms in the search (**Figure 3**). While this system appears to work well, it is available only through the *ClinicalTrials.gov* search portal, and unharmonized values persist in the raw metadata.

| **Terms**<br>Synonyms | **Search Results*** | **Entire Database**** |
|---|---|---|
| **Gastrointestinal Cancer** | 312 studies | 312 studies |
| gi cancer | 48 studies | 48 studies |
| digestive cancer | 28 studies | 28 studies |
| gastrointestinal tract cancer | 7 studies | 7 studies |
| Malignant Digestive System Neoplasm | 6 studies | 6 studies |
| Cancer of Gastrointestinal Tract | 3 studies | 3 studies |
| Cancer of the Gastrointestinal Tract | 3 studies | 3 studies |
| Malignant neoplasm of gastrointestinal tract | 3 studies | 3 studies |
| Malignant Gastrointestinal Neoplasm | 2 studies | 2 studies |
| **Cancer** | 6,467 studies | 69,347 studies |
| Neoplasm | 6,369 studies | 60,632 studies |
| Tumor | 1,112 studies | 15,457 studies |
| Malignancy | 142 studies | 2,961 studies |
| Oncology | 102 studies | 1,108 studies |
| neoplastic syndrome | 66 studies | 586 studies |
| Neoplasia | 65 studies | 589 studies |
| Neoplastic Disease | 1 studies | 19 studies |
| **Gastrointestinal** | 6,437 studies | 15,831 studies |

***Figure 3 – ClinicalTrials.gov search infrastructure***
*Synonyms automatically searched by ClinicalTrials.gov for query "gastrointestinal cancer"*

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

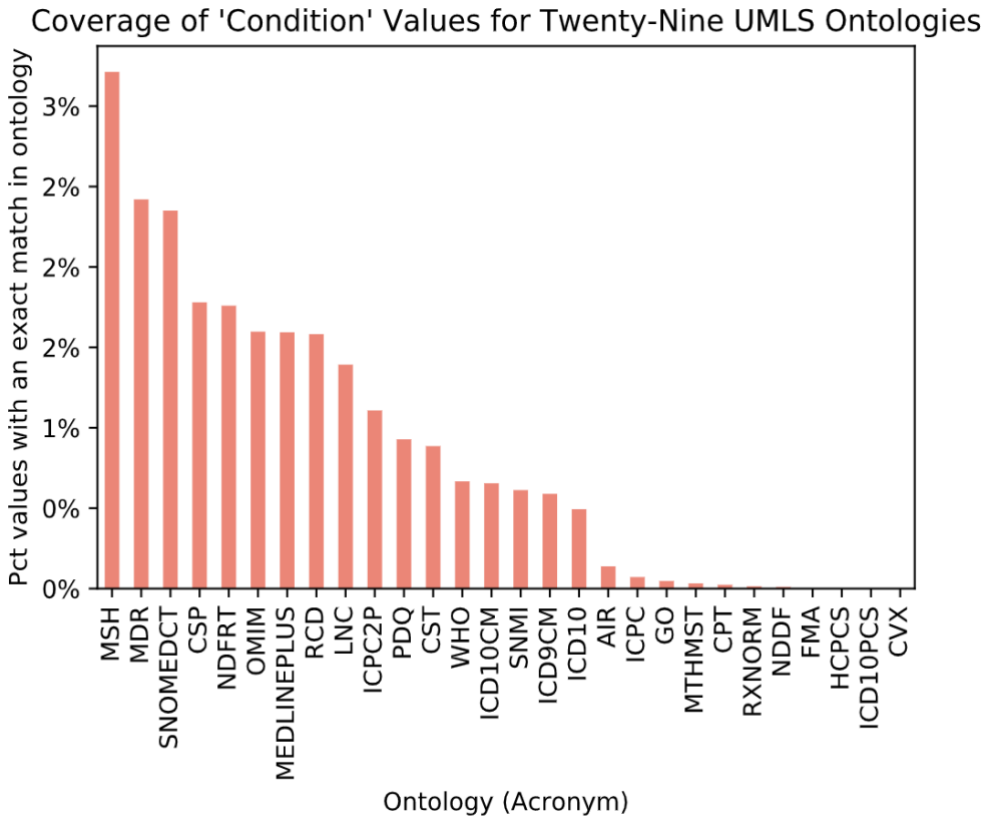


***Figure 4 – Coverage of 'condition' field by twenty-nine UMLS ontologies, measured as the percentage of 'condition' values for which ontology contains an exact match***
*The X-axis shows the 29 of 72 total UMLS ontologies which contained an exact match for at least 1 condition in ClinicalTrials.gov records. The Y-axis shows the fraction of all listed conditions that had an exact match in each ontology. MeSH, which contains matches for ~65% of listed conditions, has the best coverage, but does not significantly outperform MEDDRA or SNOMEDCT.*

We performed a comprehensive search for concepts in UMLS ontologies that matched the values given for the *condition* field (**Figure 4**). First, we checked adherence to the restrictions as they are defined. We found that only 306,197 (62%) of the 497,124 listed conditions return an exact match in MeSH, and only 402,875 (81%) have an exact match in any UMLS-mapped ontology. Second, we evaluated whether MeSH alone was sufficient to provide all terms used for the condition field. Of the 190,927 terms that have no match in MeSH, 96,678 conditions (51%) do have an exact match in another ontology. As shows, while MeSH provides the best coverage of any single ontology, it does not provide significantly more coverage than MEDDRA or SNOMED-CT.

We verified that the *intervention* field could be restricted to ontology terms without significant loss of specificity, by demonstrating that 362,546 out of 918,417 listed intervention values can be matched to terms from BioPortal ontologies, even without any pre-processing (**Figure 5**). All interventions have an associated *intervention type*, one of the eleven choices in **Figure 5**, and usage of ontology terms varies greatly between types.

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

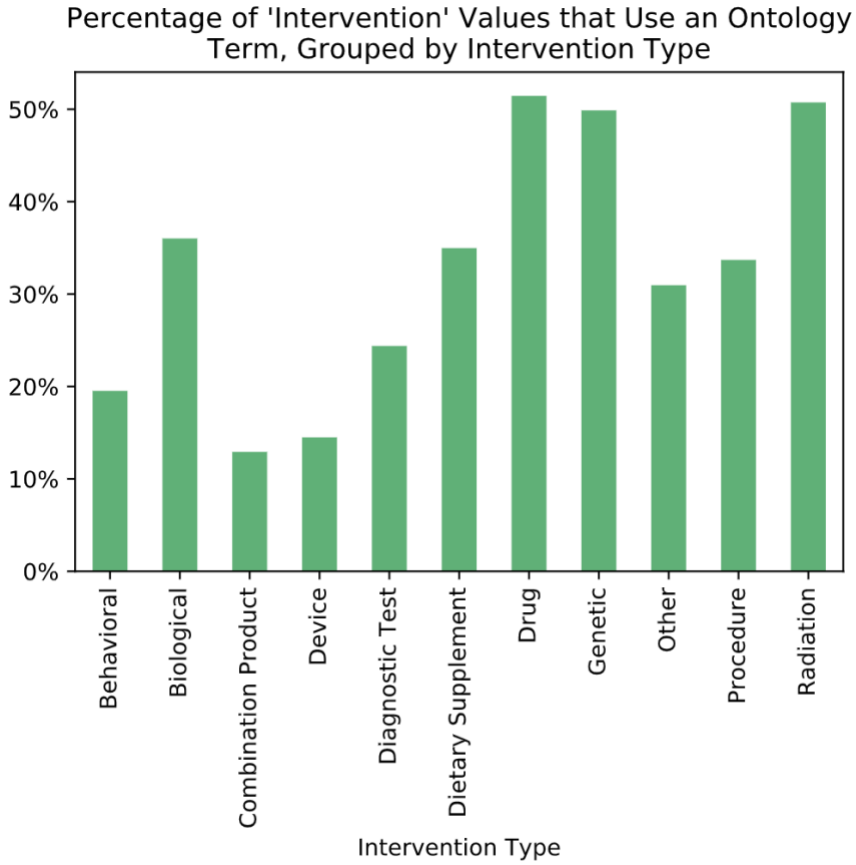


***Figure 5 - Percentage of values for the 'intervention' field for which we found an exact match in at least one ontology hosted in NCBO BioPortal, grouped by intervention type***

*Thirty-nine percent of listed values for 'intervention' are an exact match to a term from a BioPortal ontology, without any pre-parsing or normalization, indicating that this field could reasonably support ontology restrictions. Some intervention types are much better represented by ontology terms than others. More then half of all drugs and radiation therapies use ontology terms, but less than 15% of listed devices and combination products do.*

## Eligibility Criteria

The eligibility criteria element is a block of semi-formatted free text. The data dictionary says to "use a bulleted list for each criterion below the headers 'Inclusion Criteria' and 'Exclusion Criteria'", and the PRS prepopulates a textbox with the correct format (**Figure 6**). However, there is no enforced format restriction.

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

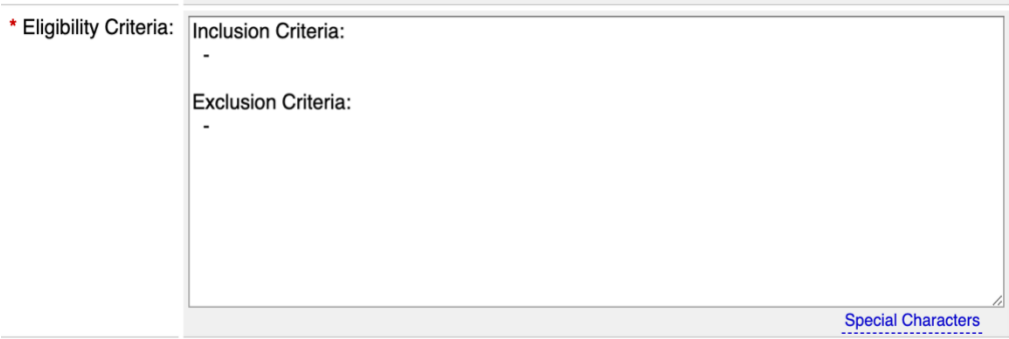


***Figure 6 – Prompt for entering eligibility criteria in the PRS***
*The PRS prepopulates the textbox for entering eligibility criteria with the correct headings and bullet style, but does not enforce the recommended format.*

We find that only 183,312 records (61%) follow the expected format for eligibility criteria (**Table 5**). Error types include:

- Missing one or both inclusion/exclusion headers
- Misspelled or alternative inclusion/exclusion headers
- Criteria not formatted, or only partially formatted as a bulleted list

***Table 5 - Percentage of records with incorrectly formatted eligibility criteria***
*Table shows the percentage of records with correctly formatted criteria, missing criteria, and the two most common incorrect formats: incorrect 'inclusion' and 'exclusion' headers, and non-bulleted criteria.*

| **Percentage of all Records (n=302,091)** | | | |
|---|---|---|---|
| Correctly Formatted Eligibility Criteria | Correct headers, but not formatted as a bulleted list | Missing or malformed headers | Missing Eligibility Criteria |
| 60.68% | 24.42% | 14.61% | .29% |

## DISCUSSION

Overall, we found that the metadata in *ClinicalTrials.gov* were of higher quality than the metadata in other biomedical repositories we have examined. For BioSample, BioSamples, and GEO, the two issues most impeding data reuse were non-standardized field names, and malformed values that failed to conform to the expected type for their field. Apart from minor irregularities in some value-set fields, *ClinicalTrials.gov* metadata were entirely free from these issues. In all cases, the design of the metadata authoring system played a key role in the quality of the metadata. BioSample, BioSamples and GEO all provide templates that suggest a particular format, but these repositories do not enforce restrictions with automated validation, placing that burden on metadata authors. In contrast, the PRS Web form system provides automated validation for most fields, and immediately displays error messages to the metadata author. A record cannot be submitted if any automatically-detected errors remain. Other metadata repositories would benefit from using similar techniques in their data-entry pipelines. The Center for Expanded Data Annotation and Retrieval (CEDAR) [26] has created such a

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

platform for metadata authoring—similar to PRS in that it enforces a schema and is based on forms—but CEDAR provides tight integration with ontologies to control field names and values.

While the trial metadata are well-formatted in some respects, several key obstacles to reuse remain. *ClinicalTrials.gov* was designed for users—who may be patients or members of the general public—to be informed of all possible treatment options.[27] It was not initially intended to be a computationally-usable source of metadata. Indeed, many design decisions prioritize usability for non-expert users of the website over ease of programmatic reuse of the raw metadata. For instance, the fact that the XML schema specification lacks type definitions for six value-set fields is irrelevant to a human browsing *ClinicalTrials.gov*, but the situation causes the fields to appear as free text to programs consuming the metadata, and therefore greatly limits the validation and downstream analyses that can be performed on the metadata.

Ontology restrictions are ill-defined and inconsistently-enforced in the case of the *condition* field, and non-existent for any other field; although we found that the *intervention* field could support ontology restrictions without significant loss of specificity. Terms in the UMLS Metathesaurus may be linked to terms in other ontologies through several kinds of relationships, not all of which indicate synonymy. Thus, the requirement that a condition term "can be mapped to MeSH" through UMLS is vague, and provides no expected range of values for a metadata-consuming program to compute over. Even when values are drawn from an ontology, they are not specified as globally unique and persistent identifiers. Many biomedical terms are present in multiple ontologies, so it is often impossible to determine the metadata author's intended source ontology. Since synonyms and syntactic variants are not harmonized within the raw metadata, searches within *ClinicalTrials.gov* depend on UMLS infrastructure that adds synonyms to each user query to retrieve all relevant records. Restricting the condition field to a single ontology would eliminate synonyms and provide a defined range of expected values, but our results in demonstrate that there is no single commonly used ontology that covers the majority of terms. Ideally, NLM should extend MeSH to include concepts from other ontologies that it is currently missing.

Missing fields in trial registrations both impair metadata quality and pose problems for the intended use of registrations as data entities in their own right. Researchers frequently use registration data to conduct meta-analyses and systematic reviews,[28] and incomplete metadata records can statistically affect such analyses. Covariances may exist between particular missing fields, although we have not analyzed whether such phenomena occur in *ClinicalTrials.gov*. Additionally, missing or underspecified contact information in almost seventy percent of all trials prevents patients from accessing elementary information about trials for which they may be eligible—the original stated purpose of *ClinicalTrials.gov*. Pre-specified outcome measures, which are critical to the ICMJE's use of registrations to prevent selective reporting of positive results, were missing or underspecified in almost half of all interventional trial records. Sub-elements of *study design info*, such as the method of patient allocation and the interventions administered to each arm, were also commonly missing, despite the fact that these fields are necessary to reproduce the trial, and that the validity of statistical inferences in clinical-trial results depends on such specifics.

Meaningful reuse of eligibility criteria is not possible given their current unstructured representation. Cohort definition and recruitment are among the most challenging aspects of conducting clinical trials,[29] and difficulties in recruitment cause delays for the majority of trials.[30,31] Structured eligibility criteria can be reused for related studies, and used to match eligible patients to clinical trials or to match patients to applicable clinical evidence.[32] Electronic health record (EHR) data from different care providers are becoming increasingly

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

standardized, and structured in ways that allow compatibly structured eligibility criteria to be used to directly query EHR databases.[33]

Several groups are developing structured representations and grammars for eligibility criteria, such as ERGO Annotation [34]—a format for eligibility criteria that includes free-text annotated with ontology terms, or systems for transforming existing free-text criteria into a structured representation, such as Criteria2Query—a system that converts eligibility criteria text into SQL queries.[33] Although no consensus on a computable representation has been reached, we propose simple improvements that move toward a more fully structured representation for clinical-trial eligibility criteria. As a first step, the element should be divided into separate fields for inclusion and exclusion criteria. This change would eliminate the need for user-defined headers, and fix 37% of existing errors, as **Table 5** shows. It is important that exclusion criteria are never mistakenly parsed as inclusion criteria due to a missing or malformed header, because they will be interpreted as expressing the logical opposite of their intended meaning.

For criteria to be reusable as queries of EHR databases, or as inputs to the systems mentioned above, they must be constrained to a collection of atomic Boolean statements. The bulleted lists in *ClinicalTrials.gov* must be parsed so that each criterion becomes a sub-field within either the 'inclusion criteria' field or the 'exclusion criteria' field. This is a challenging computational problem due to the inconsistencies we discovered in bullet type and spacing, as well as logical statements such as "patient must exhibit two or more of" followed by nested sub-criteria. In the future, we plan to investigate ways to parse and formulate these logical constructs, and to improve the data-entry system to create partially-structured criteria that can be indexed for searching the trial registration database, and that can eventually be more easily transformed into fully structured representations.

In a recent blog post, the NLM stated their intent to modernize *ClinicalTrials.gov* and to solicit feedback from users and stakeholders.[35] We hope that the findings of our analysis are useful for NLM. With this analysis of *ClinicalTrials.gov*, we intended to highlight the potential for health-care advancements from making metadata machine-readable, interoperable with other knowledge sources, and reusable by non-NLM systems. By improving the scientific rigor of metadata that describe scientific datasets, we can improve the discoverability and reusability of datasets, and thus accelerate the ability to make transformative data-driven clinical and biomedical discoveries.

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## REFERENCES


1 Federer LM, Lu Y-L, Joubert DJ, *et al.* Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. *PLoS One* 2015;**10**:e0129506. doi:10.1371/journal.pone.0129506

2 Hu W, Zaveri A, Qiu H, *et al.* Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata. *BMC Bioinformatics* 2017;**18**:415. doi:10.1186/s12859-017-1832-4

3 Barrett T, Clark K, Gevorgyan R, *et al.* BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res* 2012;**40**:D57–63. doi:10.1093/nar/gkr1163

4 Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. *Figshare* 2018;:1–26.

5 Gonçalves RS, O'Connor MJ, Martinez-Romero M, *et al.* Metadata in the BioSample Repository Are Impaired By Numerous Anomalies. 2017. doi:arXiv:1708.01286v1

6 Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018. doi:10.1038/sdata.2016.18

7 NIH U.S. National Library of Medicine. ClinicalTrials.gov. https://clinicaltrials.gov/

8 Viergever RF, Ghersi D. The Quality of Registration of Clinical Trials. *PLoS One* 2011;**6**:e14701. doi:10.1371/journal.pone.0014701

9 Chaturvedi N, Mehrotra B, Kumari S, *et al.* Some data quality issues at ClinicalTrials.gov. *Trials* 2019;**20**:378. doi:10.1186/s13063-019-3408-2

10 Bourgeois FT, Murthy S, Mandl KD. Outcome reporting among drug trials registered in ClinicalTrials.gov. *Ann Intern Med* 2010;**153**:158–66. doi:10.7326/0003-4819-153-3-201008030-00006

11 Viergever RF, Karam G, Reis A, *et al.* The Quality of Registration of Clinical Trials: Still a Problem. *PLoS One* 2014;**9**:e84727. doi:10.1371/journal.pone.0084727

12 Ross JS, Tse T, Zarin DA, *et al.* Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. *BMJ* 2012;**344**:d7292–d7292. doi:10.1136/bmj.d7292

13 Prayle AP, Hurley MN, Smyth AR. Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study. *BMJ* 2012;**344**:d7373–d7373. doi:10.1136/bmj.d7373

14 Riveros C, Dechartres A, Perrodeau E, *et al.* Timing and Completeness of Trial Results Posted at ClinicalTrials.gov and Published in Journals. *PLoS Med* 2013;**10**:e1001566. doi:10.1371/journal.pmed.1001566

15 Hartung DM, Zarin DA, Guise J-M, *et al.* Reporting Discrepancies Between the ClinicalTrials.gov Results Database and Peer-Reviewed Publications. *Ann Intern Med* 2014;**160**:477. doi:10.7326/M13-0480

16 Food and Drug Administration Modernization Act of 1997 (FDAMA). U.S.: 1997. https://www.govinfo.gov/content/pkg/PLAW-105publ115/pdf/PLAW-105publ115.pdf#page=16

17 ICMJE. Clinical Trials Registration: A Statement from the International Committee of Medical Journal Editors. http://www.icmje.org/news-and-editorials/clin_trial_sep2004.pdf

18 World Health Organization. WHO Statement on Public Disclosure of Clinical Trial Results. WHO Statement Public Discl. Clin. Trial Results. 2015.http://www.who.int/entity/ictrp/results/WHO_Statement_results_reporting_clinical_trials.pdf?ua=1

19 ClinicalTrials.gov Protocol Registration Data Element Definitions for Interventional and Observational Studies.

20 Margaritopoulos M, Margaritopoulos T, Mavridis I, *et al.* Quantifying and measuring metadata completeness. *J Am Soc Inf Sci Technol* 2012;**63**:724–37. doi:10.1002/asi.21706

21 Brazma A, Hingamp P, Quackenbush J, *et al.* Minimum information about a microarray

bioRxiv preprint doi: https://doi.org/10.1101/850578; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;**29**:365–71. doi:10.1038/ng1201-365

22 Sansone S-A, McQuilton P, Rocca-Serra P, *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nat Biotechnol* 2019;**37**:358–67. doi:10.1038/s41587-019-0080-8

23 World Health Organization. WHO Trial Registration Data Set (Version 1.3.1). https://www.who.int/ictrp/network/trds/en/

24 Food and Drug Administration Amendments Act of 2007. 2007.

25 Panahiazar M, Dumontier M, Gevaert O. Predicting biomedical metadata in CEDAR: A study of Gene Expression Omnibus (GEO). *J Biomed Inform* 2017;**72**:132–9. doi:10.1016/j.jbi.2017.06.017

26 Musen MA, Bean CA, Cheung K-H, *et al.* The center for expanded data annotation and retrieval. *J Am Med Informatics Assoc* 2015;:ocv048. doi:10.1093/jamia/ocv048

27 McCray AT, Ide NC. Design and Implementation of a National Clinical Trials Registry. *J Am Med Informatics Assoc* 2000;**7**:313–23. doi:10.1136/jamia.2000.0070313

28 Pradhan R (University of MMS. *Use of ClinicalTrials.gov Registry in Systematic Reviews and Meta-analyses: A Master's Thesis*. 2017. doi:https://doi.org/10.13028/M27H6R

29 Kadam RA, Borde SU, Madas SA, *et al.* Challenges in recruitment and retention of clinical trial subjects. *Perspect Clin Res*;**7**:137–43. doi:10.4103/2229-3485.184820

30 Thadani SR, Weng C, Bigger JT, *et al.* Electronic Screening Improves Efficiency in Clinical Trial Recruitment. *J Am Med Informatics Assoc* 2009;**16**:869–73. doi:10.1197/jamia.M3119

31 McDonald AM, Knight RC, Campbell MK, *et al.* What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. *Trials* 2006;**7**:9. doi:10.1186/1745-6215-7-9

32 Weng C, Tu SW, Sim I, *et al.* Formal representation of eligibility criteria: A literature review. *J Biomed Inform* 2010;**43**:451–67. doi:10.1016/j.jbi.2009.12.004

33 Yuan C, Ryan PB, Ta C, *et al.* Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Informatics Assoc* 2019;**26**:294–305. doi:10.1093/jamia/ocy178

34 Sim I, Tu SW, Carini S, *et al.* The Ontology of Clinical Research (OCRe): An informatics foundation for the science of clinical research. *J Biomed Inform* 2014;**52**:78–91. doi:10.1016/j.jbi.2013.11.002

35 Williams R. Engaging Users to Support the Modernization of ClinicalTrials.gov. NLM Musings from Mezzanine. 2019.https://nlmdirector.nlm.nih.gov/2019/08/13/engaging-users-to-support-the-modernization-of-clinicaltrials-gov/

36 Protocol Registration and Results System (PRS) Overview. 2015;:17.https://prsinfo.clinicaltrials.gov/trainTrainer/PRS-Overview-and-Resource-Orientation.pdf