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

Word Count: 3860

ABSTRACT

Metadata which use consistent schemas, enforce type restrictions on values, and require, where possible, that values be drawn from a well-known ontology are essential to making biomedical data findable and reusable in further analyses. One class of metadata describes the experimental protocol, funding, and scientific leadership of clinical studies. The largest publicly available source of such metadata are the registrations in *ClinicalTrials.gov*, a registry of clinical-trial study protocols whose use is required by many funding agencies and scientific publishers. Using 302,091 metadata records, we evaluated whether values adhere to type expectations for Boolean, integer, date, age, and enumerated 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. For simple fields, records contain correctly typed values, but there are anomalies in fields with enumerated values. 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.

INTRODUCTION

Over the past two decades, the scientific community has increasingly recognized the need to make protocols and results of biomedical experiments publicly accessible, so that data can be reused and analyzed. However, the findability and meaningful reuse of data are often hampered by the lack of standardized metadata that describe the data. Biomedical metadata are typically records of key–value pairs that are created when an investigator submits data to a repository such as the Gene Expression Omnibus (GEO), Protein Data Bank (PDB), or NCBI BioSamples repository. They describe the source of the data (e.g., investigators, sponsoring organizations, dates records were submitted and updated), the structure of datasets, experimental protocols, identifying and summarizing information, and other domain-specific fields.

For clinical trials, research conducted on human subjects to evaluate a new medical treatment, drug, or device, the most complete source of machine-readable metadata is *ClinicalTrials.gov.*[1] *ClinicalTrials.gov* [2] is a Web resource created and maintained by the U.S. 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, and prespecified outcome measures, among other details. *ClinicalTrials.gov* is an interesting source of metadata in that its official purpose is not to be a repository of scientific metadata, but rather to implement the requirements from a series of acts by the US Food and Drug Administration (FDA)[1] which mandated the registration of trials of controlled drugs and devices as a means to safeguard human subjects.

Nevertheless, clinical trial registries have become a crucial source of data for systematic reviews and other metanalyses. Recent studies recommend that systematic reviews include a search of clinical trial registries to identify relevant trials that are ongoing or unpublished.[3–6] Baudard et al.[7] and Hart et al.[8] found that including unpublished results from registries has a statistically significant impact on summary statistics (e.g., efficacy). Including registry metadata in systematic reviews can help identify selective reporting bias by comparing published outcomes to prespecified outcomes[9,10], and adverse events are more likely to be reported in clinical trial registries than in published literature.[11,12] Records from *ClinicalTrials.gov* are also commonly used in metanalyses about trends in sources of funding for trials[13–17], diseases and interventions studied,[14,18–22] study design,[18,20,23] time to publication following study completion[1,12,24–26], geographical availability of trials sites [27–32], and the causes of delays and early terminations in studies.[33–40]

Unfortunately, *ClinicalTrials.gov* records and metadata records from several other widely used biomedical data repositories are plagued by documented quality issues. Several studies have analyzed *ClinicalTrials.gov* records missing fields required by the Food and Drug Administration Amendments Act of 2007, which governs US trial registries, or the WHO minimum data set, which provides guidelines for registries internationally.[41–44] Chaturvedi et al.[45] found that information about the principal investigators of trials in *ClinicalTrials.gov* is inconsistent both within multiple occurrences in the same record and across records. Tse et al.[46] identify additional obstacles to data reuse: follow-up studies are not always linked to the original study, records can be modified by the responsible party at any time, standards include

[1] Food and Drug Administration Modernization Act of 1997, Food and Drug Administration Amendments Act of 2007, FDAAA801 Final Rule, released in 2016

both mandatory and optional data elements, and the presence of records in the database is biased by reporting incentives.

While these quality issues with *ClinicalTrials.gov* records are known, we have not seen other evaluations of whether study records conform to several characteristics of high-quality metadata known to be lacking in other biomedical data repositories such as GEO,[47] NCBI BioSample and EBI BioSamples,[48] and identified in the FAIR principles for making scientific data and metadata Findable, Accessible, Interoperable, and Reusable.[49] Specifically, we look for use of non-standardized field names, values that do not conform to the expected type, usage of ontology terms, missing required values, and the usage of unstructured free-text fields. We also examine the data-entry pipeline for *ClinicalTrials.gov*, the Protocol Registration System, since for both GEO and BioSamples metadata, the data-entry pipeline was shown to have a significant impact on metadata quality. Our goals are to make recommendations to NLM for the improvement of *ClinicalTrials.*gov, and alert researchers of the potential pitfalls of using *ClinicalTrials.gov* metadata.

We find that automated validation rules within the PRS have been successful at enforcing type restrictions on numeric, date, and Boolean fields, and fields with enumerated values, but fields such as contact information, principal investigators, study design, and outcome measures are frequently missing or underspecified. Values for fields commonly used in search queries such as *condition*(s) and *intervention*(s) are not restricted to ontology terms, impeding search. Eligibility criteria, which could be used to automate matching patients to applicable clinical trials if stored in a structured format, are currently stored as semi-structured free-text and cannot be used to query the repository.

BACKGROUND

History of Clinical Trials Registries

ClinicalTrials.gov is the largest of several international clinical trial registries. In 1997, 2005, and 2006, respectively, the Food and Drug Administration (FDA), the International Committee of Medical Journal Editors (ICMJE), and the World Health Organization (WHO) released statements mandating the registration of clinical trials “study plans” or “protocols” in an approved registry.[50–52] Registries would make information available to patients about studies for which they are eligible or enrolled, and help prevent selective reporting of positive results.[51] *ClinicalTrials.gov* was first released by the NLM in 2000, and it consists of a clinical-trial registration repository, the Protocol Registration System (PRS) for submitting records, and a patient-facing website. The search portal of the patient-facing website allows queries based on conditions, interventions, eligibility criteria, sponsors, locations, and other fields within the metadata.

ClinicalTrials.gov contains three kinds of records: *interventional* trials (subjects are prospectively assigned interventions), *observational* trials (outcomes are retrospectively or prospectively observed, but interventions are not prescribed; may additionally be designated *patient registries*), and *expanded access* records. Expanded access records exist in conjunction with existing interventional records, in cases where study sponsors also administer the studied interventions to patients who are ineligible for the main cohort.

Data Entry: Protocol Registration System

The data-entry system for *ClinicalTrials.gov* is the Protocol Registration System (PRS), a Web tool that provides form-based entry pages and that includes a quality-review pipeline with

both automated validation rules and 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 fields with enumerated values (**Figure 1**). Automated validation messages of four possible warning levels (**Table 1**) appear when errors or potentially wrong values are detected. The author may only submit the record for manual review when all errors are resolved.

Table 1- Warning Levels in the Protocol Registration System

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)

Characteristics of High-Quality Metadata

Several biomedical domains define a “minimum information standard”, or list of required fields, for metadata about a particular type of experiment (e.g., the Minimum Information About a Microarray Experiment (MIAME)[53]). FAIRsharing.org [54] provides a registry of such standards for metadata in various biomedical domains. High quality, "complete" metadata contain values for all fields required by the relevant minimum information standard. For clinical-trial data, two main policies govern minimum information standards: the ICMJE/WHO trial registration dataset [55], and section 801 of the FDA Amendments Act of 2007 (FDAAA801) [56]. In theory, the WHO trial registration dataset [55] applies to all interventional trials in the world, while FDAAA801 applies only to interventional trials of controlled drugs and devices within the US. In practice, however, *ClinicalTrials.gov* is the largest international repository of both types of trials, and its required element definitions correspond to those defined in FDAAA801. The ICMJE and WHO therefore accept trials which are fully registered in *ClinicalTrials.gov* as meeting their standard, and release an official mapping of WHO data elements to *ClinicalTrials.gov* data elements.[2]

Metadata schemas often require that values for certain fields be drawn from a particular biomedical ontology in order to prevent the usage of synonyms, and provide a defined range of values that can be used to query the metadata. 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.[49] 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[3]. A globally unique identifier denotes a

[2] https://prsinfo.clinicaltrials.gov/trainTrainer/WHO-ICMJE-ClinTrialsgov-Cross-Ref.pdf
[3] e.g., http://purl.bioontology.org/ontology/MESH/D003920

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. The *condition* and *intervention* fields within *ClinicalTrials.gov* records share characteristics of fields that could support and be improved by ontology restrictions on the allowed values: expected values for these fields are already likely to be found in well-known ontologies such as MeSH or RXNORM, unrestricted values for these fields are likely to introduce synonyms (e.g., the proprietary name and generic name for a drug), and they are important fields for querying the repository.

METHODS AND MATERIALS

ClinicalTrials.gov records are available as Web pages—accessible through the system's search portal—and as XML files.[4] 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. We conducted several analyses: We enumerated all fields which expect values to conform to a simple type (integer, Boolean, date, or enumerated values) or to be drawn from an ontology, and tested whether values adhere to type expectations. To evaluate completeness, we counted the numbers of records missing all fields required by FDAAA801. We conducted specific analyses on the *eligibility criteria* element using regular expressions.

Data Element Definitions and Schema

ClinicalTrials.gov data elements are defined in a free-text "data dictionary"[5] 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.

[4] https://clinicaltrials.gov/AllPublicXML.zip
[5] https://prsinfo.clinicaltrials.gov/definitions.html

Significant Fields in ClinicalTrials.gov	
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
Responsible Party	Either a "Sponsor", "Principal Investigator", or "Sponsor-Investigator"; the party responsible for submitting information about a trial
Overall Official	Person responsible for the overall scientific leadership of the protocol, including study principal investigator

Adherence to Simple 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.
- Enumerated values 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.

Usage of Ontology Terms

We used the National Center for Biomedical Ontology (NCBO) BioPortal API to find ontology terms whose preferred names are exact matches for values for the *condition* and *intervention* fields. 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."[57] 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.

Missing Fields

We counted the number of records with missing fields for 28 of the 41 fields required by FDAAA801, the statute that defines the minimum required elements for clinical trial registrations in the US. We ignored five fields that are conditionally required based on information

unavailable to us[6] (e.g., secondary outcome measures must be listed, only if they exist), three fields stored internally by *ClinicalTrials.gov* but not made public[7], three fields which were not added to ClinicalTrials.gov until November 2017 concerning FDA regulations[8], and two fields that represent administrative data present in all records[9].

Some required data element definitions were updated by the Final Rule, an amendment to FDAAA801 released on September 09, 2016. The new element definitions are legally required for all trials with start dates on or after January 18, 2017, and *ClinicalTrials.gov* released updated element definitions on January 11, 2017 to support the final rule regulations. We therefore count the number of records missing each of the 28 fields for the set of interventional records whose listed start date is before January 18, 2017, and for the set of interventional records whose start date is on or after January 18, 2017.

We also counted the number of records with no listed "Principal Investigator", required by the WHO dataset (called "Contact for Scientific Inquiries") but not required by FDAAA801. A principal investigator may be listed within a *ClinicalTrials.gov* record either as the "responsible party" when it is of type "Principal Investigator" or "Sponsor-Investigator", or in the "overall official" element.

Eligibility Criteria

The expected format for eligibility criteria is a bulleted list for each criterion below the headers 'Inclusion Criteria' and 'Exclusion Criteria'. We used regular expressions to categorize the eligibility criteria from every trial record (n=302,091) as: 1) 'correctly formatted'; 2) 'correct headers but not a bulleted list of criteria'; or 3) 'incorrectly formatted'. We manually reviewed a random sample of 300 records from group 2 and group 3 to count the number of eligibility definitions that failed to parse because they listed criteria for more than one sub-group of participants.

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").[58,59] 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.

---

[6] *I-H: Pediatric Postmarket Surveillance, I-K: Other Names for Interventions, I-Q: Post Prior to FDA Approval/Clearance, I-R: Product Manufactured in or Exported from the U.S., I-X Secondary Outcome Measure Information*

[7] *IV-C: FDA IND or IDE, IV-D: Human Subject Protection Board Review Status, IV-F: Responsible Party Contact Information*

[8] *I-N: Studies an FDA-regulated Device Product, I-O: Studies an FDA-regulated drug product, I-P: Device or Product Not Approved/Cleared by FDA*

[9] IV-A: Unique Protocol Identification Number, IV-E: Record Verification Date

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))'

Fields with Enumerated Values

The trial metadata contained very few "rogue" values (not drawn from the data dictionary) for fields with enumerated values (**Table 3**). Only nine of fifteen fields are typed within the XSD, however, and the untyped fields appear as free text to programs ingesting the raw XML files. For two fields, the enumerated permissible values in the data dictionary use 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 (or "blinding").

Table 3 – Enumerated Value Fields Validation Results

Fifteen fields have an enumerated set of permissible values 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.

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

Completeness

Table 4 shows the number of interventional trial records that are missing for each of the elements required by the FDAAA801 Final Rule. Fourteen of the 41 required fields are omitted because they are present in all records or are missing in a negligible number of records (<.03%).[10]

Several elements were added or updated in the Final Rule. "Official title", "why study stopped" (for a study that is "suspended", "terminated" or "withdrawn" prior to its planned completion), "study start date", and "study completion date" were recommended but not required for studies with start dates before the effective date of the final rule. FDAAA801 defines a single study design element whose textual definition states that study design information should include "Interventional study model, number of arms, arm information, allocation, masking, and whether a single-arm trial is controlled." This information was stored in *ClinicalTrials.gov* records as a comma-separated list. The final rule eliminates the single-arm control element, makes "interventional model", "allocation", "enrollment", and "masking" required sub-elements of "study design", and makes "number of arms" and "arm information" separate required elements.

The maintainers of *ClinicalTrials.gov* have parsed the study design information in historical records and inserted it into the new structured element. However, there is no way to confirm that this conversion was done with 100% accuracy, and any unstructured data in the original study design elements has been lost, even within the historical versions of records that *ClincalTrials.gov* provides. "Arm information" consists of a "label" (e.g., the name of the experimental intervention used or placebo), "type" ("Experimental", "active comparator", "placebo comparator", or "other"), and description. "Label" and "type" were present in all listed arms, but 10% of the records failed to list any study arms.

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.*

Required Field Name		**Number of Interventional Records Missing Field**		**Percentage of Records Missing Field**
		Trials starting before 01/18/17, effective date of Final Rule (n=192985)	Trials starting on or after 01/18/17 (n=46289)	All Interventional Trials (n=239274)
(i) Descriptive Information				
(B) Official Title		7038	10	2.9%
(D) Primary Purpose		8253	6	3.5%
(E) Study Design	interventional study model	6834	0	2.9%

[10] i-A Brief Title, i-C Brief Summary, i-F Study Phase, i-G Study Type, i-I Primary Disease or Condition Being Studied, i-J Intervention Name(s), i-M Intervention Type, ii-A Eligibility Criteria, ii-B Sex/Gender, ii-C Age Limits, ii-E Overall Recruitment Status, iii-A Name of the Sponsor, iii-B Responsible Party by Official Title, and iv-B Secondary ID

	number of arms	23,832	296	10%
	allocation	44,552	11938	24%
	masking	5423	1	2.3%
	arm information **m**	23,832	296	10%
(L) Intervention Description, for each intervention studied **m**		45,433	29	19%
(S) Study Start Date		2950	3	1.3%
(T) Primary Completion Date		15,044	0	6.3%
(U) Study Completion Date		13,516	21	5.7%
(V) Enrollment		3966	0	1.7%
(W) Primary Outcome Measure Information **m**	outcome measures *missing*	7896	0	3.3%
	time frame	9331	0	3.9%
	description	73,484	7869	34%
(ii) Recruitment Information				
(D) Accepts Healthy Volunteers		131,601	0	55%
(F) Why Study Stopped **c**		47,790	65	20%
(G) Individual Site Status **c m**		71,248	12498	35%
(H) Availability of Expanded Access		39,844	833	1.7%
(iii) Location and contact information				
(C) Facility Information	facility name **m**	26,315	5	11%
	Overall contact info OR contact info for each site	149,282	13424	68%

Most required fields are present in nearly all records submitted after the January 2017 update to *ClinicalTrials.gov*, because automated PRS validation rules prevent the submission of records with missing required fields. However, contact information was frequently missing or underspecified both before and after the final rule. Contact information consisting of a name, phone number, and email address is required either for each individual facility, or a single *overall contact* is required for the trial. No overall contact is given in 223,468 trials (74% of all trials), and of those trials only 5,853 list a primary contact for each trial location. Of all 385,279 contact details that are provided, either as the overall contact or a location-specific contact, 16,137 (4.2%) lack an email address and 20,280 (5.3%) lack a phone number.

The eighth element in the WHO dataset is a *Contact for Scientific Queries*, and its description states that there must be “clearly assigned responsibility for scientific leadership to a named Principal Investigator” and that this data element must include “Name and title, email address, telephone number, postal address and affiliation” of the PI, even if additional contact details are provided for a second contact. The FDA requires only a responsible party, which is permitted to be a sponsor. If the responsible party type is “Principal Investigator” or “Sponsor-Investigator”, then only *name*, *investigator affiliation*, and *investigator title* are required. *ClinicalTrials.gov* provides an additional element, *overall official*, which corresponds to the WHO PI element, but it is not required, and the element definition does not include contact information.

Out of 302,091 trials, 35,226 (12%) have no listed principal investigators, 22,557 (7.5%) have a responsible party of type “Principal Investigator” or “Sponsor-Investigator” but do not separately designate an overall official, 162,985 (54%) have an “overall official” and a non-

scientific responsible party (e.g., a sponsor), and 81,323 (27%) list both an overall official and investigator information for the responsible party.

We noticed irregularities in the structure of both investigator and contact-related elements. Within the XSD, an investigator has sub-fields *first_name*, *middle_name*, *last_name*, *degrees*, *role*, *affiliation*, and a contact has sub-fields *first_name*, *middle_name*, *last_name*, *degrees*, *phone*, *phone_ext*, and *email*. However, *first_name*, *middle_name*, and *degrees* are missing in all investigators and contacts in all records, and instead the individual's full name and degrees all appear within the value of the *last_name* field (e.g., "Sarah Smith, M.D."). The *responsible party* element contains sub-fields *investigator_affiliation*, *investigator_full_name*, and *investigator_title* with no field for *degrees*.

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 *ClinicalTrials.gov* records, values for *condition* appear as simple strings (e.g., "diabetes mellitus") rather than as 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 (**Table 5**). While this system mitigates some of the issues arising from synonyms and alternative spellings, it is available only through the *ClinicalTrials.gov* search portal, and unharmonized values persist in the raw metadata. Only synonyms for the query term are provided, and users cannot browse from their original query to more or less general concepts (e.g., in the MeSH hierarchy) in order to refine their search. These detected synonyms are always included and the user cannot choose to search for an exact phrase.

Table 5* - *Synonyms Added by ClinicalTrials.gov for Search Term "Cancer"
Behind the scenes, the ClinicalTrials.gov search portal adds 7 synonyms to a user query for cancer". Note: Inconsistent capitalization accurately reflects how terms are displayed in ClinicalTrials.gov.

Query Term	**Num. Search Results**
cancer	74,385 studies
Synonyms	
Neoplasm	66,572 studies
Tumor	16,473 studies
Malignancy	3,128 studies
Oncology	1,249 studies
Neoplasia	622 studies
neoplastic syndrome	592 studies
Neoplastic Disease	22 studies

We performed a comprehensive search for concepts in UMLS ontologies that matched the values given for the *condition* field. 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. MeSH provides the best coverage of any single ontology, but does not significantly outperform MEDDRA, which contains matches for 230,639 conditions (46%), or SNOMED-CT, which contains matches for 224,008 conditions (45%).

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 2**). All interventions have an associated *intervention type*, one of the eleven choices in **Figure 2**, and usage of ontology terms varies greatly between types.

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. However, there is no enforced format restriction. We find that only 183,312 records (61%) follow the expected format for eligibility criteria (**Table 6**). 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
- Criteria defined for sub-groups of participants, and/or defined for non-subjects

Table 6 - 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%

Out of a random sample of 400 records which had incorrectly formatted eligibility criteria, 55 records defined criteria for multiple sub-groups of participants (e.g., diseased patients and healthy participants, participants in part 1 and part 2 of the study), or for non-participants (e.g., parents, care providers, locations). The most common cause of criteria failing to parse according to the expected format was paragraph-style sentences interspersed with bulleted criteria.

DISCUSSION

Overall, we found that the metadata in *ClinicalTrials.gov* were of higher quality than the metadata in other biomedical repositories we have examined. All field names are standardized, and values for numeric, date, and Boolean fields, and fields with enumerated permissible values conform to type expectations. The presence of some required fields is strictly enforced by the Protocol Registration System, but contact information, principal investigators, study design information, and outcome measures are frequently missing. Two key fields for searching records, *condition* and *intervention*, do not restrict values to terms from ontologies, so users cannot easily refine or broaden search queries. Eligibility criteria are stored as semi-structured text; they are recommended to be formatted as a bulleted list of individual criteria, but nearly 49% of values fail to parse according to the expected format.

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 fields with enumerated values, *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 suggesting a particular format, but 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) [60] has created such a 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.

Restricting the allowed values for certain fields in biomedical metadata to terms drawn from a commonly used ontology such as MeSH can improve metadata computability and reusability by eliminating the usage of synonyms and preventing typographical errors, providing a defined range of values over which analyses can be performed, and enabling users to develop and refine a search query by browsing up or down in the term hierarchy. No field within *ClincialTrials.gov* records is required to use values from an ontology, and the data dictionary recommendation that the *condition* field use values that "can be mapped to MeSH" through the UMLS Metathesaurus is too vague to provide a defined set of expected values. Our results demonstrate that there is no single commonly used ontology that covers the majority of terms. Ideally, NLM should extend MeSH, which is intended to be browsable and comprehensive across all medical domains, to include concepts from other ontologies that it is currently missing.

For both the *condition* and *intervention* fields, *ClinicalTrials.gov* addresses searchability issues caused by the existing of synonyms and lack of a defined range of search terms by automatically including synonyms of the user-provided query (**Table 5**). This functionality is lost, however, if the raw XML records are downloaded from *ClinicalTrials.gov* and reused in a different system. Moreover, the inclusion of synonyms cannot be toggled off, and the user may disagree with *ClinicalTrials.gov*'s definitions of synonymy. It is, for instance, debatable whether a "tumor", which may be benign, is a synonym for "cancer", of which not all types result in tumors, yet "tumor" is always included in user queries for "cancer" (**Table 5**).

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. Persistent identifiers for each term in (a particular version of) an ontology allow the metadata to accommodate changes in terminology meaning or usage over time. Usage of globally unique identifiers to ontology terms also enable the interoperability of data with systems that expect these well-defined codes as input.

The National Library of Medicine recently announced their intent to modernize *ClinicalTrials.gov* [61], solicited feedback from stakeholders, and on April 28, 2020, posted a summary of the responses they received.[62] Among other questions, the Request for Information (RFI) asked respondents for suggestions on how to increase the use of data standards while maintaining enough flexibility that PRS users can submit accurate information. Several respondents suggested standardizing the vocabulary used in records by encouraging greater use of well-known controlled terminologies, and some suggested extending MeSH with more specific terms, as we do. Respondents also requested the ability to search for an exact phrase (i.e., without synonyms) and the ability to search by disease subtype (which restricting values to a hierarchy such as MeSH would provide).

We found that the reusability of clinical study metadata internationally is hindered by the lack of a single minimum information standard, and by the discrepancies between the 24 fields

in the WHO Trial Registration Data Set[11] and the 41 fields in FDAAA801. FDAAA801 does not require a principal investigator or contact for scientific inquiries to be listed, and 12% of interventional trials fail to list any principal investigator. Many of the fields which are shared between the WHO data set and FDAAA801 have different names and definitions within the two standards. Both standards have seen multiple updates and changes to required elements since their first publication, so older records in *ClinicalTrials.gov* (especially those added before the final rule-related site update) are often missing fields that were added or became required later. The maintainers of *ClinicalTrials.gov* have done an admirable job of parsing unstructured data such as the study design information from old records into the new structured format, but details from the original unstructured data have almost certainly been lost. Automated validation rules in the PRS system have been moderately successful at ensuring required fields are filled for trials after January 18, 2017, but important fields such as the method of allocation of patients to study arms, contact information, and individual site status for recruiting trials are still often missing.

Metanalyses involving the principal investigators of trials are further hindered by the fact that principal investigator information may be listed as part of either the "responsible party" element, the "overall official" element, both, or neither, and the representation of names within the XML records differs between the two elements (*investigator_full_name* vs. *first_name*, *middle_name*, and *last_name*). Like Chaturvedi et al.,[45] we recommend that investigator information be augmented or replaced with persistent unique identifiers such as Open Researcher and Contributor IDs (ORCID). Using ORCIDs would ensure that investigator information is consistent across multiple trials and within multiple listings in the same record. Currently within the PRS, investigator information must be manually reentered everywhere it occurs, which may validly be as the *responsible party*, as the *overall official*, and as the site-specific investigator for one or more locations for a single trial. Using ORCIDs also allows for a researcher's name, degrees, and affiliation to change with time and to be simultaneously updated in all records.

The "Summary of Responses to Request for Information (RFI)" finds that *ClinicalTrials.gov* is commonly used by patients and health care providers searching for studies that are recruiting participants. Some respondents requested the ability to search for studies by eligibility criteria, a more prominent and detailed display of eligibility criteria, or structured information for the inclusion and exclusion criteria. Developing structured representations of inclusion and exclusion criteria which may be reused in future studies, or used to automatically match eligible patients (e.g., from a hospital's patient database) is an active area of research.[63,64] Cohort definition and recruitment are among the most challenging aspects of conducting clinical trials,[65] and difficulties in recruitment cause delays for the majority of trials.[66,67] Several groups have developed structured representations and grammars for eligibility criteria,[64,68] but there is no expert consensus on a system and many remaining unsolved problems.

There are, however, several improvements that could be made to the *eligibility criteria* field in *ClinicalTrials.gov* that would facilitate searching records by this field and ease reuse of criteria for patient-trial matching. First, *inclusion criteria* and *exclusion criteria* should become separate fields. This change would eliminate the need for user-defined headers, and fix 37% of existing errors, as **Table 6** shows. Currently, exclusion criteria that are mistakenly parsed as inclusion criteria due to a missing or malformed header will be interpreted as expressing the

[11] https://www.who.int/ictrp/network/trds/en/

logical opposite of their intended meaning. Further, the PRS should allow users to enter multiple blocks of inclusion and exclusion criteria and an associated criteria group name, such as the label of the corresponding study arm. If we assume all records with eligibility criteria that followed the recommended format did not define criteria for multiple sub-groups, we can estimate that % of trials have criteria for multiple groups. However, there are almost certainly trials in which the full protocol contained eligibility criteria for multiple groups, but the metadata author entered a simplified version of the protocol due to the lack of an appropriate field, so the true percentage is likely much higher.

Although the registrations in *ClinicalTrials.gov* serve an important primary purpose of enabling FDA oversight and protecting human subjects, they are also an invaluable source of metadata about clinical trials for systematic reviews, adverse events [11,12], and analyses about funding sources[13–17], study design,[18,20,23] time to publication following study completion[1,12,24–26], geographical availability of trials,[27–32] causes of delays and early terminations,[33–40] and more. *ClinicalTrials.gov* is also an important resource for patients and health care providers search for studies for which patients are eligible. It is therefore exciting that the NLM has stated their intent to modernize *ClinicalTrials.gov*, with a focus on improving data interoperability and reuse, and on serving needs and users beyond its original purpose. We hope that the findings of our analysis are useful for NLM. With this analysis, we intended to highlight the potential for health-care advancements from making *ClinicalTrials.gov* records findable by easily constructed and specific search queries, interoperable with other knowledge sources, and reusable in statistical analyses of multiple studies.

FUNDING STATEMENT

This work was supported by American Heart Association Institute for Precision Cardiovascular Medicine grant 18IDHP34660267 and by the U.S. National Institutes of Health under grants AI117925 and GM21724.

COMPETING INTERESTS STATEMENT

This manuscript has not been published and is not under consideration for publication elsewhere, and we have no conflicts of interest to disclose.

CONTRIBUTORSHIP STATEMENT

L.M.: Study design, data collection, software implementation, study execution, data analysis, and manuscript writing. R.S.G.: Study design, software implementation, manuscript review and editing. M.M.: Conceptualization, study design, manuscript review and editing.

REFERENCES

1 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

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

3 Halfpenny NJ, Thompson JC, Quigley JM, *et al.* Clinical Trials Registries For Systematic Reviews – An Alternative Source For Unpublished Data. *Value Heal* 2015;**18**:A12. doi:10.1016/j.jval.2015.03.078

4 Keil LG, Platts-Mills TF, Jones CW. Systematic Reviews Published in Emergency Medicine Journals Do Not Routinely Search Clinical Trials Registries: A Cross-Sectional Analysis. *Ann Emerg Med* 2015;**66**:424-427.e2. doi:10.1016/j.annemergmed.2014.10.001

5 van Enst WA, Scholten RJPM, Hooft L. Identification of Additional Trials in Prospective Trial Registers for Cochrane Systematic Reviews. *PLoS One* 2012;**7**:e42812. doi:10.1371/journal.pone.0042812

6 Lanera C, Minto C, Sharma A, *et al.* Extending PubMed searches to ClinicalTrials.gov through a machine learning approach for systematic reviews. *J Clin Epidemiol* 2018;**103**:22–30. doi:10.1016/j.jclinepi.2018.06.015

7 Baudard M, Yavchitz A, Ravaud P, *et al.* Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. *BMJ* 2017;:j448. doi:10.1136/bmj.j448

8 Hart B, Lundh A, Bero L. Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses. *BMJ* 2012;**344**:d7202–d7202. doi:10.1136/bmj.d7202

9 Jones CW, Keil LG, Weaver MA, *et al.* Clinical trials registries are under-utilized in the conduct of systematic reviews: a cross-sectional analysis. *Syst Rev* 2014;**3**:126. doi:10.1186/2046-4053-3-126

10 Chan A-W, Hróbjartsson A, Haahr MT, *et al.* Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials. *JAMA* 2004;**291**:2457. doi:10.1001/jama.291.20.2457

11 Hughes S, Cohen D, Jaggi R. Differences in reporting serious adverse events in industry sponsored clinical trial registries and journal articles on antidepressant and antipsychotic drugs: a cross-sectional study. *BMJ Open* 2014;**4**:e005535–e005535. doi:10.1136/bmjopen-2014-005535

12 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

13 Ehrhardt S, Appel LJ, Meinert CL. Trends in National Institutes of Health Funding for Clinical Trials Registered in ClinicalTrials.gov. *JAMA* 2015;**314**:2566. doi:10.1001/jama.2015.12206

14 Shea M. Data Driven: Mining ClinicalTrials.gov - Institutions, Funding, And Areas of Focus. Friends Cancer Res. 2017.https://www.focr.org/blog/engaging-innovation/data-driven-mining-clinicaltrialsgov-–-institutions-funding-and-areas-focus (accessed 31 Mar 2020).

15 Ramagopalan S V, Skingsley AP, Handunnetthi L, *et al.* Funding source and primary outcome changes in clinical trials registered on ClinicalTrials.gov are associated with the reporting of a statistically significant primary outcome: a cross-sectional study. *F1000Research* 2015;**4**:80. doi:10.12688/f1000research.6312.2

16 Wagner DE, Turner L, Panoskaltsis-Mortari A, *et al.* Co-opting of ClinicalTrials.gov by patient-funded studies. *Lancet Respir Med* 2018;**6**:579–81. doi:10.1016/S2213-2600(18)30242-X

17 Giel JL. Comparison of results reporting on ClinicalTrials.gov by funding source. *Curr Med Res Opin* 2011;**27(Suppl 1**.https://www.ismpp.org/assets/docs/Education/AnnualMeeting/7thAM/OralPresentations/comparison of results reporting on clinicaltrials.gov by funding source.pdf

18 Inrig JK, Califf RM, Tasneem A, *et al.* The Landscape of Clinical Trials in Nephrology: A Systematic Review of ClinicalTrials.gov. *Am J Kidney Dis* 2014;**63**:771–80.

doi:10.1053/j.ajkd.2013.10.043
19 WU M, SIROTA M, BUTTE AJ, *et al.* CHARACTERISTICS OF DRUG COMBINATION THERAPY IN ONCOLOGY BY ANALYZING CLINICAL TRIAL DATA ON CLINICALTRIALS.GOV. In: *Biocomputing 2015*. WORLD SCIENTIFIC 2014. 68–79. doi:10.1142/9789814644730_0008
20 Hill KD, Chiswell K, Califf RM, *et al.* Characteristics of pediatric cardiovascular clinical trials registered on ClinicalTrials.gov. *Am Heart J* 2014;**167**:921-929.e2. doi:10.1016/j.ahj.2014.02.002
21 Cihoric N, Tsikkinis A, Miguelez CG, *et al.* Portfolio of prospective clinical trials including brachytherapy: an analysis of the ClinicalTrials.gov database. *Radiat Oncol* 2016;**11**:48. doi:10.1186/s13014-016-0624-8
22 Bell SA, Tudur Smith C. A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of ClinicalTrials.gov. *Orphanet J Rare Dis* 2014;**9**:170. doi:10.1186/s13023-014-0170-0
23 Zwierzyna M, Davies M, Hingorani AD, *et al.* Clinical trial design and dissemination: comprehensive analysis of clinicaltrials.gov and PubMed data since 2005. *BMJ* 2018;:k2130. doi:10.1136/bmj.k2130
24 Ross JS, Mulvey GK, Hines EM, *et al.* Trial Publication after Registration in ClinicalTrials.Gov: A Cross-Sectional Analysis. *PLoS Med* 2009;**6**:e1000144. doi:10.1371/journal.pmed.1000144
25 Ross JS, Mocanu M, Lampropulos JF, *et al.* Time to Publication Among Completed Clinical Trials. *JAMA Intern Med* 2013;**173**:825. doi:10.1001/jamainternmed.2013.136
26 Jones CW, Handler L, Crowell KE, *et al.* Non-publication of large randomized clinical trials: cross sectional analysis. *BMJ* 2013;**347**:f6104–f6104. doi:10.1136/bmj.f6104
27 Dechartres A, Boutron I, Trinquart L, *et al.* Single-Center Trials Show Larger Treatment Effects Than Multicenter Trials: Evidence From a Meta-epidemiologic Study. *Ann Intern Med* 2011;**155**:39. doi:10.7326/0003-4819-155-1-201107050-00006
28 Galsky MD, Stensland KD, McBride RB, *et al.* Geographic Accessibility to Clinical Trials for Advanced Cancer in the United States. *JAMA Intern Med* 2015;**175**:293. doi:10.1001/jamainternmed.2014.6300
29 Lu C, Yang J, Yu W, *et al.* Association between 25(OH)D Level, Ultraviolet Exposure, Geographical Location, and Inflammatory Bowel Disease Activity: A Systematic Review and Meta-Analysis. *PLoS One* 2015;**10**:e0132036. doi:10.1371/journal.pone.0132036
30 Suda KJ, Greene J, Shelton CM. Geographic location of antiretroviral clinical trials in HIV infected pediatric patients. *Int J Clin Pharm* 2013;**35**:1203–7. doi:10.1007/s11096-013-9849-x
31 Luo J, Wu M, Chen W. Geographical Distribution and Trends of Clinical Trial Recruitment Sites in Developing and Developed Countries. *J Health Inform Dev Ctries* 2017;**11**.
32 Thiers FA, Sinskey AJ, Berndt ER. Trends in the globalization of clinical trials. *Nat Rev Drug Discov* 2008;**7**:13–4. doi:10.1038/nrd2441
33 Baldi I, Lanera C, Berchialla P, *et al.* Early termination of cardiovascular trials as a consequence of poor accrual: analysis of ClinicalTrials.gov 2006–2015. *BMJ Open* 2017;**7**:e013482. doi:10.1136/bmjopen-2016-013482
34 Bernardez-Pereira S, Lopes RD, Carrion MJM, *et al.* Prevalence, characteristics, and predictors of early termination of cardiovascular clinical trials due to low recruitment: Insights from the ClinicalTrials.gov registry. *Am Heart J* 2014;**168**:213-219.e1. doi:10.1016/j.ahj.2014.04.013
35 Williams RJ, Tse T, DiPiazza K, *et al.* Terminated Trials in the ClinicalTrials.gov Results Database: Evaluation of Availability of Primary Outcome Data and Reasons for Termination. *PLoS One* 2015;**10**:e0127242. doi:10.1371/journal.pone.0127242
36 Stensland KD, McBride R, Wisnivesky JP, *et al.* Premature termination of genitourinary cancer clinical trials. *J Clin Oncol* 2014;**32**:288–288. doi:10.1200/jco.2014.32.4_suppl.288
37 Guinn D, Wilhelm EE, Shoulson I. Reasons for Premature Conclusion of Late Phase

Clinical Trials: An Analysis of ClinicalTrials.gov Registered Phase III Trials. *Ther Innov Regul Sci* 2020;**54**:232–9. doi:10.1007/s43441-019-00050-0
38 Cui Z, Seburg EM, Sherwood NE, *et al.* Recruitment and retention in obesity prevention and treatment trials targeting minority or low-income children: a review of the clinical trials registration database. *Trials* 2015;**16**:564. doi:10.1186/s13063-015-1089-z
39 Haddad RI, Chan AT, Vermorken JB. Barriers to clinical trial recruitment in head and neck cancer. *Oral Oncol* 2015;**51**:203–11. doi:10.1016/j.oraloncology.2014.12.007
40 Cook WA, Doorenbos AZ. Indications of Recruitment Challenges in Research with U.S. Military Service Members: A ClinicalTrials.gov Review. *Mil Med* 2017;**182**:e1580–7. doi:10.7205/MILMED-D-16-00225
41 Moja LP, Moschetti I, Nurbhai M, *et al.* Compliance of clinical trial registries with the World Health Organization minimum data set: a survey. *Trials* 2009;**10**:56. doi:10.1186/1745-6215-10-56
42 Miller JE, Korn D, Ross JS. Clinical trial registration, reporting, publication and FDAAA compliance: a cross-sectional analysis and ranking of new drugs approved by the FDA in 2012. *BMJ Open* 2015;**5**:e009758. doi:10.1136/bmjopen-2015-009758
43 Califf RM. Characteristics of Clinical Trials Registered in ClinicalTrials.gov, 2007-2010. *JAMA* 2012;**307**:1838. doi:10.1001/jama.2012.3424
44 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
45 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
46 Tse T, Fain KM, Zarin DA. How to avoid common problems when using ClinicalTrials.gov in research: 10 issues to consider. *BMJ* 2018;:k1452. doi:10.1136/bmj.k1452
47 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
48 Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. *Figshare* 2018;:1–26.
49 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
50 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
51 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
52 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
53 Brazma A, Hingamp P, Quackenbush J, *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;**29**:365–71. doi:10.1038/ng1201-365
54 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
55 World Health Organization. WHO Trial Registration Data Set (Version 1.3.1). https://www.who.int/ictrp/network/trds/en/
56 Food and Drug Administration Amendments Act of 2007. 2007.
57 ClinicalTrials.gov Protocol Registration Data Element Definitions for Interventional and Observational Studies.
58 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

59 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
60 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
61 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/
62 Summary of Responses to Request for Information (RFI): ClinicalTrials.gov Modernization. 2019. https://prsinfo.clinicaltrials.gov/SummaryResponsesToRFI.pdf
63 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
64 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
65 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
66 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
67 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
68 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

Figure Legends

Figure 1 - 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').

Figure 2 - 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 than half of all drugs and radiation therapies use ontology terms, but less than 15% of listed devices and combination products do.

Help Definitions

* Study Type: Interventional

* § Primary Purpose: Treatment

* Study Phase: Phase 1

Use "N/A" for trials that do not involve drug or biologic products.

WARNING: Phase 1 studies typically have at least one Intervention Type of Drug, Biological/Vaccine or Combination Product.

* § Interventional Study Model: --Select--

WARNING: Interventional Study Model has not been entered.

Model Description:

* § Number of Arms: 3.

ERROR: Number of Arms needs to be a whole number.
WARNING: Number of Arms has not been entered.

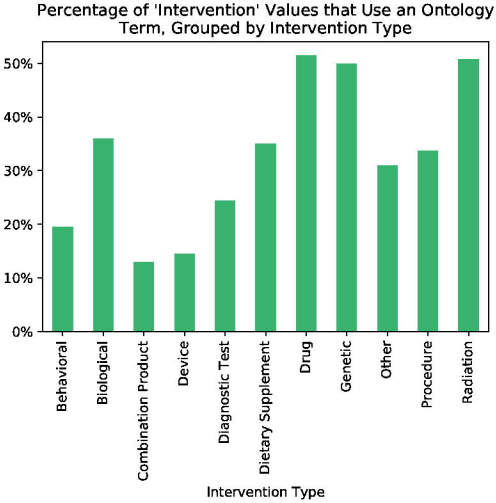
* § Masking:

- [ ] Participant
- [ ] Care Provider
- [ ] Investigator
- [ ] Outcomes Assessor

- [ ] None (Open Label)

Check all roles that are masked or check None (Open Label).

NOTE: Masking has not been entered.


Percentage of 'Intervention' Values that Use an Ontology Term, Grouped by Intervention Type
50%
40%
30%
20%
10%
0%
Behavioral
Biological
Combination Product
Device
Diagnostic Test
Dietary Supplement
Drug
Genetic
Other
Procedure
Radiation
Intervention Type