Article

CLINICAL TRIALS

Clinical Trials
2022, Vol. 19(3) 337–346

DOI: 10.1177/17407745221087456
journals.sagepub.com/home/ctj
SAGE

Results publications are inadequately linked to trial registrations: An automated pipeline and evaluation of German university medical centers

Maia Salholz-Hillel, Daniel Strech and Benjamin Gregory Carlisle

Abstract

Background/Aims: Informed clinical guidance and health policy relies on clinicians, policymakers, and guideline developers finding comprehensive clinical evidence and linking registrations and publications of the same clinical trial. To support the finding and linking of trial evidence, the World Health Organization, the International Committee of Medical Journal Editors, and the Consolidated Standards of Reporting Trials ask researchers to provide the trial registration number in their publication and a reference to the publication in the registration. This practice costs researchers minimal effort and makes evidence synthesis more thorough and efficient. Nevertheless, trial evidence appears inadequately linked, and the extent of trial links in Germany remains unquantified. This cross-sectional study aims to evaluate links between registrations and publications across clinical trials conducted by German university medical centers and registered in ClinicalTrials.gov or the German Clinical Trials Registry. Secondary aims are to develop an automated pipeline that can be applied to other cohorts of trial registrations and publications, and to provide stakeholders, from trialists to registries, with guidance to improve trial links.
Methods: We used automated strategies to download and extract data from trial registries, PubMed, and results publications for a cohort of registered, published trials conducted across German university medical centers and completed between 2009 and 2017. We implemented regular expressions to detect and classify publication identifiers in registrations, and trial registration numbers in publication metadata, abstracts, and full-texts.
Results: In breach of long-standing guidelines, 75% (1,418) of trials failed to reference trial registration numbers in both the abstract and full-text of the journal article in which the results were published. Furthermore, 50% (946) of trial registrations did not contain links to their results publications. Seventeen percent (327) of trials had no links, so that associating registration and publication required manual searching and screening. Overall, trials in ClinicalTrials.gov were better linked than those in the German Clinical Trials Registry; PubMed and registry infrastructures appear to drive this difference. Trial registration numbers were more likely to be transferred to PubMed metadata from abstracts for ClinicalTrials.gov trials than for German Clinical Trials Registry trials. Most (78%, 662/849) ClinicalTrials.gov registrations with a publication link were automatically indexed from PubMed metadata, which is not possible in the German Clinical Trials Registry.
Conclusions: German university medical centers have not comprehensively linked trial registrations and publications, despite established recommendations. This shortcoming threatens the quality of evidence synthesis and medical practice, and burdens researchers with manually searching and linking trial data. Researchers could easily improve this by copy-and-pasting references between their trial registrations and publications. Other stakeholders could build on this practice, for example, PubMed could capture additional trial registration numbers using automated strategies (like those developed in this study), and the German Clinical Trials Registry could automatically index publications from PubMed.

Keywords
Clinical trials, registration, reporting, meta-research, research transparency

Introduction

Linking trial registrations and results publications makes each more findable and improves research transparency at minimal effort to researchers. Threaded

QUEST Center for Responsible Research, Berlin Institute of Health (BIH), Charité—Universitätsmedizin Berlin, Berlin, Germany

Corresponding author:
Maia Salholz-Hillel, QUEST Center for Responsible Research, Berlin Institute of Health (BIH), Charité—Universitätsmedizin Berlin, Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany.
Email: maia.salholz-hillel@bih-charite.de

evidence also empowers readers to cross-check sources, for example, for potential outcome switching.[1] If trial results cannot be found or linked to their registration, systematic reviewers may miss relevant data and draw invalid conclusions.[2] Health policy decisions and clinical guidelines rely on such evidence synthesis, and incomplete evidence can misinform subsequent clinical trials, drive the misallocation of healthcare resources, and risk patient wellbeing.[3,4]

To inform clinical decision-making with comprehensive trial evidence, legal regulations and ethical guidelines advocate results transparency.[5–9] Trial results should be bidirectionally linked, meaning both a reference to publications in the registration, and trial registration numbers (TRNs) in the full-text, abstract, and metadata of publications. The Consolidated Standards of Reporting Trials (CONSORT) as well as the International Committee of Medical Journal Editors (ICMJE) ask trialists to report the "trial registration number and name of the trial register" in both the full-text and abstract of trial results publications.[10,11] Reporting a TRN solely in the full-text does not suffice, since readers may not have access to the full-text or may screen a trial on the abstract alone.[12] Abstracts may also be published independently of a full-text publication, such as for conferences.[11]

The inclusion of TRNs in publication and bibliographic metadata especially enhances machine readability and discoverability. Trial results can then be found more efficiently using TRNs to search publication databases.[13] Publications can then also be automatically linked within trial registrations, as is currently done with ClinicalTrials.gov using TRN metadata from PubMed.[14] Such references to results publications within a registration allow readers to quickly identify and navigate to the trial findings. Results references may also be entered manually in selected registries, with varying degrees of structure.[2,15,16] ClinicalTrials.gov, for example, provides fields for digital object identifier (DOI) and PubMed identifier (PubMed ID), whereas the German Clinical Trials Registry (DRKS) offers an unstructured free-text field. While registry readers may be able to find the referenced publication from the free-text, unique identifiers provide a more robust and structured link that allows for automated retrieval.[13]

Trial evidence appears to be insufficiently linked in both registrations and publications. Previous studies on trial registrations found links to results publications in as few as 13% to as many as 60% of trials.[13,17,18] TRN reporting in either the metadata, abstract, or full-text of trial publications has been found to be as low at 8% and as high as 97%, varying widely depending on how the trial population was defined.[17–21] The sampling strategies used in previous studies limit our ability to draw conclusions about registration-publication linking. Publication-based cohorts relied on inaccurate PubMed clinical trial filters[22] and included publications beyond trial results (e.g. protocols), as well as trials with unknown registration rates, as not all trials are registered.[23–25] Registration-based cohorts included trials that may not have an associated publication and thus should not be expected to have a publication link.[26–28] Furthermore, these studies of publication links in registrations have primarily focused on ClinicalTrials.gov, and we are not aware of previous studies that investigate publication links in DRKS. In this study, we limit our sample to registered clinical trials with published results which can be expected to link the registration and publication, and which allows us to draw conclusions about the prevalence of this responsible research practice.

Objectives

In this exploratory study, we evaluate the links between registrations and results publications in a cross-section of published clinical trials conducted by German university medical centers and registered in either ClinicalTrials.gov or DRKS with completion dates between 2009 and 2017. We also looked at the relationship between different link types, as well as registry and change in practices over time. We developed an automated and scalable approach for data collection and extraction using regular expressions to detect and classify publication identifiers and TRNs, which may be applied to other trial cohorts.

Methods

Automated pipeline for data collection

We used two cohorts of registered clinical trials and associated results previously developed by Wieschowski et al.[29] and Riedel et al.[30] and referred to as the "IntoValue" data set.[31] The data set consists of clinical trials registered on ClinicalTrials.gov or DRKS, conducted by a German university medical center, and completed between 2009 and 2017. Corresponding results publications were found via manual searches.

We downloaded data from both registries on 15 August 2021. We queried ClinicalTrials.gov using the Clinical Trials Transformation Initiative's Aggregate Content of ClinicalTrials.gov via its PostgreSQL database application programming interface.[32] As DRKS does not provide an application programming interface, we built a webscraper to capture the necessary fields.[33]

We queried the PubMed Entrez Programming Utilities application programming interface on 15 August 2021 for all trial results PubMed IDs.[34,35] From the PubMed Extensible Markup Language (XML), we extracted bibliometric information, including the publication abstract and secondary identifier (or databank) metadata.

We used DOIs and PubMed IDs to search for full-text publications as PDFs, using a combination of automated and manual strategies, including contacting the corresponding author as a final step.[36,37] We then used the Grobid machine learning library for technical and scientific publications (v. 0.6.1) via Python to parse the PDFs into machine-readable XMLs.[38,39] To isolate the main body from the publication abstract for our analysis, we extracted the <body> sections of the papers.

Inclusion and exclusion criteria

After updating registry data, we reapplied the IntoValue inclusion criteria: study completion date between 2009 and 2017, interventional, complete based on study status, and conducted by a German university medical center. Trials were considered to be conducted by a German university medical center if one or more were included as a trial sponsor, overall official, and/or responsible party in ClinicalTrials.gov or provided in study addresses in DRKS; trials with a German university medical center as only a facility in ClinicalTrials.gov or recruitment location in DRKS were excluded. See Riedel et al.[30] for further details on these criteria. We limited our sample to trials with results with a PubMed ID and full-text publication.

Detection and classification of publication identifiers in registrations

As each registry formats references to publications differently, we developed parallel approaches to extract publication identifiers for each. For ClinicalTrials.gov, we retrieved the reference type, citation, and PubMed ID fields. We then used a regular expression (regex) to extract any DOIs from the citations. For DRKS, we scraped the reference type, citation, and URL, when available. We used regexes to extract any DOIs and PubMed IDs from the citation and URL. In the rare cases when conflicting DOIs or PubMed IDs were found in the citation versus the URL, we manually reviewed to determine which was valid; if both were valid, we preferred the identifier provided in the citation. In addition, we determined whether a reference was manually or automatically indexed in the registry. DRKS allows only manually added references; for ClinicalTrials.gov, references of type “derived” were marked as automatically indexed. We then used DOIs and PubMed IDs to match the referenced publications to those in our trial data set. We did not attempt to match publications without identifiers (i.e. publication title only).

Detection and classification of TRNs in publications

We developed regexes for the TRN patterns for all registries indexed by PubMed and in the International Clinical Trials Registry Platform registry network (available at https://github.com/maia-sh/ctregistries/blob/master/inst/extdata/registries.csv) and used these to detect and classify TRNs in the PubMed secondary identifier metadata, abstract, and full-text. To gauge the sensitivity and specificity of these regexes, we visually inspected all PubMed secondary identifier metadata. For each source, we first extracted all unique TRN patterns within a source. Our regex allowed for minor formatting errors in the TRN patterns (such as additional punctuation). We then cleaned the TRN patterns to remove these formatting errors, and then deduplicated the corrected TRNs to exclude duplicates within a source uncovered through the cleaning process. We also merged our sources to produce a list of unique TRNs by publication along with the source(s) in which each was reported.

The output of TRN extraction, cleaning, and deduplication included TRNs beyond those of the known registrations. These additional TRNs could be cross-registrations of the same trial or provided as background or discussion. For this study, we were interested in whether the known registrations were reported in the publication.

Analysis

We generated descriptive statistics on trial and publication characteristics, overall and by registry. We calculated the number and proportion of trials linked via the publication (full-text, abstract, and metadata) and the registration. To explore change over time, the relationships between types of links, and differences between registries, we ran logistic regressions for each link type, with all other link types as well as registry and completion year as explanatory variables. Regressions were performed as exploratory analyses, and all variables were included in each regression model. There was no model selection or fitting, or correction for multiple testing. In particular, since PubMed metadata may be generated from the publication abstract or full-text, we examined the relationship between TRN reporting in either the abstract or full-text, and TRN inclusion in the metadata. In addition, since ClinicalTrials.gov registrations automatically reference publications with the TRN in the PubMed metadata (whereas DRKS does not), we examined the proportion of automatically versus manually linked publications in ClinicalTrials.gov. To explore registry differences for only manually linked publications, we excluded automated links and calculated the number and proportion of trials with the publication linked in the registration.

Software, code, and data

Data collection, preparation, and analysis were performed in R (Version 4.1.0).[40] Code to recreate the analysis data set, rerun the analysis, and generate this manuscript is available at https://github.com/maia-sh/reg-pub-link. Raw data (with the exception of the full-text of publications) are available at https://doi.org/10.5281/zenodo.5506434, and code for generating the raw data is available at https://github.com/maia-sh/intovalue-data. A STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) checklist for reporting cross-sectional studies in provided in the Supplemental Material.

Results

The IntoValue data set includes all trials conducted by a German university medical center that were registered on ClinicalTrials.gov or DRKS and completed between 2009 and 2017 (n = 3790). After applying our exclusion criteria, our sample included 1895 trials with 1861 unique results publications indexed in PubMed and available as full-text, as some publications include results from more than one trial. Supplemental Figure 1 provides a flow diagram of the trial and publication screening. Table 1 shows summary descriptive details of the trials with results publications by registry.

Per our inclusion criteria all trials in our sample (n = 1895) were registered and had a publication, however only 373 (19.7%) trials had the most comprehensive registration-publication linking, meaning the publication linked in the registration as well as the TRN in the publication full-text, abstract, and PubMed metadata. Disregarding metadata, which is largely beyond trialists' control, an additional 12 trials had comprehensive linking for a total of 385 (20.3%). An additional 92 (4.9%) met the CONSORT and ICMJE guidelines to include TRNs in both the full-text and the abstract. In contrast, we found 327 (17.3%) trials with no links in either the registration or the publication. The most common linking practice was reporting of the TRN in the full-text only, accounting for 476 (25%) trials. The inclusion of the various link types ranged from 715 (38%) in abstracts to 1137 (60%) in full-text. Table 2 shows registration-publication links overall and by registry. Figure 1 shows the percentage of trials with each combination of links between registration and publication. As PubMed incorporated DRKS as a databank source in 2014, Supplemental Table 1 shows the linking practices in trials published as of 2014 (n = 1400), which reflect similar rates to Table 2.[41]

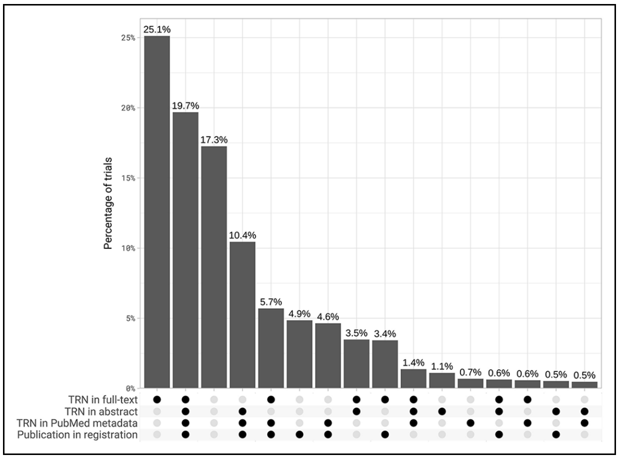

Figure 1. Percentage of trials with each combination of links between registration and publication.
TRN: trial registration number.

Table 3 shows the crude univariate and adjusted multivariate odds ratios (cORs and aORs) for each type of publication–registration link across all explanatory variables. Completion year did not have a strong relationship with linking practices, although trials completed more recently were more likely to report the TRN in the abstract (aOR 1.13 (1.07, 1.19)) and full-text (aOR 1.07 (1.03, 1.12)). Figure 2 shows the rate of registration-publication links over time for ClinicalTrials.gov and DRKS.

Across the adjusted multivariate models, trials with one link type were generally more likely to have another type of link. In particular, TRNs were more likely to appear in the PubMed metadata if included in the abstract (aOR 24.2 (16.2, 37.1)), but not more likely if included in the full-text (aOR 1.33 (0.95, 1.87)). DRKS TRNs were less likely than ClinicalTrials.gov TRNs to appear in the PubMed metadata (aOR 0.22 (0.14, 0.33)). Similarly, for trials with a TRN in either abstract or full-text, ClinicalTrials.gov TRNs appeared at a higher rate than DRKS TRNs in the metadata (62% vs 18%). Supplemental Table 2 shows the number and proportion of trials with a TRN in the metadata given a TRN in the abstract, full-text, or either, both overall and by registry.

Trials registered in DRKS were also less likely than trials registered in ClinicalTrials.gov to reference the primary outcome publication in the registration (aOR 0.39 (0.27, 0.55)). Most (78%, 662/849) ClinicalTrials.gov trials that reference the publication in the registration had the link automatically derived from the PubMed metadata, which is not currently possible for DRKS. Excluding trials with automatically linked publications, ClinicalTrials.gov and DRKS had similar rates of referencing publication in the registration (24%, 187/786 vs 22%, 97/447); the number of manually linked publications in ClinicalTrials.gov may, however, be an underestimate of researcher efforts, since a researcher may attempt to manually link a publication in ClinicalTrials.gov, only to see it was already automatically indexed.

In our manual evaluation of the TRN regular expressions for classifying PubMed secondary identifiers (n = 1296), we found that most ids were clear true positive TRNs (n = 1288). A handful (n = 3) were clear true negative other ids, meaning ids from non-

Table 1. Characteristics of German trials with published results.

	Overall *n* = 1895	ClinicalTrials.gov *N* = 1448	DRKS *n* = 447
Time from trial completion to publication (days), median (IQR)	643 (336, 1015)	644 (323, 1016)	643 (364, 1013)
Time from trial completion to summary results (days), median (IQR)	536 (333, 1017)	544 (338, 1028)	344 (241, 407)
Unknown	1738	1296	442
Registry summary results available, *n* (%)	157 (8%)	152 (10%)	5 (1%)
Prospective registration, *n* (%)	882 (47%)	688 (48%)	194 (43%)
Unknown	2	2	0
Randomized, *n* (%)	1335 (84%)	1018 (89%)	317 (71%)
Unknown	298	298	0
Multicentric trial, *n* (%)	654 (35%)	551 (38%)	103 (23%)
Unknown	1	0	1
Industry sponsor, *n* (%)	273 (14%)	246 (17%)	27 (6%)
Trial enrollment, median (IQR)	69 (33, 156)	70 (34, 160)	60 (31, 150)
Unknown	4	4	0
Phase, *n* (%)			
I	104 (13%)	86 (12%)	18 (26%)
I–II	59 (7%)	57 (8%)	2 (3%)
II	259 (32%)	242 (32%)	17 (24%)
II–III	35 (4%)	34 (5%)	1 (1%)
III	184 (23%)	169 (23%)	15 (21%)
IV	175 (21%)	158 (21%)	17 (24%)
Unknown	1079	702	377
Top six journals, *n* (%)			
PLoS One	72 (41%)	55 (45%)	17 (32%)
Deutsches Arzteblatt International	24 (14%)	5 (4%)	19 (36%)
BMC Cancer	21 (12%)	16 (13%)	5 (9%)
BMC Anesthesiology	19 (11%)	11 (9%)	8 (15%)
Lancet	19 (11%)	15 (12%)	4 (8%)
The New England Journal of Medicine	19 (11%)	19 (16%)	0 (0%)
Trial completion date, *n* (%)			
2009	107 (6%)	97 (7%)	10 (2%)
2010	163 (9%)	131 (9%)	32 (7%)
2011	191 (10%)	156 (11%)	35 (8%)
2012	203 (11%)	161 (11%)	42 (9%)
2013	195 (10%)	160 (11%)	35 (8%)
2014	250 (13%)	183 (13%)	67 (15%)
2015	287 (15%)	208 (14%)	79 (18%)
2016	266 (14%)	179 (12%)	87 (19%)
2017	233 (12%)	173 (12%)	60 (13%)

DRKS: German Clinical Trials Registry; IQR: interquartile range.

A trial was considered randomized if allocation included randomization. A trial was considered prospectively registered if registered in the same or previous months to start date. Summary results were taken from a structured data field in ClinicalTrials.gov, and determined based on manual inspection for terms, such as *Ergebnisbericht* or *Abschlussbericht* in DRKS. Top journals refer to the journals with the greatest number of trial publications in our sample.

Table 2. Registration-publication links overall and by registry.

	Overall *n* = 1895	ClinicalTrials.gov *n* = 1448	DRKS *n* = 447
TRN in full-text	1137 (60%)	882 (61%)	255 (57%)
TRN in abstract	715 (38%)	581 (40%)	134 (30%)
TRN in PubMed metadata	826 (44%)	759 (52%)	67 (15%)
Publication in registration	946 (50%)	849 (59%)	97 (22%)

DRKS: German Clinical Trials Registry; TRN: trial registration number.

registry databanks, such as molecular sequences and open data repositories (i.e. figshare, Dryad).[41] We also found five ids which we manually classified as true negative non-TRNs; however, using additional PubMed

Table 3. Crude (cOR) and adjusted (aOR) odds ratios for factors associated with publication–registration links.

	TRN in full-text		TRN in abstract		TRN in PubMed metadata		Publication in registration	
	cOR	aOR	cOR	aOR	cOR	aOR	cOR	aOR
Completion year	1.08 (1.04, 1.12; < 0.001)	1.07 (1.03, 1.12; < 0.001)	1.07 (1.03, 1.11; < 0.001)	1.13 (1.07, 1.19; < 0.001)	0.98 (0.94, 1.02; 0.3)	0.97 (0.91, 1.04; 0.4)	0.96 (0.92, 0.99; 0.017)	0.95 (0.89, 1.00; 0.067)
Registry								
ClinicalTrials.gov	1 (Ref)	1 (Ref)	1 (Ref)	1 (Ref)	1 (Ref)	1 (Ref)	1 (Ref)	1 (Ref)
DRKS	0.85 (0.69, 1.06; 0.15)	0.76 (0.60, 0.96; 0.020)	0.64 (0.51, 0.80; < 0.001)	4.56 (3.13, 6.75; < 0.001)	0.16 (0.12, 0.21; < 0.001)	0.22 (0.14, 0.33; < 0.001)	0.20 (0.15, 0.25; < 0.001)	0.39 (0.27, 0.55; < 0.001)
TRN in full-text	NA	NA	1.58 (1.30, 1.92; < 0.001)	1.76 (1.36, 2.28; < 0.001)	1.22 (1.02, 1.47; 0.034)	1.33 (0.95, 1.87; 0.10)	0.92 (0.76, 1.10; 0.4)	0.54 (0.40, 0.72; < 0.001)
TRN in abstract	1.58 (1.30, 1.92; < 0.001)	1.78 (1.37, 2.30; < 0.001)	NA	NA	24.3 (18.9, 31.3; < 0.001)	24.2 (16.2, 37.1; < 0.001)	11.4 (9.06, 14.4; < 0.001)	2.37 (1.59, 3.49; < 0.001)
TRN in PubMed metadata	1.22 (1.02, 1.47; 0.034)	1.25 (0.90, 1.74; 0.2)	24.3 (18.9, 31.3; < 0.001)	26.7 (18.0, 40.8; < 0.001)	NA	NA	64.6 (47.8, 88.9; < 0.001)	37.4 (26.0, 54.9; < 0.001)
Publication in registration	0.92 (0.76, 1.10; 0.4)	0.55 (0.42, 0.74; < 0.001)	11.4 (9.06, 14.4; < 0.001)	2.57 (1.72, 3.80; < 0.001)	64.6 (47.8, 88.9; < 0.001)	34.5 (24.0, 50.6; < 0.001)	NA	NA

TRN: trial registration number; DRKS: German Clinical Trials Registry.
Odds ratio (95% CI; p-value).

data, we determined these were severely misformatted DRKS ids (i.e. missing the preceding letters "DRKS" and just a string of numbers, such as "00000711"). Our regexes correctly classified all well-formatted TRNs and non-TRNs, resulting in a sensitivity and specificity of 100%. If we had instead categorized the five severely misformatted DRKS ids as true positives, our regexes would have sensitivity of 99.6% and a specificity of 100%.

Discussion

Linking of trial registrations and results publications plays an important role in research transparency and facilitates comprehensive evidence synthesis and informed health policy decision-making. Poor linking poses a barrier to identifying trial publications via automated approaches and instead requires researchers to perform intensive manual searches to attempt to match publications to trials.[17,26,29] This responsible research practice comes at minimal costs to researchers, from seconds for pasting TRNs in papers, to minutes for adding a publication link to the registration.

Our study shows that German university medical centers can improve in both TRN reporting in publications and references to publications in the registration. In our sample (n = 1895), 17% (327) of trials had no links between registration and publication and only 20% (373) of trials had the most comprehensive registration-publication links. Furthermore, only 25% (477) of trials in our sample fully met the CONSORT and ICMJE guidelines to include TRNs in both the full-text and the abstract. Linking practices showed at best minimal improvement over time. The upward trend in reporting in full-text and abstracts suggests this practice is gaining traction, however, more slowly than advisable per CONSORT and ICMJE guidelines. Trials registered in ClinicalTrials.gov were overall better linked than trials registered in DRKS. These differences are in part beyond trialists' control and reliant on bibliometric databases (i.e. PubMed) and registries, namely, (1) generating PubMed metadata from TRNs in the abstract or full-text and (2) automated indexing of publications in the registry.

Our findings suggest that PubMed's current approach to capturing TRNs in metadata misses TRNs from the full-text as well as DRKS trials. TRNs in PubMed metadata may be either provided by publishers or manually assigned by National Library of Medicine PubMed staff who copy-and-paste TRNs found in the abstract and full-text (personal communication via PubMed Helpdesk Ticket CAS-552810-T3H7 V5). As such, we expect a TRN from any registry to be included in the metadata, if the trialist includes it in either the abstract or the full-text. However, we found that while trials with a TRN in the abstract were

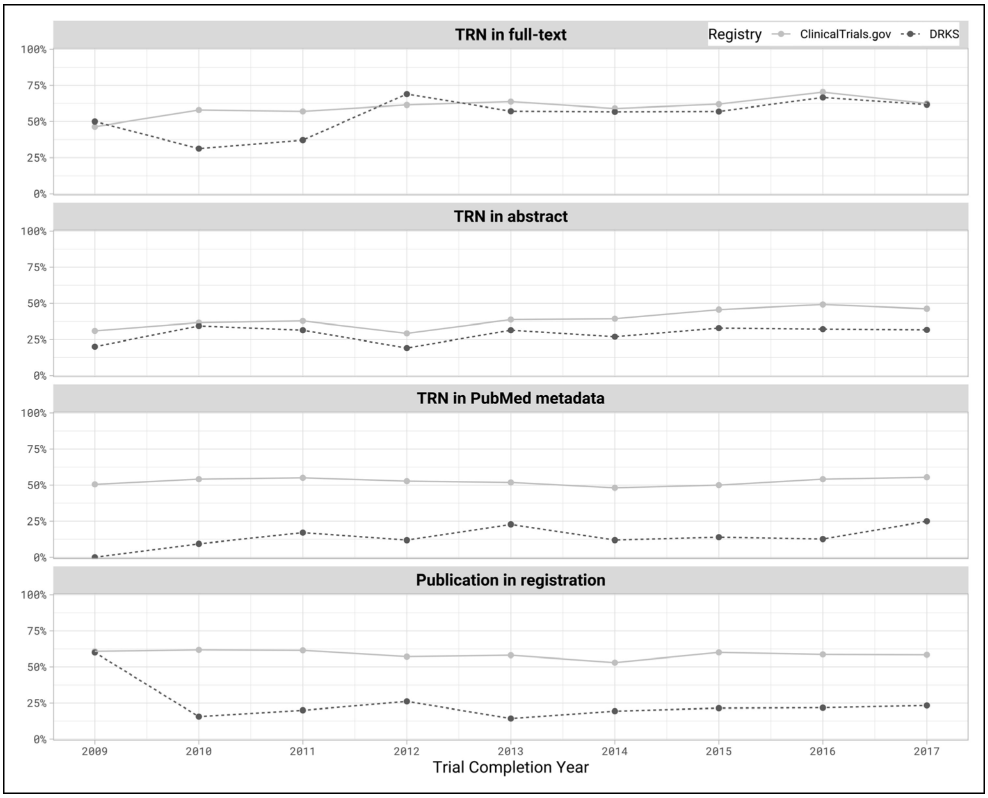

Figure 2. Percentage of trials with linked registrations and publications by trial completion year in ClinicalTrials.gov and DRKS. Completion year from registry.
DRKS: German Clinical Trials Registry; TRN: trial registration number.

indeed more likely to include the TRN in the metadata (aOR 24.2 (16.2, 37.1)), TRNs in the full-text were not more likely to appear in the metadata (aOR 1.33 (0.95, 1.87)). Furthermore, our data suggest that PubMed staff are better at extracting ClinicalTrials.gov than DRKS TRNs (aOR 0.22 (0.14, 0.33)).

Automated indexing of publications in ClinicalTrials.gov accounts for most (78%, 662/849) publication references in the registry and drives the almost three-fold discrepancy with DRKS (59%, 849/1,448 vs 22%, 97/447). Currently, DRKS allows for only manual submission of references by trialists and does not index publications. Furthermore, registrations may reference non-results publications (e.g. background) as well as multiple results publications (e.g. primary, subgroup analyses); registry metadata should encode publication type to support quick identification of primary results. While some publication type metadata is currently available in both registries, it is not systematically used, and many publications are categorized generically as a "paper" (DRKS) or "reference" (ClinicalTrials.gov).

Accurate TRN formatting in publication data is critical for machine-readability, which in turn enables automated indexing of publications in the registration. While our regular expressions allow for and correct minor formatting errors (such as erroneous punctuation), egregious misformatting may make the TRN undetectable. For example, in our visual inspection of the PubMed metadata TRNs, we found five severely misformatted DRKS TRNs (i.e. numbers only with no preceding letters), which we could only identify as DRKS TRNs using additional metadata and which prevented the regex from classifying them as TRNs based on the pattern alone.

Strengths and limitations

This approach has numerous strengths. In contrast with previous studies relying on PubMed queries to identify potential (randomized) clinical trial results publications, we relied on a sample of bona fide results publications from registered trials, allowing us to evaluate the rate of structured links to known results publications in its registration, and the rate of reporting a trial's known TRN in its publication full-text, abstract, and metadata. Furthermore, we used an automated approach including regular expressions with high sensitivity and specificity to identify and classify publication identifiers and TRNs which allowed for a larger sample

Table 4. Recommended stakeholder actions to improve links between trial registrations and publications.

Stakeholder	Recommendation	Example
Researchers	• Include TRN in abstract and full-text per CONSORT and ICMJE guidelines • Link publication in registration	ClinicalTrials.gov provides a step-by-step tutorial for linking publications in the registration.[43]
Registries	• Provide guidance on TRN formatting and reminder to include TRN in publications, and publications in registration • Link publications automatically using TRNs in bibliographic database metadata • Provide metadata on linked publication type (i.e., result, background, protocol)	ClinicalTrials.gov uses PubMed TRN metadata to automatically link publications in the registration[14]. Our study shows this automated publication linking is responsible for the majority of linked publications in ClinicalTrials.gov and implies similar potential for DRKS
Research institutions	• Educate, support, and incentivize researchers on linking	University Medical Centers' ethics review offices or core facilities could give researchers an info sheet with trial registration and reporting guidelines when reviewing trial. Bonuses, that is, *Leistungsorientierte Mittel* (*LOM*), could be disseminated for trial registration, reporting, and linking
Publishers/journals	• Request TRN in specialized metadata field • Review abstract and full-text for TRN inclusion, using automated regexes and/or manual strategies • Provide TRN as metadata to bibliographic databases	Taylor & Francis extracts TRNs from the abstract and full-text, submits this metadata to CrossRef and PubMed, and displays the linked trial via Crossmark on the article page[44,45]
Bibliographic databases	• Integrate publisher-provided TRNs into metadata • Extract TRN as metadata from abstract and full-text, using automated regexes and/or manual strategies	The National Library of Medicine relies on publisher-provided data and manual indexers to create the TRN PubMed metadata if the TRN appears in the abstract or full-text. Our study shows this manual-only strategy misses TRNs and could be semi-automated to detect more TRNs for manual verification

TRN: trial registration number; ICMJE: International Committee of Medical Journal Editors; DRKS: German Clinical Trials Registry; LOM: *Leistungsorientierte Mittel.*

size than manual data extraction would have permitted. This automated strategy is scalable and can be applied to other trial sets.

The approach also faces limitations. Our input IntoValue data set comprised trials conducted by German university medical centers and registered in ClinicalTrials.gov and DRKS and may not reflect practices across other registries and/or countries. Future projects should look at multinational samples of bona fide results publications. Furthermore, we relied on IntoValue for trial deduplication and bona fide results publications. The data set may have had a small number of unaccounted for cross-registrations (e.g. DRKS00004156 and NCT00215683) and publications that are not trial results (e.g. systematic reviews, conference abstract books, etc.). Finally, this automated approach faces limitations of the software on which it is built. While PubMed updated their website in 2020, their application programming interface reflects the previous backend and has subtle differences to the web version (personal communication via PubMed Helpdesk Ticket CAS-575119-Y6D6Y4). While Grobid algorithms are trained on academic papers and have been used in large-scale bibliometric projects (e.g. Wang et al.[42]), parsing PDFs to XMLs may introduce some errors.

Implications for policy and practice

This study reveals shortcomings of German university medical centers in linking trial registrations and results publications and highlights several promising avenues forward. In contrast with other responsible clinical research practices (such as data protection or trial registration itself), registration-publication linking is straightforward and can be improved with action across stakeholders. Table 4 outlines recommendations for stakeholders across clinical research.

With improved TRN inclusion in bibliographic metadata and increased automatic indexing of publications based on this metadata, full linking could be achieved with negligible work by researchers: simply pasting the TRN into the publication abstract and full-text. In our sample alone, an additional 47% (650/1375) of trials with TRN in abstract or full-text could have the TRN included in PubMed's metadata, and an additional 79% (53/67) of DRKS trials with TRN in the PubMed metadata could be automatically indexed in the registration. By adopting the recommended actions, stakeholders can improve trial registration-publication links and foster more comprehensive evidence synthesis and well-informed clinical guidelines and health policy decisions.

Acknowledgements

The authors thank Kerstin Rubarth from the Charité Institut für Biometrie und Klinische Epidemiologie for her consultation on the statistical analysis, and Nicholas Devito for earlier versions of the TRN regular expressions. They also thank their colleague Till Bruckner for useful comments on an earlier version of this manuscript.

Author contributions

The authors made the following contributions. Maia Salholz-Hillel contributed to conceptualization, methodology, software, investigation, data curation, formal analysis, visualization, validation, project administration, writing—original draft preparation, writing—review and editing. Daniel Strech contributed to funding acquisition, resources, conceptualization, writing—review and editing, supervision. Benjamin Gregory Carlisle contributed to writing—review and editing, methodology, software, validation, and supervision.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: MSH, BGC, and DS declare that they have no direct conflict of interest related to this work. MSH is employed as a researcher under the project funding and additional grants from the German Bundesministerium für Bildung und Forschung (BMBF). DS is a member of the Sanofi Advisory Bioethics Committee and receives an honorarium for his contribution to meetings.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded under a grant from the Federal Ministry of Education and Research of Germany (Bundesministerium für Bildung und Forschung—BMBF) [01PW18012]. The funder had no involvement in the study design, data collection, analysis, or interpretation, writing of the manuscript, or the decision to submit for publication.

ORCID iDs

Maia Salholz-Hillel https://orcid.org/0000-0003-1934-9504
Daniel Strech https://orcid.org/0000-0002-9153-079X
Benjamin Gregory Carlisle https://orcid.org/0000-0001-8975-0649

Supplemental material

Supplemental material for this article is available online.

References

1. Altman DG, Furberg CD, Grimshaw JM, et al. Linked publications from a single trial: a thread of evidence. *Trials* 2014; 15: 369.
2. Miron L, Gonçalves RS and Musen MA. Obstacles to the reuse of study metadata in ClinicalTrials.gov. *Sci Data* 2020; 7: 443.
3. Chan A-W, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. *Lancet* 2014; 383: 257–266.
4. Council for International Organizations of Medical Sciences, World Health Organization. *International ethical guidelines for health-related research involving humans*. Geneva: CIOMS, 2016, https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf
5. World Medical Association. Declaration of Helsinki: ethical principles for medical research involving human subjects, 2013, https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/
6. World Health Organization. WHO joint statement on public disclosure of results from clinical trials, 2017, https://www.who.int/ictrp/results/ICTRP_JointStatement_2017.pdf
7. Altman DG and Moher D. Importance of transparent reporting of health research. In: Moher D, Altman D, Wager E, et al. (eds) *Guidelines for reporting health research: a user's manual*. New York: John Wiley & Sons, Ltd, 2014, pp. 1–13.
8. Borysowski J, Wnukiewicz-Kozłowska A and Górski A. Legal regulations, ethical guidelines and recent policies to increase transparency of clinical trials. *Br J Clin Pharmacol* 2020; 86(4): 679–686.
9. Chalmers I, Glasziou P and Godlee F. All trials must be registered and the results published. *BMJ* 2013; 346: f105.
10. International Committee of Medical Journal Editors (ICMJE). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals, 2019, http://www.icmje.org/icmje-recommendations.pdf
11. Hopewell S, Clarke M, Moher D, et al. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS Med* 2008; 5: e20.
12. Hopewell S, Clarke M, Moher D, et al. CONSORT for reporting randomised trials in journal and conference abstracts. *Lancet* 2008; 371: 281–283.
13. Huser V and Cimino JJ. Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials. *PLoS ONE* 2013; 8(7): e68409.
14. National Library of Medicine. How to find results of studies, 2017, https://clinicaltrials.gov/ct2/help/how-find/find-study-results
15. 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.
16. Venugopal N and Saberwal G. A comparative analysis of important public clinical trial registries, and a proposal for an interim ideal one. *PLoS ONE* 2021; 16(5): e0251191.

17. Bashir R, Bourgeois FT and Dunn AG. A systematic review of the processes used to link clinical trial registrations to their published results. *Syst Rev* 2017; 6: 123.
18. van de, Wetering FT, Scholten RJPM, Haring T, et al. Trial registration numbers are underreported in biomedical publications. *PLoS ONE* 2012; 7(11): e49599.
19. Carlisle BG. Non-existent ClinicalTrials.gov identifiers in abstracts indexed by PubMed. *medRxiv* 2020, https://www.medrxiv.org/content/10.1101/2020.02.24.20027300v1.full.pdf
20. Al-Durra M, Nolan RP, Seto E, et al. Prospective registration and reporting of trial number in randomised clinical trials: global cross sectional study of the adoption of ICMJE and Declaration of Helsinki recommendations. *BMJ* 2020; 369: m982.
21. Huser V and Cimino JJ. Evaluating adherence to the International Committee of Medical Journal Editors' policy of mandatory, timely clinical trial registration. *J Am Med Inform Assoc* 2013; 20(e1): e169–e174.
22. Glanville J, Kotas E, Featherstone R, et al. Which are the most sensitive search filters to identify randomized controlled trials in MEDLINE? *J Med Libr Assoc* 2020; 108: 556–563.
23. Trinquart L, Dunn AG and Bourgeois FT. Registration of published randomized trials: a systematic review and meta-analysis. *BMC Med* 2018; 16: 173.
24. Gopal AD, Wallach JD, Aminawung JA, et al. Adherence to the international committee of medical journal editors' (ICMJE) prospective registration policy and implications for outcome integrity: a cross-sectional analysis of trials published in high-impact specialty society journals. *Trials* 2018; 19: 448.
25. Denneny C, Bourne S and Kolstoe SE. Registration audit of clinical trials given a favourable opinion by UK research ethics committees. *BMJ Open* 2019; 9: e026840.
26. Chen R, Desai NR, Ross JS, et al. Publication and reporting of clinical trial results: cross sectional analysis across academic medical centers. *BMJ* 2016; 352: i637.
27. Dwan K, Gamble C, Williamson PR, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias an updated review. *PLoS ONE* 2013; 8(7): e66844.
28. Ross JS, Tse T, Zarin DA, et al. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. *BMJ* 2012; 344: d7292.
29. Wieschowski S, Riedel N, Wollmann K, et al. Result dissemination from clinical trials conducted at German university medical centers was delayed and incomplete. *J Clin Epidemiol* 2019; 115: 37–45.
30. Riedel, N, Wieschowski, S, Bruckner, T, et al. Results dissemination from completed clinical trials conducted at German university medical centers remained delayed and incomplete. The 2014-2017 cohort. *J Clin Epidemiology* 2021; 144:1–7.
31. Riedel N, Wieschowski S, Bruckner T, et al. *Dataset for the IntoValue 1 + 2 studies on results dissemination from clinical trials conducted at German university medical centers completed between 2009 and 2017* [Data set]. Zenodo.
32. Clinical Trials Transformation Initiative (CTTI). Aggregate Content of ClinicalTrials.gov (AACT) Database, https://aact.ctti-clinicaltrials.org/
33. Federal Institute für Drugs and Medical Devices (BfArM). Deutsches register klinischer studien (German Clinical Trials Register) (DRKS), https://www.drks.de/
34. National Library of Medicine. The 9 E-utilities and associated parameters. *The Insider's Guide to Accessing NLM Data*, 30 July 2021, https://dataguide.nlm.nih.gov/eutilities/utilities.html
35. Winter DJ. rentrez: an r package for the NCBI eUtils API. *R J* 2017; 9: 520–526.
36. Our Research. Unpaywall: an open database of 20 million free scholarly articles, https://unpaywall.org/
37. Jahn N. Roadoi: find free versions of scholarly publications via unpaywall, 2021, https://CRAN.R-project.org/package=roadoi
38. Grobid: GeneRation Of BIbliographic Data, https://github.com/kermitt2/grobid
39. Lopez P. Grobid Python client, https://github.com/kermitt2/grobid_client_python
40. R Core Team. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2021, https://www.R-project.org/
41. National Library of Medicine. MEDLINE Databank Sources, https://www.nlm.nih.gov/bsd/medline_databank_source.html
42. Wang LL, Lo K, Chandrasekhar Y, et al. CORD-19: the covid-19 open research dataset. *arXiv* 2020: 200410706, http://arxiv.org/abs/2004.10706
43. National Library of Medicine. ClinicalTrials.gov protocol registration and results system (PRS): entering references information. *ClinicalTrials.gov*, August 2020, https://prsinfo.clinicaltrials.gov/tutorial/content/index.html#/lessons/GE_igGejMjFu9WtErAxXw9-qdeUggVBX
44. Crossref. Linked clinical trials. *Crossref*, https://www.crossref.org/documentation/crossmark/linked-clinical-trials/
45. Taylor & Francis Group. Making a clear link between clinical trials and journal articles, https://authorservices.taylorandfrancis.com/linking-journal-articles-to-clinical-trials/