Clinical trial registries as Scientometric data: A novel solution for linking and deduplicating clinical trials from multiple registries

Christian Thiele¹ · Gerrit Hirschfeld¹ · Ruth von Brachel²

Received: 15 June 2020 / Accepted: 16 July 2021 / Published online: 18 August 2021
© The Author(s) 2021

Abstract

Registries of clinical trials are a potential source for scientometric analysis of medical research and serve important functions for the research community and the public at large. Clinical trials that recruit patients in Germany are usually registered in the German Clinical Trials Register (DRKS) or in international registries such as ClinicalTrials.gov. Furthermore, the International Clinical Trials Registry Platform (ICTRP) aggregates trials from multiple primary registries. We queried the DRKS, ClinicalTrials.gov, and the ICTRP for trials with a recruiting location in Germany. Trials that were registered in multiple registries were linked using the primary and secondary identifiers and a Random Forest model based on various similarity metrics. We identified 35,912 trials that were conducted in Germany. The majority of the trials was registered in multiple databases. 32,106 trials were linked using primary IDs, 26 were linked using a Random Forest model, and 10,537 internal duplicates on ICTRP were identified using the Random Forest model after finding pairs with matching primary or secondary IDs. In cross-validation, the Random Forest increased the F1-score from 96.4% to 97.1% compared to a linkage based solely on secondary IDs on a manually labelled data set. 28% of all trials were registered in the German DRKS. 54% of the trials on ClinicalTrials.gov, 43% of the trials on the DRKS and 56% of the trials on the ICTRP were pre-registered. The ratio of pre-registered studies and the ratio of studies that are registered in the DRKS increased over time.

Keywords Clinical trials · Clinical Trial Registries · DRKS · ClinicalTrials.gov · ICTRP · Random Forest

✉ Christian Thiele
christian.thiele@fh-bielefeld.de

¹ CareTech, OWL – Center for Health, Welfare and Technology, Faculty of Business, University of Applied Sciences Bielefeld, Interaktion 1, 33619 Bielefeld, Germany

² Mental Health Research & Treatment Center, Ruhr-University Bochum, Bochum, Germany

Introduction

Scientometrics is concerned with the measurement of science and scientific research. In order to gain a picture of the research output that is as comprehensive as possible, it is often not enough to scrutinize published articles in scientific journals and books. Instead, in different fields dissertations (Andersen & Hammarfelt, 2011), and patents (Narin, 1995) are routinely used to measure scientific activities. While dissertations and patents are ubiquitous in many domains of research, in medicine, two unique data sources are available—research ethics committees and clinical trial registries. The former derive their importance from the principle that all medical research involving humans needs to be checked by a research committee. In 2008, the seventh Declaration of Helsinki (General Assembly of the World Medical Association, 2014) already included the requirement for every clinical trial to be registered in a publicly accessible registry before enrolling the first patient. Using these data sources, it has become apparent that about a quarter of trials that were planned or performed do not find their way into journal articles (Wieschowski et al., 2019). Also, comparisons between the two have been performed (Denneny et al., 2019). The present study aims to provide an overview of registered clinical trials in Germany and trends in registration by linking relevant clinical trial registries.

The registration of clinical trials is a vital element for improving clinical research (De Angelis et al., 2005; Hillienhof, 2018; Zarin & Keselman, 2007). Specifically, if trials are prospectively registered before the enrollment of the first patient and before running any analyses, i.e., if they are pre-registered, two types of biases are reduced. First, it prevents publication bias, which refers to a higher probability of being published if a study yields favorable results. Secondly, it prevents reporting bias according to which only those results of a study are reported that are desirable (Easterbrook et al., 1991). Of course, these biases are still at play, but the ability to compare pre-registered study information and subsequently released scientific articles makes it possible to quantify these biases. The past research using information from ClinicalTrials.gov has highlighted systematic mismatches between registrations and published articles that are indicative of reporting bias (Anderson et al., 2015; Chan et al., 2004; Hartung et al., 2014; Ramagopalan et al., 2014, 2015; Riveros et al., 2013). In most of these studies, the practice of outcome-switching, where authors change the primary outcome variable from the one in the registration to one that produces more favorable results, has been studied. It is hoped that the potential to be exposed will deter researchers from using this practice and thus will lead to fewer biases in study reporting.

From a research perspective, it is important to note that while the need for clinical trial registries is universally accepted, there is no comprehensive worldwide registry for clinical trials. Although the WHO released a list of mandatory pieces of information in trial data sets (World Health Organization, 2009), which should accordingly be found in the registries, the registries differ in the amount of data that is supplied and especially in the way it is structured. For example, information on outcomes is provided as two unstructured data fields in DRKS and ICTRP but as a relational database in ClinicalTrials.gov via AACT (Table 1). Of course, this affects the type of sociometric analysis that can be readily performed within the databases. It would be straightforward to count the number of outcomes within ClinicalTrials.gov but not on trials registered in DRKS. Differences between the different data sources are described in the methods section.

Despite these limitations, there are numerous examples of studies that use registry data in the context of scientometric analyses. It has been found that articles that are cited in

Table 1 Example of data structuring of outcomes in AACT, DRKS and ICTRP

DRKS/CTP	ClinicalTrials.gov		
	Primary and secondary outcome	Outcome_type	Measure
<p>Primary Outcome: "Change in provocation thresholds from baseline to day 70 in urticaria patients after treatment with omalizumab compared to placebo." "Duration of the visits, threshold will be assessed by using a standardised dermographometer, the Fric Test device."</p> <p>Secondary Outcome: "Safety of patients treated with omalizumab: This includes physical examination, routine safety laboratory assessments, vital signs and adverse event reporting." "Change in quality of life scores, in number of symptom free days, and in physician and patient global assessment of disease severity from baseline to day 70 after treatment with omalizumab compared to placebo." "Long-term effects of omalizumab in UF patients on change in friction thresholds from day 70 (week 10) to day 112 (week 16)." "Quality of life: DLQI, disease activity (Patient and Investigator global assessment)"</p>	<p>[1] "primary"</p> <p>[2] "secondary"</p> <p>[3] "secondary"</p> <p>[4] "secondary"</p> <p>[5] "secondary"</p> <p>[6] "secondary"</p> <p>[7] "secondary"</p>	<p>[1] "Change in Provocation Thresholds From Baseline to Day 70 in Urticaria Patients After Treatment With Omalizumab Compared to Placebo"</p> <p>[2] "To Assess the Effects of Omalizumab in Urticaria Patients on Quality of Life"</p> <p>[3] "To Assess the Effects of Omalizumab in UF Patients on Number of Symptom Free Days"</p> <p>[4] "To Assess the Effects of Omalizumab in UF Patients on Physician Global Assessment of Disease Severity"</p> <p>[5] "To Assess the Effects of Omalizumab in UF Patients on Patient Global Assessment of Disease Severity"</p> <p>[6] "To Assess Long-term Effects of Omalizumab in UF Patients"</p> <p>[7] "To Assess the Safety of Omalizumab in UF Patients"</p>	<p>[1] "Patients receive provocation test by FricTest (standardized stroking of the skin). FricTest ratings are from 0 (no wheal development to the longest pin) to 4 (wheal development to all four pins). The development of wheals within 30 min after provocation is monitored."</p> <p>[2] "Change in quality of life scores assessed by Dermatology Life Quality Index (DLQI) and UF specific life quality questions from baseline to day 70 after treatment with omalizumab compared to placebo."</p> <p>[3] "Change in number of symptom free days as assessed by a patient diary from baseline to day 70 after treatment with omalizumab compared to placebo"</p> <p>[4] "Change in physician global assessment of disease severity assessed by visual analogue scale by a physician from baseline to day 70 after treatment with omalizumab compared to placebo."</p> <p>[5] "Change in patient global assessment of disease severity assessed by visual analogue scale by the patient from baseline to day 70 after treatment with omalizumab compared to placebo."</p> <p>[6] "To assess long-term effects of omalizumab in UF patients, change in friction thresholds from day 70 (week 10) to day 112 (week 16) will be assessed"</p> <p>[7] "Safety of patients treated with omalizumab: This includes physical examination, routine safety laboratory assessments, vital signs and adverse event reporting"</p>

records on ClinicalTrials.gov have higher citation impacts than other articles from the same journals. This could be due to several reasons, for example, simply that highly cited articles are generally more likely to be cited in a clinical trial. The type of citation could not be extracted automatically because that information is not available in the registry (Thelwall & Kousha, 2016). Network analyses on cooperations within trials on ClinicalTrials.gov have revealed that although international trials have become larger in scale, international collaboratives have not. Developing nations have collaborated more within themselves while Europe remained the major contributor to multinational trials (Hsiehchen et al., 2015). ClinicalTrials.gov is arguably the most popular registry for running analyses, probably due to the granularity of its data and availability of processed repositories such as the Aggregate Analysis of ClinicalTrials.gov (AACT). There are many further databases of clinical trials, but an obstacle for simultaneous analyses of multiple databases is the identification of cross-registered trials.

Research questions

The aim of the present study is to scrutinize whether and how registries for clinical trials can be combined into a database that can be used for further analysis to evaluate clinical trials performed in Germany. The research questions are as follows:

1. How many unique trials are registered on which database and how large is the overlap between the databases? Since trials may be registered in one or more registries, a critical first step is to develop robust methods to link identical trials that are registered in different registries and to identify trials that are duplicates. Specifically, we focus on the two primary registries ClinicalTrials.gov and DRKS, and the meta-registry ICTRP to search for and link trials in order to generate a database that is as comprehensive as possible.
2. Can a machine learning model aid the process of linking these databases? Manual identification of duplicates is feasible only for small samples.
3. What are trends regarding the overall number of registered trials, the share of all trials that are preregistered, and the share of trials registered on DRKS? To illustrate the utility of the combined database, we will report on some possible results.

Methods

Data sources

For aggregating a set of trials that have recruited patients in Germany and that is as comprehensive as possible, data from ClinicalTrials.gov via AACT (Clinical Trials Transformation Initiative, 2021), DRKS (DRKS—German Clinical Trials Register, 2021), and ICTRP (ICTRP FullExport Data Set, 2021) were merged. No further filtering was applied, so this includes not only interventional, but also registered observational, patient registry or expanded access studies. There are further primary registries that could be included, but we decided to include DRKS as the register that might contain the largest share of contemporary German studies, ClinicalTrials.gov as probably the most relevant registry that might contain older studies, and the ICTRP to capture further trials. All of the three

(meta-) registries identify trials via a primary ID and also list secondary IDs for many of the trials. The possibilities for automatic download of data from the registries differ quite a lot with regards to the requirement to write custom functions, download data manually and to pre-process that data for further analysis. To obtain fundamental information about the studies and predictors for the following machine learning model, we extracted, among others, the variables pertaining to the study titles in short and official form, primary and secondary outcomes, countries, start date, completion date, registration date, number of patients, study status (e.g., completed), study type, sponsors, and the e-mail addresses of the contact persons. When applicable, we refer to variables from a specific table in the form of *table/variable*. The raw data from the individual registries can be found in the accompanying project on OSF at <https://osf.io/pw4rc/>. Based on the registration date and the start date, it was determined whether a study was pre-registered, which parallels the way the DRKS labels studies as pre-registered.¹

The study type is not consistently reported by the individual registries. Here, the DRKS contains only the values "Interventional" and "Non-interventional". For the purpose of later evaluation, the studies in the category "Non-interventional" were classified as "Observational". The ICTRP contains the categories "Interventional", "Observational", "Diagnostic Test", "Expanded Access", "PMS", "Other", and "Observational [Patient Registry]". We combined the latter category with the "Observational" category. ClinicalTrials.gov separates studies into the categories "Interventional", "Expanded Access", and "Observational [Patient Registry]". We again combined "Observational" and "Observational [Patient Registry]".

ClinicalTrials.gov

ClinicalTrials.gov, operated by the National Library of Medicine (NLM) of the National Institutes of Health (NIH), is the oldest and arguably the most important registry of clinical trials. It was founded in 2000 by the National Institutes of Health (NIH) in association with the Food and Drug Administration (FDA) to provide simple access to clinical trial data for patients, relatives, and researchers. At the end of 2020, it contained about 363,000 studies and during the previous years about 30,000 new studies were registered per year. Also, studies that enroll patients in other countries than the USA may be registered on ClinicalTrials.gov. The registry has undergone several changes during its lifetime, the most important being the inclusion of results data, such as survival rates or baseline measurements, that can be entered after completion of the study. All data on ClinicalTrials.gov can also be accessed via the Aggregate Analysis of ClinicalTrials.gov (AACT) that is provided and updated daily by the Clinical Trials Transformation Initiative (CTTI). It aggregates the data from ClinicalTrials.gov in a highly structured format that is suited very well for statistical analysis. The pieces of information that are mandatory are subject to change and have, for example, been augmented in 2017 to include primary outcomes when a study is an applicable clinical trial in terms of the Public Health Service Act.

A monthly archive with studies registered up to 2021-01-01 was downloaded from AACT during the first weeks of 2021. The AACT provides separate tables that can be easily read in using standard statistical software and joined as needed using the primary ID. These tables represent, for example, the basic study description, measured data, or contacts. The relevant tables for this study are *studies*, *sponsors*, *countries*, *id_information*, *central_contacts*, *result_contacts*, and *design_outcomes*. We filtered for German studies by searching for "Germany" in *countries/countries*. We merged the before-mentioned tables

¹ Personal communication via e-mail on 2020-08-24.

using the *nct_id* variable. As an aid for subsequent linkage of studies, we added *studies/acronym* to the secondary ID. When there was more than one record per unique ID, for example in the case of the recruiting countries, we merged the respective records into a single one so that the result is an aggregated dataset with one row per unique ID. The final AACT data set contained 21,996 studies.

DRKS

The German Clinical Trials Register (DRKS) was founded in 2008 and funded by the Federal Ministry of Education and Research. At the beginning of 2021, it contained around 11,000 studies, while around 1,200 new studies are registered per year (Hillienhof, 2018). Study records can be viewed and downloaded in both German and English. To facilitate matching with the other databases, we downloaded the English entries. Data from the DRKS was downloaded manually via the web interface, where a query is limited to 1000 results, in csv-format. Using multiple queries over different date ranges, all available data from DRKS was downloaded between 2020–09-08 and 2021–02-09. Some care has to be taken when formatting the date variables, since they include inconsistent formats, and when coding missing values because they are represented by an unusual character string. In contrast to the XML-data, which can also be obtained from the DRKS, the csv-files contain a large amount of numbered columns, for example, *recruitmentCountries.value0* to *recruitmentCountries.value19*. We summarised these columns into one by concatenating the individual values to obtain a data set that is less wide and to be able to have only one column per variable. We filtered for German studies by searching for “Germany” in the columns starting with *recruitmentCountries.value*, which applied to 10,219 of 10,990 studies. Similar to AACT, we added entries from the *acronym* column to the secondary ID. Again, the data set contains one row per unique ID. The final DRKS data set contained 10,219 studies that were conducted in Germany.

ICTRP

The International Clinical Trials Registry Platform (ICTRP) by the WHO was founded in 2006 as a meta-registry to provide a single point of access to worldwide clinical trial activity. It aggregates data from primary registries, which have to fulfill certain criteria, partner registries, and data providers (World Health Organization, 2009). The ICTRP currently lists 17 primary registries, two partner registries, and 18 data providers. DRKS is among the primary registries. All of these are also listed as data providers, with ClinicalTrials.gov being the only data provider that is not also a primary registry. The ICTRP likewise supported the manual download of data in XML format via a web interface. However, at the time of the writing of this article, the ICTRP search interface has not been available for several months. Instead, the ICTRP has made available a csv-file that contains a full export of the database (ICTRP FullExport Data Set, 2021), which was downloaded on 2021-02-06 and included studies that were registered until 2021–01-12. Again, some care has to be taken when formatting date and study type variables because the formats are imported from the partner registries and are inconsistent. We used the *Countries* variable to filter for

German studies. Additionally, we kept only studies from clinicaltrialsregister.eu (EUCTR) that ended on “-DE”, because EUCTR lists protocols ending on the various country codes for international studies that usually contain the full list of recruiting countries so that filtering on *Countries* does not suffice. Additionally, around 450 EUCTR studies in ICTRP list Germany as one of the recruiting countries without having a respective protocol ending on “-DE”. We excluded these studies, assuming that they did not actually recruit patients in Germany. The final ICTRP data set had 46,407 records, with one row per primary ID. As a meta-register, the ICTRP uses the partner registries’ IDs as the primary IDs. Acronyms from partner registries can usually already be found in the secondary ID.

Notably, since ICTRP aggregates data from various partner registries, the ICTRP data set contains many internal duplicates. Among those were 41 studies that were not found in AACT. Of those 41 studies, 22 were internal duplicates on ClinicalTrials.gov that are not contained in the *studies* table from AACT because these duplicates were already accounted for and can be found as having the “alias” attribute in *id_information*. The others were, upon manual inspection, not available anymore on ClinicalTrials.gov, and thus, these 41 studies were dropped from ICTRP.

A random forest model for joining studies from AACT, DRKS and ICTRP

It was demonstrated in scientometric studies that Fuzzy Matching and text-based methods for calculating similarity scores, such as the edit-distance, n-grams, or hashing, are suitable for disambiguating database entries (Abdulhayoglu & Thijs, 2018). Typically, these methods are used to compare a specific entry or a set of entries to potential matches or among each other. A problem that arises quickly with this procedure is the complexity of the task that increases exponentially with the size of the database(s). For that reason, studies that use large databases often resort to Hashing to circumvent that problem. In the present study, we were interested in an interpretable model that could be used to identify duplicates and similar articles between registries and that would yield variable importances. We chose to create a Random Forest model for disambiguating potential duplicates because it needs practically no tuning.

To quantify the similarity of two studies in a first step, similarity scores based on multiple variables that were available in all three databases were calculated. For easier interpretability all similarity scores are being normalized to the range [0, 1]. For the study titles a similarity score based on the Levenshtein-Distance was calculated as $\text{stringsim}(s_1, s_2) = 1 - \frac{\text{LED}(s_1, s_2)}{\text{MAXLEN}(s_1, s_2)}$ where LED is the Levenshtein-Distance and MAXLEN the maximal length of the character strings (Wang & Ling, 2012). For the start dates a similarity score was calculated as $\text{datesim}(d_1, d_2) = 1 - \frac{|d_1 - d_2|}{365 \times 25}$ where d_1 and d_2 are the start dates of the studies to be compared. A similarity score for the sample sizes was calculated as $\text{sizesim}(n_1, n_2) = \frac{\min(n_1, n_2)}{\max(n_1, n_2)}$ where n_1 and n_2 are the sample sizes of the two studies.

To create training data for the Random Forest, we calculated the three above-mentioned similarity scores for pairings between AACT and ICTRP where some IDs were missing. That way, we can assess the discrimination quality of the model depending on the presence of matching IDs. Due to the fact that ICTRP contains the vast majority of all DRKS studies, as will be shown later, the model should also generalize well to links between AACT and DRKS. Details on further variables and the importances of all variables in the final model can be found in Table 3. To create a training set with a considerable number of positives, we drew multiple random samples from the pairings with relatively high similarity

scores. The final training data set was labeled manually and contained 472 correct links, 115 incorrect links, and some missing values. This turned out to be quite an imbalance while additionally the real data contained a much larger proportion of negatives. However, it has been shown that for name disambiguation tasks, the Random Forest is not very sensitive to imbalances in the training data (Kim & Kim, 2018). A Random Forest was trained on this manually labeled data set to be able to automatically link studies in all databases. We chose to train the Random Forest with $n_{tree} = 10000$ trees and $m_{try} = 3$ randomly chosen variables per split, where the latter is roughly a third of the total number of predictors. Missing values were imputed with the median.

Join sequence

The three databases have been merged starting from the AACT by gradually adding ICTRP and DRKS entries based on primary IDs, matches as classified by the Random Forest, and ID matches in combination with string similarities, see Fig. 1. By preserving all previously extracted variables of the databases, the entries in different databases can be compared or combined for each study in a later step. The ICTRP as a meta-registry should already contain a large number of studies from the DRKS and ClinicalTrials.gov, but it cannot be assumed that all links are present, neither via the primary nor via the secondary IDs.

During the merging process some typical problems were encountered which made it difficult to merge the databases (see Table 2). In order to achieve the best possible join of the studies, other variables such as the sponsor were used as predictors in addition to the titles.

We define a duplicate study as a record that refers to the same study. This excludes, for example, extension studies or studies such as NCT01247324 and NCT01412333 (OPERA-I and OPERA-II), that have identical protocols but different acronyms and whose resulting data were analyzed individually, as non-duplicates.

The merging of the databases is shown schematically in Fig. 1. All data processing and analysis was conducted using R v4.0.3 (R Core Team, 2018) and the packages tidyverse v1.3.0 (Wickham et al., 2019), lubridate v1.7.9.2 (Grolemund & Wickham, 2011), randomForest 4.6–14 (Liaw & Wiener, 2002), stringdist v0.9.6.3 (Van der Loo, 2014), furrr v0.2.2 (Vaughan & Dancho, 2018), caret v6.0–86 (Kuhn, 2008), and UpSetR v1.4.0 (Conway et al., 2017).

Joining AACT and ICTRP using primary IDs

In a first step, 21,908 studies could already been linked between the AACT database and the ICTRP database using the primary IDs.

Joining remaining AACT studies and ICTRP using the Random Forest model

Another 24 links to still unmatched studies from AACT were identified by the Random Forest model (implying 2,115,912 tested study comparisons). One of which, CTRI/2012/12/003272, was an internal duplicate in ICTRP. All of these links had an ID match in the primary or secondary IDs. This observation speaks both for the ability of the Random Forest to handle extremely imbalanced data and for the relative completeness of links between the databases, if all of the primary and secondary IDs are compared. No

Joining remaining ICTRP studies and aggregate database using secondary IDs and the Random Forest model

 Springer

Table 2 Examples of typical problems when merging the databases

Problem	Example
Identical trials with differing titles	DRKS00010005: A Prospective Cohort Study on Employment and Quality of Life in Patients With Malignant Melanoma NCT02668887: Employment and Quality of Life in Patients With Malignant Melanoma
Multi-part trials, which are no duplicates in terms of this study	NCT01261793: Study of Epratuzumab Versus Placebo in Subjects With Moderate to Severe General Systemic Lupus Erythematosus (SLE) (EMBODY2) NCT01262365: Study of Epratuzumab Versus Placebo in Subjects With Moderate to Severe General Systemic Lupus Erythematosus (EMBODY1)
Different trials with highly similar titles	CTRI/2012/05/002692: Efficacy and Safety Study of Botulinum Toxin Type A Against Placebo to Treat Spasticity in the Leg After a Stroke NCT01392300: Efficacy and Safety Study of Botulinum Toxin Type A Against Placebo to Treat Spasticity in the Arm After a Stroke
Typos	2006–005,473–21: An open label extention [sic] study to evaluate safety, tolerability, and efficacy of 18 weeks of valsartan treatment in children 6 months -5 years old with hypertension
Inconsistent registration of multi-part trials on different registries	NCT01425957, which has parts A and B and can be found under the same ID on ICTRP, has two entries on DRKS, one for each part (DRKS00003721 and DRKS00003705)

secondary IDs, we matched the candidate study to the record from AACT with the highest probability as estimated by the Random Forest model.

We checked only study pairings with matches in any of the IDs in this join step, because the large number of still unmatched studies from ICTRP in combination with the exercise of comparing these to all ~22,000 studies that were in the overall database made the calculation of all necessary similarity metrics computationally expensive. However, since we observed in the previous join step and during the cross-validation of the Random Forest that the most important variable was matching IDs, we decided to proceed as detailed above. 10,537 links were established between still unmatched studies from ICTRP and the overall database. 41 studies that constituted alias-studies from ClinicalTrials.gov were dropped. Lastly, we added the remaining 13,897 studies from ICTRP as new rows, many of which were studies from the DRKS.

Joining DRKS and the aggregated database using primary IDs

When DRKS was joined with the aggregated database using the primary IDs, 10,198 links were identified, leaving 21 unmatched studies from DRKS. We encountered ten studies from DRKS that had multiple matches in this step, which we decided to resolve manually due to the low number. One of these was an internal duplicate in DRKS (DRKS00007604) that was caused by automatic data import from ClinicalTrials.gov for a study that was

separately registered on DRKS. We dropped the former one. Some studies from DRKS have various “internal links” to related studies, which are no duplicates. For example, DRKS00004172 is linked via its secondary IDs to the related studies DRKS00004195, DRKS00004196, DRKS00004197, and DRKS00004198 (and to NCT01625325).

Join remaining studies from DRKS to the aggregate database using the Random Forest model

Of the remaining 21 studies, 2 could be linked to studies from the aggregated database using the Random Forest model (implying 231,273 tested study combinations), one of those links (NCT04571801, DRKS00022782) without having any identical primary or secondary IDs. The remaining 19 studies from DRKS were added as new rows, leading to 35,912 studies in the overall database. This database is available for viewing and download at <https://osf.io/pw4rc/>.

Results

The Random Forest achieved an accuracy of 95.4%, a sensitivity of 96.4%, and a specificity of 91.0% on the manually labeled data in a tenfold cross-validation. This would usually not suffice, however, since the cross-validation was performed on data that was selected on the basis of high similarities, that data contained many observations that were relatively difficult to classify. In comparison to a simple linkage using the secondary IDs, the Random Forest increased the cross-validated F1-score from 96.4% to 97.1% and Cohen’s Kappa from 83.9% to 85.6%. The training set was rather small at $n=587$ study pairings, so more data and practical applications with manual checks are needed to confirm these metrics.

From the variable importances of the Random Forest (see Table 3) it can be seen that, in addition to matching IDs, high importance was given to similarities in titles, sample sizes, and starting dates. Only minor importance was attributed to variables based on e-mail addresses, sponsor names, and recruiting countries.

We made a total of 43,001 links between trials in the three databases. 32,106 of those links were made using primary IDs from ClinicalTrials.gov and DRKS and the primary IDs from the partner registries of ICTRP. 10,869 internal links in ICTRP were made by first matching trials with any matches in the primary or secondary IDs and subsequently checking these links using the Random Forest model. 24 links were made between ClinicalTrials.gov and ICTRP and 2 links between DRKS and ICTRP using that model, one of which without any matching IDs.

The resulting complete database consisted of 35,912 studies. As shown in Fig. 2, 5367 studies were registered in only one register, while the vast majority of 30,545 studies can be found in more than one database. Any of which were studies that were imported from DRKS or ClinicalTrials.gov to ICTRP. Of 10,213 studies from DRKS in the aggregated database 1579 can also be found on ClinicalTrials.gov. Of 11,492 studies on EUCTR 8016 were also registered on ClinicalTrials.gov, and 704 were also registered on DRKS. 3476 studies from EUCTR were not cross-registered in any other registry. While it is still possible that studies were conducted that were not registered at all or registered in another database not indexed by ICTRP, we believe this is a very

comprehensive estimate of studies. The overlap between ICTRP and both of DRKS and ClinicalTrials.gov is nearly complete, but as discussed in a previous chapter, there are some issues regarding updates of the data, so that it is advisable to use the primary registries where possible and if data quality is critical. In total, about 28% or 10,219 of the studies conducted in Germany in the entire period were registered in the DRKS. This is related to the relatively late introduction of the DRKS in 2008. Of all studies with a start of enrollment in 2020 and a recruiting location in Germany, over 50% were registered in the DRKS.

Regarding the number of studies over time, it seems that the total number of studies registered per year has started to flatten off after a phase of almost exponential growth in the early 2000s. Since 2010, the number of new studies per year ranged from 1952 to 2742 in 2020, the latter representing an increase of 212 studies compared to 2019. However, the pattern of databases in which the studies are registered is changing. For example, since the inception of the DRKS there has been a clear trend towards more studies being registered there. This includes retrospectively registered studies, as can be seen by studies with a start date prior to 2008. At the same time, the number of studies with a recruiting location in Germany that are exclusively registered on ClinicalTrials.gov or one of the ICTRP partner registries other than DRKS has been decreasing (Fig. 3).

Finally, some further results can be arrived at using the aggregated database. First, there are differences in the proportion of studies that are actually pre-registered. For example, over the entire period, 54% of the studies on ClinicalTrials.gov and 56% of the studies on the partner registries of the ICTRP were entered into the databases before the start of recruitment, whereas in DRKS this was 43% overall and 45% for the subset of studies that were registered and started after 2009-01-01. So, in terms of pre-registration rates, the studies found on ClinicalTrials.gov and ICTRP are ahead of the DRKS, but the latter has caught up in recent years, see Fig. 4. There is also a tendency for studies to be pre-registered more frequently the larger the number of patients. In general, interventional trials were pre-registered more frequently than observational studies, since the trend towards pre-registration began earlier there, but a steadily increasing pre-registration rate is also apparent for observational studies. Regarding study types, the share of interventional trials among all available studies with a start date in 2000 is about 90%. Studies with a start date in 2018 were interventional to about 50% in the DRKS, about 70% on the ICTRP and about 78% on ClinicalTrials.gov.

In terms of the amount of data available, the registries differ mainly in the degree of data structuring. Using the primary and secondary endpoints as examples, Table 1 shows how the AACT divides this information into different variables, while the DRKS and ICTRP contain the information as a single variable. Therefore, in the latter databases, for example, the information on how many secondary endpoints a study contained would be more difficult to extract and potentially error-prone. In addition, results data are practically only available on ClinicalTrials.gov.

Discussion

The Random Forest model proved useful for disambiguating trials, because some type of classification was necessary when a trial had ID-matches with multiple other trials. It also conveniently combines multiple possible heuristics that might otherwise be used (similarity of titles, sample sizes, start dates, etc.). As we have seen in our data, the usefulness of

Table 3 Variable importances from the Random Forest

Variable	Mean decrease Gini
any_matched_id	87.68
best_stringsim_titles	51.44
start_date_similarities	16.62
n	10.88
size_similarities	10.17
best_stringsim_name	5.44
best_stringsim_email	2.34
any_matched_email	0.01

For details on the similarity functions *stringsim*, *datesim*, and *sizesim* please refer to the methods chapter. any_matched_id: At least one matching primary or secondary ID; best_stringsim_titles: maximum of string_similarities_b_b, string_similarities_off_b, string_similarities_off_off and string_similarities_b_off; string_similarities_b_b: *stringsim* of the brief titles; string_similarities_off_b: *stringsim* of the official title from database A and brief title from database B; string_similarities_off_off: *stringsim* of the official titles; string_similarities_b_off: *stringsim* of brief title from database A and official title from database B; start_date_similarities: *datesim*; size_similarities: *sizesim*; best_stringsim_name: Maximum of *stringsim* for all combinations of sponsor names; best_stringsim_email: Maximum of *stringsim* for all combinations of e-mail addresses; any_matched_email: True, if there was a match in E-mail contact addresses

Fig. 2 Upset plot of overlaps between the databases

Fig. 3 Number of registered studies over time

the model and whether it is worth the additional effort depends on the databases that are to be matched. For example, if only AACT should be matched to the ICTRP, links using primary and secondary IDs are virtually complete. For new database pairings, it may be interesting to check possible links using a machine learning model, if completeness of de-duplications is critical. If the datasets that are to be joined are small, it may be an alternative to flag possible pairings using simple heuristics and to check possible duplicates manually. The model attained good accuracy scores, but we did additional manual checks because we were aiming for as few incorrect links as possible and because we had to assess the correctness of the links that were made by the Random Forest. Additionally, in a multi-step join process with intermediate results, errors early in the process may lead to downstream errors in later join steps.

In principle, registries of clinical studies are supposed to provide a comprehensive overview of all clinical studies that are being carried out. As such, they are a potentially highly relevant source for scientometric analysis. In contrast to published articles, registries are, firstly, much more structured and thus enable a faster and possibly automatic evaluation of key study variables. For example, it is straightforward to calculate how many patients were enrolled in all clinical trials based on registry data, while this information is either absent from more structured data sources such as Web of Science or only found in unstructured text form in scientific articles. Secondly, in contrast to journal articles which are sometimes behind paywalls, the information in registries is freely and publicly accessible. Thirdly, registries suffer less from the otherwise ubiquitous problem of non-publication of studies, since studies can be registered regardless of results or perceived novelty. Fourthly, registries allow a glimpse into the future because they contain information on studies that will often not be published for several years (Easterbrook et al., 1991). The only other source that would be similarly comprehensive for interventional trials would be ethics committees. According to the Helsinki Declaration, an ethics committee must review the ethical safety of every medical study involving humans. However, there is currently neither a list of all ethics committees in Germany nor the possibility to request structured data from them. Therefore, registries remain a practical way to gain an overview of clinical research.

Fig. 4 Number of pre-registered trials over time. Top: Trials with less than 100 patients; Middle: Trials with 100 to 999 patients; Bottom: Trials with at least 1000 patients

However, one must be aware that, strictly speaking, there is a legal requirement in Germany to register only a small part of the clinical studies, namely those that test medicinal products, and many studies are conducted outside the regulatory framework of the German Medicines Act (Pigeot et al., 2019). The other clinical trials might be registered voluntarily, because of journal policies (Taichman et al., 2017) or because registration is necessary for subsequent publication.

However, a closer look at the registrations actually shows that in practice some problems imply that registries cannot, yet, be used as intended. For example, the goal of the DRKS "to provide a complete and up-to-date overview of clinical studies conducted in Germany" (Dreier et al., 2016) cannot be fully achieved because there are a number of other registries in addition to the DRKS in which studies that are conducted in Germany are being registered. For patients and physicians looking for relevant clinical trials, it is regrettable that only about half of all studies recruiting in Germany offer information in German via the DRKS. At the same time, there are several problems associated with cross-registering trials both in the native language in a country-specific as well as in general registers in English or importing them. Cross-Registrations pose a challenge regarding de-duplication, since secondary IDs contained virtually all necessary links, but may also refer to related, non-identical trials. Additionally, internal duplicates may not be readily identifiable (van Valkenhoef et al., 2016). Cross-Registrations furthermore pose a challenge regarding data quality, because many trials that could be found on multiple databases were automatically imported and did not necessarily contain the most recent data.

The present study also makes it clear that it may not suffice to query ClinicalTrials.gov only when studies that recruit in Germany are sought to be analyzed. If studies investigating registries (Anderson et al., 2015; Chan et al., 2004; Hartung et al., 2014; Ramagopalan et al., 2014, 2015; Riveros et al., 2013) refer to ClinicalTrials.gov alone, it seems necessary to replicate the results of the studies in Germany and search several registries, which was indeed carried out for specific subsets of studies (Wieschowski et al., 2019). However, this raises the problem of matching the different entries.

Conclusion

To summarize our findings, we arrived at the following conclusions with regards to the three research questions:

1. We obtained a total number of 35,912 studies with a recruiting location in Germany that were contained in the DRKS, on ClinicalTrials.gov and on the ICTRP. Of those, 21,910 were registered on ClinicalTrials.gov and 10,213 on DRKS. Both registries are virtually completely contained in the ICTRP. 5,367 studies were registered in only one registry, the majority of those (3,476) on the EU Clinical Trials Register.
2. It was possible to deduplicate most of the cross-registered registered studies using IDs. However, in addition to these, there was a considerable number of studies that could only be matched using machine learning or string comparisons. This way, we made a total of 10,563 links. The machine learning model was helpful in disambiguating studies, since the databases overlap quite strongly and records about secondary IDs necessary for identifying duplicates are in rare cases not complete and sometimes refer to non-duplicates. Future improvements could include models that use Natural Language Processing in order to differentiate study titles more reliably and possibly to use further, more comprehensive variables, like inclusion and exclusion criteria or primary and secondary outcomes, which are difficult to classify using only simple string similarity metrics. Depending on the scope of the application, the number of possible study pairings and thus comparisons continues to pose a challenge, especially if more complex methods are to be implemented, so a balance between scalability and classification accuracy has to be struck. Furthermore, manual inspection of the (meta-) registries confirmed that internal duplicates also exist, so a more scalable model for automatically identifying these could prove useful.
3. A look at the temporal trends shows that the number of clinical trials is growing much less rapidly overall than it was about ten to twenty years ago. This is probably due to the fact that the increases in the early noughties were not primarily based on the fact that significantly more studies were actually carried out, but rather on the fact that the proportion of registered studies increased. More important from the perspective of German patients is certainly the fact that the proportion of studies published in the DRKS—and thus in German—has increased over the last ten years. In view of the goal to make clinical research accessible to the public (Dreier et al., 2016), it is important to continue this trend. Our results regarding the share of pre-registered studies are in line with previous results (Trinquart et al., 2018).

In summary, combining different registries and a meta-register while de-duplicating their entries is a viable route for gaining an overview of clinical trials in Germany.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was funded by the German Federal Ministry of Education and Research (BMBF #01PU17010).

Availability of data and material The final combined database and raw data are available at <https://osf.io/pw4rc/>

Code availability R-Code for producing the data and results is available at <https://osf.io/pw4rc/>

Declarations

Conflict of interest The authors declare no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdulhayoglu, M. A., & Thijs, B. (2018). Use of locality sensitive hashing (LSH) algorithm to match Web of Science and Scopus. *Scientometrics*, 116(2), 1229–1245. <https://doi.org/10.1007/s11192-017-2569-6>
- Andersen, J. P., & Hammarfelt, B. (2011). Price revisited: On the growth of dissertations in eight research fields. *Scientometrics*, 88(2), 371–383. <https://doi.org/10.1007/s11192-011-0408-8>
- Anderson, M. L., Chiswell, K., Peterson, E. D., Tasneem, A., Topping, J., & Califf, R. M. (2015). Compliance with results reporting at ClinicalTrials.gov. *The New England Journal of Medicine*, 372(11), 1031–1039. <https://doi.org/10.1056/NEJMsa1409364>
- Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, 291(20), 2457–2465.
- Clinical Trials Transformation Initiative. (2021). *AACT Database*. Retrieved February 16, 2021, from <https://aact.ctti-clinicaltrials.org/download>
- Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18), 2938–2940.
- De Angelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J. P. M., Schroeder, T. V., Sox, H. C., & Van Der Weyden, M. B. (2005). Is This Clinical Trial Fully Registered?: A Statement from the International Committee of Medical Journal Editors*. *Annals of Internal Medicine*, 143(2), 146. <https://doi.org/10.7326/0003-4819-143-2-200507190-00016>
- Denneny, C., Bourne, S., & Kolstoe, S. E. (2019). Registration audit of clinical trials given a favourable opinion by UK research ethics committees. *BMJ Open*, 9(2), e026840. <https://doi.org/10.1136/bmjopen-2018-026840>
- Dreier, G., Jena, S., & Kunzweiler, K. (2016, March 11). *Internetportal: Das Deutsche Register Klinischer Studien*. Deutsches Ärzteblatt. <https://www.aerzteblatt.de/archiv/175249/Internetportal-Das-Deutsche-Register-Klinischer-Studien>
- DRKS - German Clinical Trials Register. (2021). Retrieved February 17, 2021, from https://www.drks.de/drks_web/
- Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, 337(8746), 867–872.
- General Assembly of the World Medical Association. (2014). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *The Journal of the American College of Dentists*, 81(3), 14–18.
- Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3), 1–25.
- Hartung, D., Zarin, D. A., Guise, J.-M., McDonagh, M., Paynter, R., & Helfand, M. (2014). Reporting Discrepancies between the ClinicalTrials.gov Results Database and Peer Reviewed Publications. *Annals of Internal Medicine*, 160(7), 477–483. [/10.7326/M13-0480](https://doi.org/10.7326/M13-0480)
- Hillienhof, A. (2018). *Studienregister: Zugriff auf mehr als 7 000 klinische Studien*. Deutsches Ärzteblatt. <https://www.aerzteblatt.de/archiv/199772/Studienregister-Zugriff-auf-mehr-als-7-000-klinische-Studien>
- Hsiehchen, D., Espinoza, M., & Hsieh, A. (2015). The Cooperative Landscape of Multinational Clinical Trials. *PLOS ONE*, 10(6), e0130930. <https://doi.org/10.1371/journal.pone.0130930>

- ICTRP FullExport data set. (2021). Retrieved February 6, 2021, from <https://bit.ly/3nyMyCK>
- Kim, J., & Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics*, 117(1), 511–526.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1–26.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.
- Narin, F. (1995). Patents as indicators for the evaluation of industrial research output. *Scientometrics*, 34(3), 489–496. <https://doi.org/10.1007/BF02018015>
- Pigeot, I., Foraita, R., Frömke, C., Haller, B., Hoffmann, V. S., & Zapf, A. (2019). Ethische Bewertung von Studien am Menschen außerhalb des regulatorischen Rahmens: Nicht bindend, aber von großer Wichtigkeit. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 62(6), 722–728.
- R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ramagopalan, S. V., Skingsley, A. P., Handunnetthi, L., Klingel, M., Magnus, D., Pakpoor, J., & Goldacre, B. (2014). Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: A cross-sectional study. *F1000Research*, 3. <https://doi.org/10.12688/f1000research.3784.1>
- Ramagopalan, S. V., Skingsley, A. P., Handunnetthi, L., Magnus, D., Klingel, M., Pakpoor, J., & Goldacre, B. (2015). 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*, 4(80), 80.
- Riveros, C., Dechartres, A., Perrudeau, E., Haneef, R., Boutron, I., & Ravaud, P. (2013). Timing and completeness of trial results posted at ClinicalTrials.gov and published in journals. *PLoS Medicine*, 10(12).
- Taichman, D. B., Sahni, P., Pinborg, A., Peiperl, L., Laine, C., James, A., Hong, S.-T., Haileamlak, A., Gologly, L., Godlee, F., Frizelle, F. A., Florenzano, F., Drazen, J. M., Bauchner, H., Baethge, C., & Backus, J. (2017). Data sharing statements for clinical trials: A requirement of the International Committee of Medical Journal Editors. *Annals of Internal Medicine*, 167(1), 63–65. <https://doi.org/10.7326/M17-1028>
- Thelwall, M., & Kousha, K. (2016). Are citations from clinical trials evidence of higher impact research? An Analysis of ClinicalTrials.gov. *Scientometrics*, 109(2), 1341–1351.
- Trinquart, L., Dunn, A. G., & Bourgeois, F. T. (2018). Registration of published randomized trials: A systematic review and meta-analysis. *BMC Medicine*, 16(1), 173. <https://doi.org/10.1186/s12916-018-1168-6>
- Van der Loo, M. P. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1), 111–122.
- van Valkenhoef, G., Loane, R. F., & Zarin, D. A. (2016). Previously unidentified duplicate registrations of clinical trials: An exploratory analysis of registry data worldwide. *Systematic Reviews*, 5(1), 116. <https://doi.org/10.1186/s13643-016-0283-8>
- Vaughan, D., & Dancho, M. (2018). Furr: Apply mapping functions in parallel using futures. *R Package Version 0.2.2*.
- Wang, X., & Ling, J. (2012). Multiple valued logic approach for matching patient records in multiple databases. *Journal of Biomedical Informatics*, 45(2), 224–230. <https://doi.org/10.1016/j.jbi.2011.10.009>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686.
- Wieschowski, S., Riedel, N., Wollmann, K., Kahrass, H., Müller-Ohlraun, S., Schürmann, C., Kelley, S., Kszuk, U., Siegerink, B., Dirnagl, U., Meerpohl, J., & Strech, D. (2019). Result dissemination from clinical trials conducted at German university medical centers was delayed and incomplete. *Journal of Clinical Epidemiology*, 115, 37–45. <https://doi.org/10.1016/j.jclinepi.2019.06.002>
- World Health Organization. (2009). *WHO | WHO ICTRP Registry Criteria*. WHO; World Health Organization. <https://www.who.int/clinical-trials-registry-platform/network/registry-criteria>
- Zarin, D. A., & Keselman, A. (2007). Registering a clinical trial in ClinicalTrials.gov. *Chest*, 131(3), 909–912.