Visualizing clinicaltrials.gov Data for Rapid Assessment of the Clinical Trial Landscape

Jeffery Painter

ABSTRACT

Currently, accessing clinical trial data through the clinicaltrials.gov website is a tedious and cumbersome process, requiring the user to read large quantities of text before being able to find what they need.

The aim of this project is to build a data visualization tool which aids the user in understanding vast quantities of clinical trial data.

By creating a visual inspection tool, knowledge should be more easily gleaned from this public data to enable a better understanding of the clinical trial landscape and, more specifically, any adverse drug reactions (ADRs) [18] disclosed in those trials.

KEYWORDS

clinical trials, drug safety, adverse events, data visualization

1 INTRODUCTION

1.1 Project Goals

This project will scrape clinical trial (CT) data from clinicaltrials.gov (published in XML format) and allow users to explore this data in a graphical representation.

By utilizing the UMLS (Unified Medical Language System) [2], we will connect adverse events into "concept codes" and a network graph which will provide improved search capability and data aggregation using hierarchical and semantic relations.

Data will be stored in a SQLite database, and the UI will be presented as interactive dashboards using Tableau. Further, we will refer to Tang, et al. on best practices for extracting relevant adverse event information, filtering and customizing dashboards [20].

The dashboards will allow users to search and display clinical trials by disease target, therapeutic class or potential adverse events; presenting results in a rapid assessment for understanding the CT landscape.

1.2 Current limitations

Searching for trials using clinicaltrials.gov is difficult at best, requiring much manual effort. Although the website makes it easy to download the data, it is only in XML format making it difficult for non-technical medical staff to utilize [1].

In addition, the search interface on clinicaltrials.gov does not provide any way to search for adverse drug events. The search capability it does have is limited and does not take advantage of ontologies which our tool will provide.

After starting this project, deeper investigation revealed a similar project begun in 2012 called the *The Database for Aggregate Analysis of ClinicalTrials.gov* (AACT) [21]. This project aimed to collect and organize CT data into a structure database available as a download directly from the AACT website.

Our project is distinguished from the AACT primarily in the construction of our data model compared to the AACT. The AACT has major deficiencies that fail to help us answer questions around demographic makeup of clinical trials as well as safety and adverse drug reaction data captured in a standardized way. The AACT also does not provide any data visualization or UI to interact with their database, making it inaccessible to non-technical, medical professional staff.

2 PROPOSED METHOD

2.1 Intuition: Our approach and why it will be successful

There are no tools available today that can help visualize clinicaltrials.gov data. What does exist is mostly geared towards supporting ongoing trials [3] [8]. In a recent article, Du et al expressed "Given the rapid growth of COVID-19 clinical trials, there is an urgent need for a better clinical trial information retrieval tool that supports searching by specifying criteria, including both eligibility criteria and structured trial information." [5]

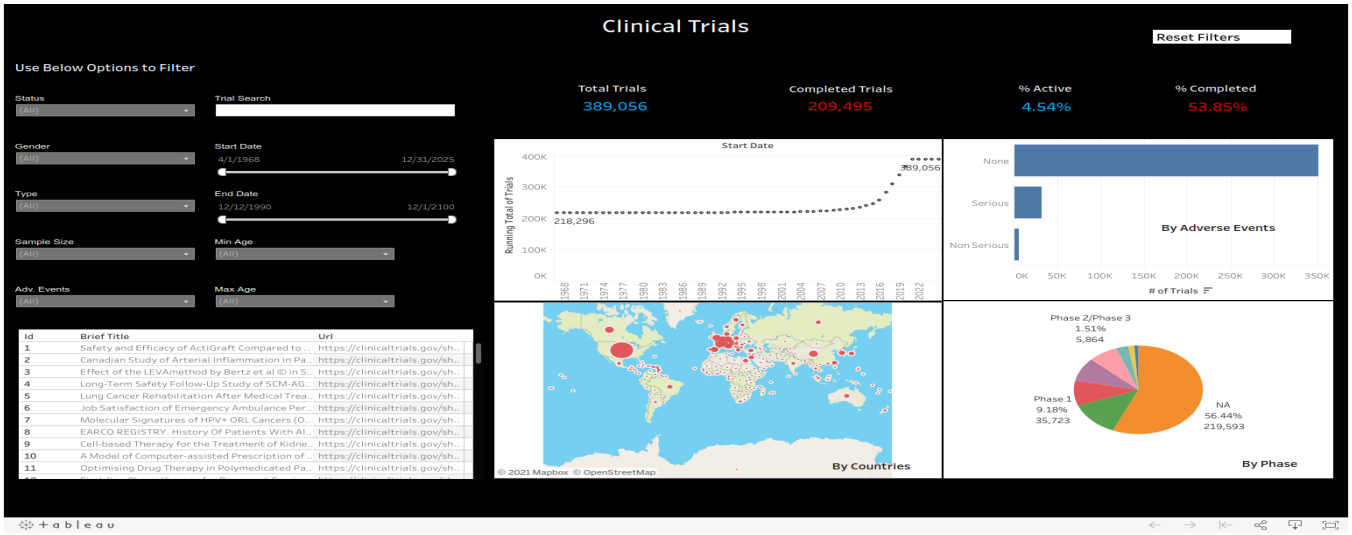

Figure 1: Tableau Dashboard

Our approach is to solve that problem explicitly by analyzing trial data from clinicaltrials.gov (approx. 380K trials). Our tool will identify correlations among variables and present those in a visual display, increasing the understanding of current clinical trials. Through our data driven approach, we are optimistic this visualization tool will allow for more robust and rapid understanding of published clinical trial data [12] [9] [19] and help researchers more rapidly assess the eligibility criteria and structured trial information just as Du et al. suggest.

By utilizing statistical methods to identify outliers, meta-analysis of studies matching a users search criteria (e.g., all trials about Parkinson's disease that have adverse events) may be possible using a methodology similar as proposed in [24].

Tableau was used to generate the interactive dashboards. Tableau is widely accepted in the medical community, and we are more likely to gain acceptance from users [10] [22] [23]. We have applied a hierarchical analysis by way of constructing an ontology on the clinical trial database similar as shown in [16] and [15].

2.1.1 Who cares and what difference will it make? All involved in clinical trial research (e.g. commercial pharma) should care about this tool. Inspecting clinical trial data for adverse drug reactions is not possible without extracting data from XML files. However, the ultimate beneficiary will be patients who need treatment and minimize patient exposure to adverse drug events in future clinical trials.

The tool should also aid researchers in understanding more about CT populations, giving greater insight into control (or placebo) groups than otherwise possible. Understanding not only test, but control populations is fundamental in experimental design [14] [7] [13]. This should also make clearer the current state of patient recruitment [11] [4].

2.2 Data Ingestion & Data Modeling

2.2.1 Data Parsing & Creating the Object Model. A substantial effort was required to transform the raw data from XML format into CSV files suitable for ingestion into our database. Our primary goal in this project is to be able to identify with a reasonable level of confidence the demographics of a particular clinical trial (e.g. age ranges, gender and ethnicity) as well as extract whether or not any of those patients experienced any adverse events as a result of the trial (that is – the patient had a safety issue while participating in the trial).

Furthermore, we identified for each trial, whether or not there were healthy patients involved. Investigators planning for future clinical trials are interested in understanding what the demographics are of patients willing to participate in a clinical trial and this information will help our users discover the pool of potential future participants.

The XML parser was written in Java and utilized the JAXB (Jakarta XML Binding) libraries [6] to reverse engineer the clinicaltrials.gov XSD (XML Schema Definition) into generated java objects in the de-marshalling process. We extracted all possible strings used to represent ethnic categories, and took an approach to try and unify these into high level categories (e.g. Caucasian, Hispanic, black, etc).

While the clinicaltrials.gov system has attempted to map the adverse event terms into MeSH, the XML documents themselves state that this algorithm is not perfect; and there is a recommendation to proceed with caution in using these maps. Since we are interested primarily in defining safety events, and most of the safety community processes adverse event data using MedDRA (Medical Dictionary for Regulatory Affairs), we chose MedDRA as our standard terminology to map adverse event and condition terms rather than MeSH. A two-phase approach is applied to map conditions and adverse events to the MedDRA dictionary. First, maps are found using exact term matching (e.g. 'nausea' maps directly to the MedDRA code 10028813 'Nausea').

However, exact term matching does not yield a complete map of reported AE and conditions in CT data, thereby requiring us to extend the mapping to a second stage using fuzzy term matching similar to the process shown in [17]. The fuzzy term matching utilizes a bigram analysis to probabilistically align the terms found in the CT adverse event and conditions (which may contain misspellings or non-standard terminology) to MedDRA terms where possible. Below is an example of the results achieved by searching for reported adverse events that contain the term 'nausea' – the query returns all of the reported AE terms in the CT data with a link to the MedDRA code/term pair that it was mapped to:

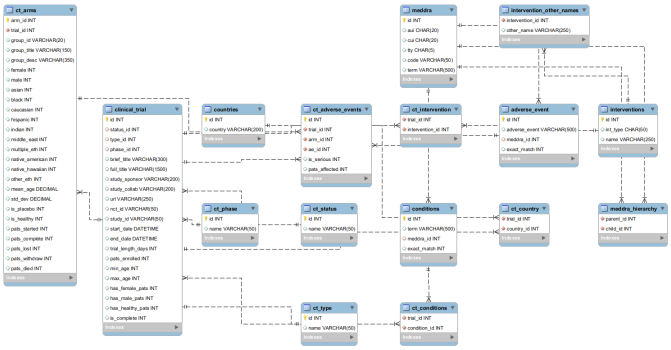

Figure 2: Quaranteam: Clinical Trial Data Model

```
SELECT  reported_ae_term, meddra_code, meddra_term
  FROM    ctdb.adverse_event
  WHERE   reported_ae_term like 'nausea';
```

reported_ae_term	meddra_code	meddra_term
Nausea?	10028813	Nausea
Nause	10028813	Nausea
Nauseas	10028813	Nausea
Nausea*	10028813	Nausea
Nausea (NAUSEA)	10028813	Nausea
Naseau	10028813	Nausea
Nauseau	10028813	Nausea
Nasusea	10028813	Nausea
Nauseau AE	10028813	Nausea
Nausea G1	10028813	Nausea
Nausea**	10028813	Nausea
LLT Nausea	10028813	Nausea
Nausea+	10028813	Nausea
Nausea G2	10028813	Nausea
Nausea G3	10028813	Nausea

There are 15 different forms of reported AE terms that contain the term 'nausea' and we see from the SQL output above that we were able to align all of these variations into a single MedDRA term. This mapping process will allow us to search across the clinical trials data in a more uniform and systematic way rather than trying to accommodate each of the reported AE terms.

2.2.2 Data Model. Our data model (Figure 2) grew organically as we explored the data from the raw XML format and started to think about how we would like to organize and structure the data to help solve our project goals.

The data model starts of course with the **clinical_trial** table which tells us some identifying information around the trial (title, description and registration identifiers – e.g. NCT ID, clinicaltrials.gov URL, etc). From this table, we have linked several supporting tables including **ct_country** (where the trials are taking place), **ct_arms**

(the arm represents who is participating in a trial), **ct_condition** (what is the study trying to treat?), **ct_intervention** (what treatment is the study applying?).

For overall trial eligibility, we have included some information at the **clinical_trial** table level to help quickly identify and speed up querying of the data. This includes details such as minimum and maximum age, whether the trial includes males or females and if the trial includes healthy patients or not. The real details around the population under study however are stored in the **ct_arms** table.

It is at the **ct_arms** table level that the mean and standard deviation of the age is reported, as the counts of male/female and ethnicity of each arm when it is provided.

Further, each trial arm is linked to tables that describe adverse events patients may have experienced, along with count data which is in the **ct_adverse_events** table. Arms can have more than one adverse event linked, and these include both serious and non-serious outcomes (defined in the adverse event table descriptor itself). Due to the nature of CT data, patients are blinded, and we do not have exact details on who was affected (i.e there is no personal identifiable information about the patient stored in our database).

2.3 Data Ingestion

We downloaded the entire set of XML files from clinicaltrials.gov as of 2021-Sep-06. The archive is downloaded as a single ZIP file and when expanded, generated 389,766 XML files where each XML file represents a single clinical trial available at the time of download. The archive was approximately 11gb in size.

After converting the XML files into our data model, we excluded trials that failed to meet some basic criteria consistent with other researchers. This resulted in the total number of trials being reduced to 389,056 (710 trials - 0.18%).

The reasons for excluding a trial include the following:

- The study had a status of "Withheld" meaning that the study sponsor or investigators have chosen not to share results of the trial.
- The second reason to exclude a trial is that no eligibility criteria were defined for a patient cohort. Without any eligibility criteria, that would mean that the trial has not specified a population to study and therefore, is not of interest to our use cases.
- Thirdly, a study must have some conditions to study. If the clinical trial did not specify what they were studying, we excluded it.

In all, these three exclusion criteria resulted in 710 trials not being entered into our database (less than 0.2% of all trial data available).

Our fully processed data files can be downloaded from the following URL: https://tinyurl.com/quaranteam-data

Once the XML data was parsed and converted into our data model, the next step was to store the data in a way that all team members could access the data for further project activities. A python script was written using the sqlite3 library to generate the database, tables, and requirements shown in Figure 2.

Once the DDL was in place, we could then load the CSV files generated by our XML parsing module into their respective tables using sqlite3, pandas and the csv library. Logging of the entire data ingestion process is accomplished by using python's logging library. This method of creating the database generates a local copy of the SQLite database to work with while developing the data visualization and dashboards.

2.4 Data Visualization

After creating our SQLite database representing the data available from clinicaltrials.gov, our next step was to perform some basic data discovery and identify the list of data attributes required for our dashboard development as shown in the rendering of our final Tableau Dashboard in Figure 1. The table below enumerates the fields required for the dashboard, along with the source table and any transformations needed before incorporating those elements into the data visualization tool.

After completing the data discovery step, a SQL view was created **vw_clinical_trial**. The view exists in our SQLite database instance and joins the tables listed in Figure 3. This view allows us to produce all of the data needed for dashboard development.

In this process, we started with main table **clinical_trial** and left joined other tables, applying data transformations as necessary to enrich the data for visualization.

BusinessField	TableName	FieldName(s)	Transformation
Total Trials	clinical_trial	Id	# of unique Ids
Completed Trials	clinical_trial	Id, status	# of unique Ids where status = 'Completed'
% On-going	clinical_trial	Id, status	% of unique Ids where status = 'Active, not recruting'
% Completed	clinical_trial	Id, status	% of unique Ids where status = 'Completed'
Min Age	clinical_trial	min_age	categorized this into different age buckets
Max Age	clinical_trial	max_age	categorized this into different age buckets
Sample Size	clinical_trial	pats_enrolled	categorized this into different count buckets
Trial Title	clinical_trial	brief_title	
Trial Url	clinical_trial	url	
Gender Male	clinical_trial	has_male_pats	
Gender Female	clinical_trial	has_female_pats	
Trial Start Date	clinical_trial	start_date	
Trial End Date	clinical_trial	end_date	
Study Status	ct_status	name	
Phase	ct_phase	name	
Country	countries	country	
Type	ct_type	name	
Adv Events	ct_adverse_events	is_serious	for a trial if sum(is_serious)>0 then 'Serious' else 'Non Serious'

Figure 3: Data Elements and Transformations for Dashboard

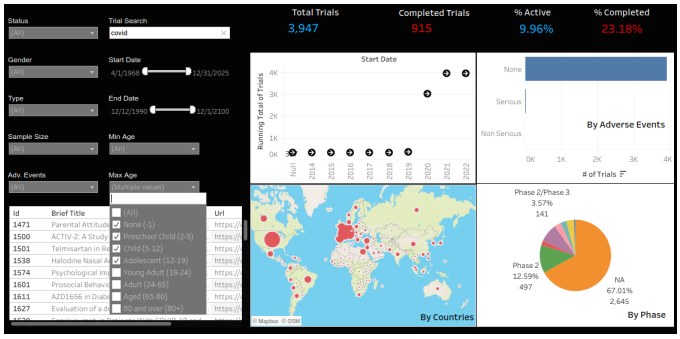

Figure 4: COVID-19 Studies of Children

Once this SQL view was tested for completeness and accuracy, we setup a data connection in data visualization tool named Tableau and linked to this SQL view **vw_clinical_trial** and imported data into Tableau.

For each of the chart components in the dashboard, a separate worksheet was created in Tableau. Later, these were combined into a single worksheet in order to generate the final dashboard seen in Figure [x]. To enable public access to anyone, this dashboard has been published in a public Tableau server which can be accessed by the following URL: https://tinyurl.com/quaranteam-dashboard

3 EXPERIMENTS

In a recent publication, Zippel et al attempted to scrape the clinicaltrials.gov website to ascertain the prevalence of Machine Learning methods as used in clinical trials [25]. This is a perfect example of the type of activity our tool could more easily and quickly facilitate. We may use the search criteria presented in this paper to see if it is possible for us to duplicate their results and findings, and could also influence how our dashboard and visualizations may evolve as the project evolves.

In another publication, Du et al stated that with the rapid increase in COVID-19 clinical trials, it is imperative to be able to quickly search and retrieve data related to these trials in order to better understand the work being carried out across the globe in order to reduce duplicative work and also to understand when new results are reported more easily [5].

Another experiment was to see how the tool could help us better understand clinical trials related to COVID-19, and more specifically to children. In Figure 4, you can see that in just a few clicks, we can search for all clinical trials mentioning COVID-19 (n=5,727), and then adjust our filter by max age to include only those under age 19 to see that 3,947 trials (69%) had a maximum age which excluded adults in the trial. This type of analysis would be impossible directly through the clinicaltrials.gov website, and would take significant programming effort to extract from the raw XML files.

4 EVALUATION

For the first publication, Zippel et al conducted an in depth review in order to ascertain the prevalence of Machine Learning methods as used in clinical trials. Through our dashboard, this same study can be completed with much less effort. Due to the construction of our data model, we can easily assess many of the same elements that were covered in this report simply by entering the search term “machine learning” into the dashboard, and generating a report.

In Figure 5, we see the trend is increasing in recent years around the mention of “Machine Learning” in clinical trials. This is consistent with the increase in availability of ML methods and high end computing required to perform these types of trials, as well as the increased interest in machine learning in the last few years. With minimal effort, we could update our search to include the additional search terms used by Zipple to come to the same conclusions reached in their paper, but with much less effort.

5 CONCLUSION

This project demonstrates that by creating a dashboard view for the rapid assessment of clinicaltrials.gov data,

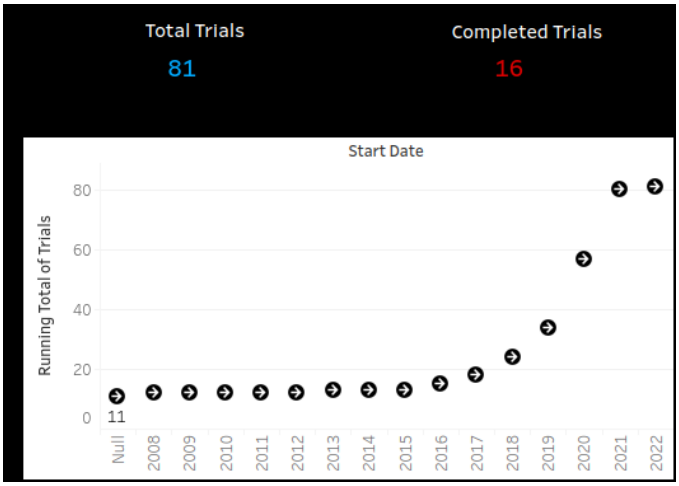

Figure 5: Clinical Trials involving Machine Learning

we are able to enable any user to quickly assess the clinical trial landscape without having to have programming expertise. Furthermore, the dashboard provides a much richer experience for the user to understand the current state of all clinical trials in a way that is not currently possible with any other tool, including the clinicaltrials.gov website itself. In an instant, the user can quickly search, evaluate and understand clinical trials based on simple keyword searches, see the distribution of those clinical trials across the globe through our geographic mapping module, and also see a bar chart which alerts them to the frequency of which adverse safety events are happening in those clinical trials.

This tool should prove a valuable companion to anyone involved in researching and understanding how clinical trials are being conducted and reported through the clinicaltrials.gov website.

REFERENCES

[1] Dimitris Bertsimas, Allison O'Hair, Stephen Relyea, and John Silberholz. 2013. An analytics approach to designing clinical trials for cancer. *Submitted for publication* (2013).

[2] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.

[3] Nathan Bryant and Jeremy Wildfire. 2016. Webcharts–A Web-based Charting Library for Custom Interactive Data Visualization. *Journal of Open Research Software* 4, 1 (2016).

[4] Richard E Deichmann, Marie Krousel-Wood, and Joseph Breault. 2016. Bioethics in practice: Considerations for stopping a clinical trial early. *Ochsner Journal* 16, 3 (2016), 197–198.

[5] Jingcheng Du, Qing Wang, Jingqi Wang, Prerana Ramesh, Yang Xiang, Xiaoqian Jiang, and Cui Tao. 2021. COVID-19 Trial Graph: A Linked Graph for COVID-19 Clinical Trials. *Journal of the American Medical Informatics Association* (2021).

[6] Joseph Fialli and Sekhar Vajjhala. 2003. The Java architecture for XML binding (JAXB). *JSR Specification, January* (2003).

[7] Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. 2015. Basic study design. In *Fundamentals of Clinical Trials*. Springer, 89–121.

[8] David Gotz and David Borland. 2016. Data-driven healthcare: challenges and opportunities for interactive visualization. *IEEE computer graphics and applications* 36, 3 (2016), 90–96.

[9] The Lancet Haematology. 2017. Clinical trial reporting: an evolving process.

[10] Timothy C Huber, Arun Krishnaraj, Dayna Monaghan, and Cree M Gaskin. 2018. Developing an interactive data visualization tool to assess the impact of decision support on clinical operations. *Journal of digital imaging* 31, 5 (2018), 640–645.

[11] Rashmi Ashish Kadam, Sanghratna Umakant Borde, Sapna Amol Madas, Sundeep Santosh Salvi, and Sneha Saurabh Limaye. 2016. Challenges in recruitment and retention of clinical trial subjects. *Perspectives in clinical research* 7, 3 (2016), 137.

[12] Noreen Kamal, Samuel Wiebe, JD Engbers, and MD Hill. 2014. Big data and visual analytics in health and medicine: from pipe dream to reality. *Journal of Health & Medical Informatics* 5, 5 (2014).

[13] S Sam Lim, Alan J Kivitz, Doug McKinnell, M Edward Pierson, and Faye S O'Brien. 2017. Simulating clinical trial visits yields patient insights into study design and recruitment. *Patient preference and adherence* 11 (2017), 1295.

[14] Charles S Mayo, Martha M Matuszak, Matthew J Schipper, Shruti Jolly, James A Hayman, and Randall K Ten Haken. 2017. Big data in designing clinical trials: opportunities and challenges. *Frontiers in oncology* 7 (2017), 187.

[15] Gary H Merrill. 2008. The meddra paradox. In *AMIA annual symposium proceedings*, Vol. 2008. American Medical Informatics Association, 470.

[16] Gary H Merrill, Patrick B Ryan, and Jeffery L Painter. 2008. Construction and annotation of a UMLS/SNOMED-based drug ontology for observational pharmacovigilance. *Methods* (2008).

[17] Jeffery L Painter. 2010. Toward automating an inference model on unstructured terminologies: Oxmis case study. In *Advances in Computational Biology*. Springer, 645–651.

[18] Zachary Seagrave and Sonya Bamba. 2017. Adverse drug reactions. *Disease-a-Month* 2, 63 (2017), 49–53.

[19] Philipp Storz-Pfennig. 2017. Potentially unnecessary and wasteful clinical trial research detected in cumulative meta-epidemiological and trial sequential analysis. *Journal of clinical epidemiology* 82 (2017), 61–70.

[20] Eve Tang, Philippe Ravaud, Carolina Riveros, Elodie Perrodeau, and Agnes Dechartres. 2015. Comparison of serious adverse events posted at ClinicalTrials. gov and published in corresponding journal articles. *BMC medicine* 13, 1 (2015), 1–8.

[21] Asba Tasneem, Laura Aberle, Hari Ananth, Swati Chakraborty, Karen Chiswell, Brian J McCourt, and Ricardo Pietrobon. 2012. The database for aggregate analysis of ClinicalTrials. gov (AACT) and subsequent regrouping by clinical specialty. *PloS one* 7, 3 (2012), e33677.

[22] Maurine Tong, William Hsu, and Ricky K Taira. 2016. Evaluating a novel summary visualization for clinical trial reports: a usability study. In *AMIA Annual Symposium Proceedings*, Vol. 2016. American Medical Informatics Association, 2007.

[23] Eric Yang, Christopher O'Donovan, JodiLyn Phillips, Leone Atkinson, Krishnendu Ghosh, and Dimitris K Agrafiotis. 2018. Quantifying and visualizing site performance in clinical trials. *Contemporary clinical trials communications* 9 (2018), 108–114.

[24] Jing Zhang, Haoda Fu, and Bradley P Carlin. 2015. Detecting outlying trials in network meta-analysis. *Statistics in medicine* 34, 19 (2015), 2695–2707.

[25] Claus Zippel and Sabine Bohnet-Joschko. 2021. Rise of Clinical Studies in the Field of Machine Learning: A Review of Data Registered in ClinicalTrials. gov. *International Journal of Environmental Research and Public Health* 18, 10 (2021), 5072.