MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

© 2022 International Medical Informatics Association (IMIA) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220085

Interactive Similarity-Based Search of Clinical Trials

Yingcheng Sun[a], Jiaqi Tang[b], Alex Butler[ac], Cong Liu[a], Yilu Fang[a], Chunhua Weng[a]

[a] Department of Biomedical Informatics, [b] Data Science Institute, [c] Department of Medicine Columbia University, New York, NY, USA

Abstract

The rapid growth of clinical trials launched in recent years poses significant challenges for accurate and efficient trial search. Keyword-based clinical trial search engines require users to construct effective queries, which can be a difficult task given complex information needs. In this study, we present an interactive clinical trial search interface that retrieves trials similar to a target clinical trial. It enables user configuration of 13 clinical trial features and 4 metrics (Jaccard similarity, semantic-based similarity, temporal overlap and geographical distance) to measure pairwise trial similarities. Among 1,007 coronavirus disease 2019 (COVID-19) trials conducted in the United States, 91.9% were found to have similar trials with the similarity threshold being 0.85 and 43.8% were highly similar with the threshold 0.95. A simulation study using 3 groups of similar trials curated by COVID-19 clinical trial reviews demonstrates the precision and recall of the search interface.

Keywords:

Clinical Trial, Similarity Search, Information Retrieval

Introduction

Clinical trials, the well-regarded gold standard for generating medical evidence [1], have been growing exponentially recently [2]. For example, in ClinicalTrials.gov, one of the largest clinical trial registries in the world, there are more than 370,000 registered clinical studies as of April, 2021, with an average annual growth rate of 13.7%. The vast volume of trials can be overwhelming to clinical trial seekers [3]. Existing clinical trial retrieval tools such as Janssen Global Trial Finder [4], ResearchMatch [5], or SearhClinicalTrials.org, require users to manually construct effective queries to find relevant trials, which is hard given the complexity of clinical trials [6].

A potential approach to this problem is content-based clinical trial retrieval that allows users to provide a target clinical trial and to identify trials similar to this target [7]. With no need for manually constructing complex queries, this approach can enable end users without technical skills to perform highly customized searches. For example, stakeholders might search for clinical studies related to a new trial [8]. Investigators could search for collaboration opportunities among clinical trials with similar study populations [9]. After the COVID-19 pandemic started, plentiful clinical trials emerged in a short time, many of which studied similar treatments or medications. Identifying the competing trials and recommending collaboration opportunities might help investigators reduce research costs and boost patient enrollment [10]. For clinical trial volunteers, other similar trials can be detected and provided as recommendations if a trial of interest is closed, already full of eligible participants or the recruitment site is too far away [11].

Previous studies measured similarity using specific features of clinical trials, such as condition name [12], outcome [13] or eligibility criteria [7], but no prior work measured clinical trial similarity using all and any possible trial features such as geographic location, condition, eligibility criteria, outcome, etc. In this study, we extracted 13 features from full-text clinical trial summaries, including structured elements with categorical values such as "study type" or "intervention type" and unstructured elements in free text such as "outcome measure" or "eligibility criteria". Four different metrics including Jaccard similarity, semantic-based similarity, temporal overlap and geographical proximity, were used to determine the pairwise clinical trial similarity using various clinical trial features. We assessed the similarity of 1,007 COVID-19 trials in the US exported from ClinicalTrials.gov and found over 90% of trials had trials similar to them according to some features.

We built an interactive, similarity-based clinical trial search interface that enables flexible selection of various features for similar trial retrieval. It takes a target clinical trial as input and returns a ranked list of similar trials. The recruiting sites of the returned trials are marked on a location map and the number of sites is visualized by a heat map. The system is accessible online (http://apex.dbmi.columbia.edu/trialmatcher/), as well as its source code (https://github.com/WengLabInformaticsResearch/COVID19-TrialMatcher). We collected three groups of similar trials measured by different features from COVID-19 trial reviews. A simulation study using the three groups of similar trials demonstrates the system's precision and recall.

Methods

Features and Metrics

In ClinicalTrials.gov, a clinical trial summary includes the descriptive, recruitment, tracking and administrative information and other data elements [14]. We worked with a medical domain expert (AB) and extracted 13 common data elements to use as similarity features, which are Study Type, Masking, Phase, Primary Purpose, Intervention/Observation Model, Allocation, Intervention Type represented by categorical variables; Condition Name, Intervention Name, Outcome Measure and Eligibility Criteria represented by free text; Location represented by geographical addresses and Study Period represented by a temporal interval. Various similarity metrics were applied to different types of features. All features and metrics are described as follows. For each categorical feature, its value distribution in 1,007 COVID-19 trials is also provided in percentage after the category name.

Study Type: the nature of a clinical study, whose allowed values include Interventional Studies (70.71%), Observational Studies (Including Patient Registries) (27.71%), and Expanded Access (1.59%).

Masking: a clinical trial design strategy, in which one or more parties involved in the trial for interventional studies, including: Open Label (42.98%), Quadruple (19.52%), Double (18.68%), Triple (12.22%) and Single (6.60%) Blind Masking.

Phase: stage of a clinical trial studying a drug or biological product for interventional studies, including: Early Phase 1 (2.11%), Phase 1 (9.41%), Phase 1/Phase 2 (6.46%), Phase 2 (34.83%), Phase 2/Phase 3 (5.48%), Phase 3 (13.34%) and Phase 4 (4.63%). Not Applicable (23.74%) is used to describe trials of devices or behavioral interventions.

Primary Purpose: the main reason for the clinical trial, including: Treatment (72.33%), Prevention (11.80%), Supportive Care (5.20%), Diagnostic (2.53%), Device Feasibility (0.56%), Screening (0.70%), Health Services Research (2.39%), Basic Science (1.12%), and Other (3.37%).

Allocation: a method used to assign participants to an arm of a clinical study, including Randomized (92.28%) and Nonrandomized (7.72%).

Intervention Type: general types of the interventional study, including: Drug (48.86%), Biological (10.57%), Behavioral (7.17%), Device (5.44%), Diagnostic Test (3.76%), Dietary Supplement (2.69%), Procedure (1.49%) , Combination Product (0.72%), Radiation (0.66%), Genetic (0.12%) and others (18.52%). A clinical trial may contain multiple intervention types.

Intervention/Observation Model: the general design of the strategy for assigning interventions to participants in a clinical study, including Parallel (71.49%), Single Group (19.52%), Sequential (5.20%), Cross-Over (1.97%), and Factorial Assignment (1.83%) for intervention models, and Cohort (63.80%), Case-Only (13.26%), Other (8.96%), Case-Control (7.53%), Ecologic Or Community (6.09%) and Family-Based (0.36%) for observation models.

For the above 7 features represented by categorical variables, Jaccard Index is used as the similarity metric:

$$f(x_1, x_2) = \frac{|S(x!) \cap S(x")|}{|S(x!) \cup S(x")|}$$

where $f(x!, x")$ is the similarity between clinical trial $x!$ and$x"$, and (x) is the set of categorical values in trial x.

Condition Name: the disease, disorder, syndrome, illness, or injury that is being studied.

Intervention Name: a process or action taken to treat or cure a condition. "Placebo" is not included when comparing the similarity of two trials since it is not a real treatment.

Outcome Measure: variables monitored during a clinical trial to assess how they are affected by the treatment taken or by other parameters.

For these 3 features (Condition Name, Intervention Name, and Outcome Measure), all text are first converted to vectors by word embedding methods, and then measured by cosine similarity algorithm to obtain the semantic-based similarity.

Study Period: the whole study period from the actual start date to the final completion date. If the study is ongoing, the completion date is estimated. The temporal overlap between two study periods is used as the similarity metric, where the similarity value is 1 if there is overlap, 0 otherwise.

Eligibility Criteria: the qualification criteria for study participants. Eligibility Criteria usually include multiple complex inclusion or exclusion rules [15]. To avoid overfitting, only frequently used criteria are extracted for measuring eligibility criteria similarities. In this study, 7 most frequently used criteria across 1,007 COVID-19 trials about Age, Gender, High-Risk Status, COVID-19 Status, Current Hospitalization Status, Pregnancy Status, and Healthy Status were extracted to measure Eligibility Criteria similarity. For Age similarity, temporal overlap is used as the metric. For other criteria, Jaccard Index is used as the similarity metric. The overall similarity is the weighted mean of all frequent criteria similarity values.

Location: the geographical locations where clinical trials are conducted. A clinical study might have multiple recruiting sites. The closest two sites of any pair of trials are used to measure the physical proximity as follows:

$$di(x_1, x_2) = \min\{p4l_{\#}, l_{\$}67\ l_{\#}\ in\ L_{\%}, l_{\$}\ in\ L_{\%}\}$$

where $dis(x_1, x_2)$ represents the shortest distance between the locations of two trials x_1 and x_2 , and $p4l_{\#}, l_{\$}6$ measures the geographical distance of sites $l_{\#}$ and $l_{\$}$. $L_{\%}$ is the set of recruiting sites in trial x. Location similarity $f_{\&}$is computed by a piecewise function:

$$f_{\&}(x_1, x_2) = \begin{cases} 1, & if\ d \le D_1 \\ 1 - \left(\frac{d}{D_2}\right)^n, & if\ D_1 < d < D_2 \\ 0, & if\ d \ge D_2 \end{cases}$$

where d equals to $dis(x_1, x_2)$, D_1, D_2 and n are parameters. When the shortest distance d is less than D_1, such as 10 miles, the distance can be ignored and the similarity value is 1. Contrarily, if the shortest distance is more than D_2, such as 3,000 miles, participants usually would not like to travel over such a long distance to the other site, so the similarity is 0. For other conditions, the similarity metric is a decreasing function that initially decreases slowly and then fast after a point.

The clinical trial similarity metric is the weighted average over all selected features:

$$sim(x_1, x_2) = \frac{1}{\sum w_i} \sum w_i f_i(x_1, x_2)$$

where w is the weight for similarity features. For the similarity metric whose value is a continuous variable, the similarity threshold determines the desired lower limit for the similarity of two clinical trials. To be convenient, we use **SCTs** as the abbreviation for "**similar clinical trials**". A trial with SCTs means we can find other trials similar to it. Figure 1 shows the ratio of trials with SCTs against different thresholds.

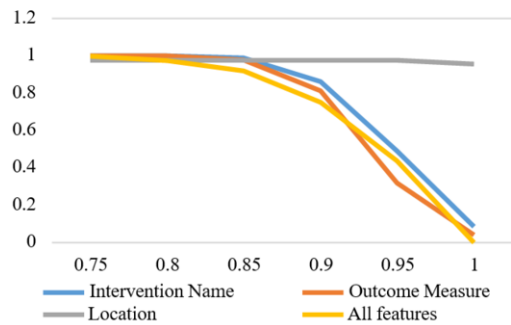

Figure 1 – Ratio of trials with SCTs against different threshold values

In Figure 1, the ratio of trials with SCTs measured by Intervention Name, Outcome measure, Location and the combination of all features with different similarity thresholds were presented. For Location feature, we set D_1 as 10 miles, D_2 as 3,000 miles and n as 1. We found 95.6% trials have SCTs within 10 miles, so the ratio is up to 0.95 even the threshold is1. For the other two features and the feature combination, the ratio decreases obviously faster when the threshold is 0.85. A balance needs to be preserved between the number of trials with SCTs and the degree of similarity, so we set the threshold as 0.85 to assess the pairwise similarities of 1,007 COVID-19 trials. Table 1 lists the number and percentage of trials that could find SCTs based on the given feature.

Table 1– Clinical trial similarity measured by different features.

Feature *(t: similarity threshold)*	**Trials with SCTs**	
	Count	**Percentage**
Masking	1,007	100.00%
Phase	1,007	100.00%
Study Type	1,007	100.00%
Primary Purpose	1,007	100.00%
Allocation	1,007	100.00%
Intervention Type	1,007	100.00%
Condition Name	1,007	100.00%
Intervention/Observation Model	1,006	99.00%
Location *(t=0.85)*	963	95.60%
Study Period	949	94.20%
Eligibility Criteria	937	93.00%
Intervention Name *(t=0.85)*	996	98.90%
Outcome Measure *(t=0.85)*	982	97.60%
All features (t=0.85)	925	91.90%
All features (t=0.95)	441	43.80%

For the categorical features, all trials can find SCTs measured by Masking, Phase, Study Type, Primary Purpose, Allocation or Intervention Type.

There is a trial with the Observation Model type "Family-Based" without any similar trial, making the percentage of trials with SCTs to 99%. All the trials have the similar "Condition Name" since they all study COVID-19.The percentage of trials with SCTs measured by the other five features or the combination of all features are all above 90%. If we increase the similarity threshold up to 0.95 and choose all features, there are still 43.8% trials that have SCTs, generating 990 pairs of similar trials. It demonstrates that the features we extracted are effective in searching similar clinical trials.

Clinical Trial Search Interface

We developed an interactive clinical trial search interface to assist in finding similar COVID-19 trials given an example target clinical trial. It supports user-specified similarity measure based on different clinical trial similarity features with customized weights, and provides visualization for the searched results. Figure 2 shows the architecture.

Clinical trials are exported from ClinicalTrials.gov and then all features are extracted from the trials. For categorical features, original category names were mapped to numbers to increase the computation efficiency. For features represented by free text, a deep learning model "Bio_Clinical BERT" [16] pretrained on clinical notes in a large clinical database is used to convert the free text into embedded vectors for semantic-based similarity computation. For the Location feature, Google Map API is used to calculate the geographical distance between two sites given their zip code information.

Users are able to adjust the weights applied to each extracted feature via sliders on the search interface. Zero weight means the feature is eliminated while any larger number (up to 10) is applied to the relevant metric (Jaccard Index, semantic-based similarity, temporal overlap or geographical distance) to compute the trial similarity based on this feature set.

Given an submitted trial ID, a ranked list of similar trials based on a default or user-defined similarity features will be returned. The returned results include a detailed table with all the feature values, and these features are visualized by an interactive location map that marks all the recruiting sites of found trials and a heat map that shows the density of all the recruiting sites. Each trial has been assigned a hyper link to ClinicalTrials.gov. Users interested in learning more about the study will be able to access a link to ClinicalTrials.gov.

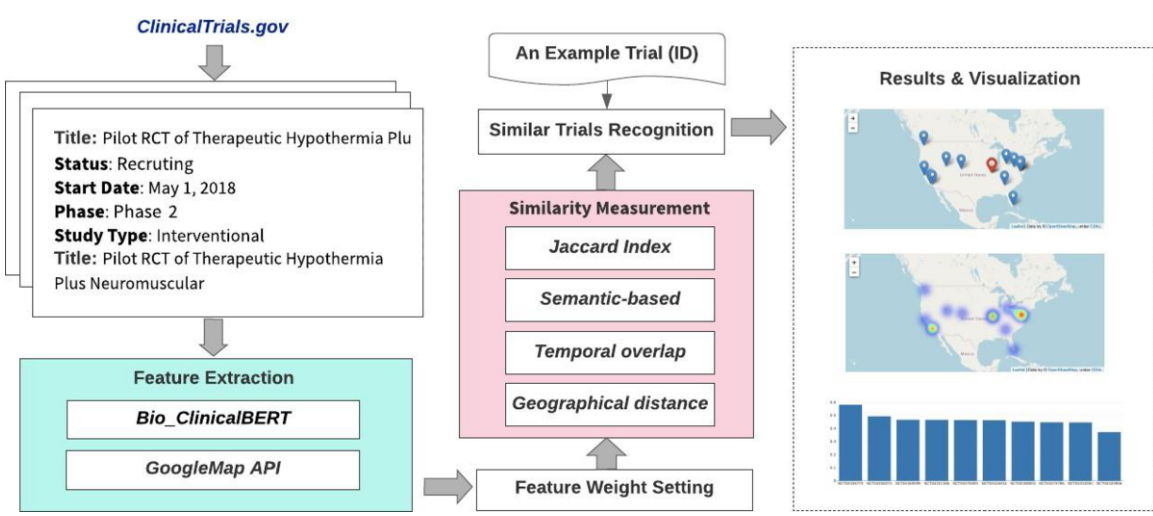

Figure 2– Architecture of An Interactive User Interface for Similarity-Based Search of Clinical Trials

Table 2–Effectiveness evaluation of finding similar clinical trial with different feature combinations.

#	Included Similarity Features	# of target SCT pairs	# of found SCT pairs	Precision	Recall	F1 Score
1	Primary Purpose, Phase, Allocation, Masking	8	8	1.0	1.0	1.0
2	Primary Purpose, Eligibility Criteria, Intervention Name	295	220	0.73	0.75	0.74
3	Intervention Type, Intervention Name	690	596	0.76	0.86	0.81
Average				0.83	0.87	0.85

Results

We evaluated the effectiveness of this similarity-based trial search user interface by assessing its precision and recall in finding similar trials with various similarity features. We chose three groups of similar trials clustered by three different combinations of similarity features and run simulations on our search interface. Each group of trials is curated by a systematic review report of COVID-19 trials [17] [18] [19], respectively. To make them comparable with our dataset, only clinical trials in the US by the same acquisition date were included for all groups of clinical trials. Table 2 lists the results.

In case 1, all selected features are categorical, and the 8 pairs of similar clinical trials were all successfully identified by querying one trial to find another making the F1 score up to 1.0. In case 2 and 3, around 25% pairs of similar trials were not correctly identified, resulting in lower F1 scores. Among the five features used in case 2 and 3, values of the features "Primary Purpose", "Intervention Type" and "Eligibility Criteria" all need to be exactly matched to be considered "similar", while the representations of free text in "Intervention Name" could be different by different text vectorization methods or comparison metrics. To be specific, the similarity metric used by the review reports for case 2 and 3 to compare "Intervention Name" is: if there is any common intervention name except "Placebo" between two trials, they are similar, otherwise not, which is different from our semantic-based similarity metric.

Next, we evaluated the "edge cases" of the search interface with extreme conditions. We selected all similarity features and set the similarity threshold to the highest value 0.99 (no results for 1.0), and searched SCTs for each trial to find the most similar pairs of trials. Table 3 listed all the identified pairs of trials and their sponsors.

Table 3–The most similar pairs of trials

#	Clinical Trial	Similar Trials	Sponsor
1	NCT04589117	NCT04589104	Duke University
2	NCT04662060	NCT04662073, NCT04662086	Stanford University
3	NCT04393311	NCT04570501	Stanford University
4	NCT04424446	NCT04334954	National Institutes of Health Clinical Center, NIAID
5	NCT04524663	NCT04346628	Stanford University
6	NCT04583969	NCT04583956	NIAID
7	NCT04292730	NCT04292899	Gilead Sciences
8	NCT04551378	NCT04650178	M.D. Anderson Cancer Center

After manually reviewing all trials in Table 3, it was confirmed that each pair of trials is very similar, and even sponsored by the same agency. For example, clinical trial "NCT04662060" has two similar trials "NCT04662073" and "NCT04662086", all sponsored by Stanford University. They belong to three different protocols but are all randomized, double masking, phase 2 and interventional COVID-19 outpatient pragmatic platform studies with closed eligibility criteria, and each of their interventions contains the drug "Acebilustat" or "Camostat" for treatment. "Sponsor" is not a similarity feature used by the search interface, but similar trials with the same sponsor were successfully identified. It shows that the system is highly precise with the maximum threshold value. If the threshold is set as 0, the system returned 506, 521 pairs of similar trials, that equal to all possible computations of the 1,007 trials. It means all similar trials were successfully retrieved by the system and the recall is 1.0 with the minimum threshold value.

We finally checked the geographical distribution of recruiting sites for all 925 trials with SCTs (threshold=0.85). Although a few trials allow remote participation by sending supplies, most trials need volunteers go to the research sites, so similar trials located in closed areas would be more attractive for volunteers to choose to participate. Since all the trails we used are from the US, we computed the number of similar trial pairs in each state and plotted the geographical distribution in Figure 3.

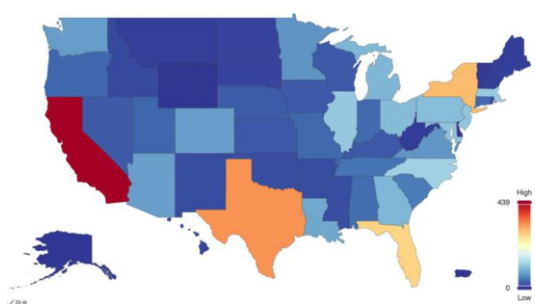

Figure 3–Distribution of pairs of similar trials in the US

There are 3,413 pairs of similar trials with recruiting sites located in the same state. In Wyoming, Puerto Rico, North Dakota and Alaska (deep blue), there are not any pair of similar trials. In California (deep red), Texas (orange), New York (yellow) and Florida (yellow) states, there are over 200 pairs of similar trials in each state, corresponded to the large COVID19 clinical studies in these states. In most states in the Midwest, like Illinois, Ohio, Iowa, Kansas, Michigan, Minnesota or Indiana (light blue), there are less than 100 pairs of similar trials. If the search condition becomes stricter with a higher similarity threshold, there might not be enough similar trials for participants to choose in these states.

Discussion

We presented 13 frequently used features extracted from clinical trials to compute the pairwise similarity for full-text clinical trial summaries. The analysis of 1,007 COVID-19 trials showed that more than 90% of trials have SCTs measured by the presented features. The number of SCTs will decrease when the threshold increases. With the highest similarity threshold up to 0.99, we can still find 9 pairs of similar trials. It demonstrates the feasibility for similarity-based search of clinical trials. A clinical trial search interface was developed providing interactive search for similar trials given an example trial and accessible visualization of geographic information.

There are a few limitations in this study. Although trained on large clinical notes, "Bio_ClinicalBert" might not accurately represent the semantics in clinical trials with different linguistic characteristics than the clinical notes. A domain specific finetuning model trained on clinical trials might improve the power of semantic similarity measure. The eligibility criteria were represented by only the most frequently used rules, but users might have different requirements in selecting eligibility criteria for similarity analyses. A flexible representation allowing users to choose any criteria rules may further improve the usability of the system. Finally, we conducted simulation studies to evaluate the system. Involvements of real users like clinical trial participants or investigators will help to obtain more valuable results for evaluation experiments.

Conclusions

We developed a method to measure clinical trial similarity and an interactive user interface to facilitate similarity-based clinical trial search by accepting a target clinical trial as input, and demonstrated its effectiveness through multiple experiments. The system currently only included COVID-19 trials conducted in the US, but it is feasible to apply the system for any trials. More clinical trial similarity features and metrics will be explored and tested in clinical trials in other disease domains in the future.

Acknowledgements

This work was supported by the National Library of Medicine grant R01LM009886-11 (Bridging the Semantic Gap Between Research Eligibility Criteria and Clinical Data) and National Center for Advancing Clinical and Translational Science grants UL1TR001873 and 3U24TR001579-05.

References

[1] Liu, H., Chi, Y., Butler, A., Sun, Y. and Weng, C., 2021. A knowledge base of clinical trial eligibility criteria. *Journal of Biomedical Informatics*, *117*, p.103771.

[2] Sun, Y., Butler, A., Stewart, L.A., Liu, H., Yuan, C., Southard, C.T., Kim, J.H. and Weng, C., 2021. Building an OMOP common data model-compliant annotated corpus for COVID-19 clinical trials. *Journal of biomedical informatics*, *118*, p.103790.

[3] Sun, Y., Butler, A., Lin, F., Liu, H., Stewart, L.A., Kim, J.H., Idnay, B.R.S., Ge, Q., Wei, X., Liu, C. and Yuan, C., 2021. *The COVID-19 Trial Finder*. Journal of the American Medical Informatics Association, 28(3), pp.616-621.

[4] Amy R. Global Trial Finder: *Why It Just Got Easier to Enroll in a Janssen Clinical Study*. 2016..

[5] Pulley JM, Jerome RN, Bernard GR, et al. *Connecting the public with clinical trial options: the research match trials today tool*. Journal of Clinical and Translational Science. 2018; 2(4), pp.253-257.

[6] Kury, F., Butler, A., Yuan, C., Fu, L.H., Sun, Y., Liu, H., Sim, I., Carini, S. and Weng, C., 2020. *Chia, a large annotated corpus of clinical trial eligibility criteria*. Scientific data, 7(1), pp.1-11.

[7] A Hao, T., Rusanov, A., Boland, M.R. and Weng, C., 2014. *Clustering clinical trials with similar eligibility criteria features*. Journal of biomedical informatics, 52, pp.112-120.

[8] Park, J., Park, S., Kim, K., Hwang, W., Yoo, S., Yi, G.S. and Lee, D., 2020. *An interactive retrieval system for clinical trial studies with context-dependent protocol elements*. PloS one, 15(9), p.e0238290.

[9] Moss, A.J., Francis, C.W. and Ryan, D., 2011. Collaborative clinical trials. The New England journal of medicine, 364(9), pp.789-791.

[10] Glasziou, P.P., Sanders, S. and Hoffmann, T., 2020. Waste in covid-19 research. The BMJ.

[11] Helmer, T.T., Lewis, A.A., McEver, M., Delacqua, F., Pastern, C.L., Kennedy, N., Edwards, T.L., Woodward, B.O. and Harris, P.A., 2021. Creating and Implementing a COVID-19 Recruitment Data Mart. *Journal of Biomedical Informatics*, p.103765.

[12] Atal, I., Zeitoun, J.D., Névéol, A., Ravaud, P., Porcher, R. and Trinquart, L., 2016. Automatic classification of registered clinical trials towards the Global Burden of Diseases taxonomy of diseases and injuries. *BMC bioinformatics*, *17*(1), pp.1-14.

[13] Koroleva, A., Kamath, S. and Paroubek, P., 2019. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *Journal of Biomedical Informatics: X*, *4*, p.100058.

[14] ClinicalTrials.gov Protocol Registration Data Element Definitions for Interventional and Observational Studies. https://prsinfo.clinicaltrials.gov/definitions.html. Accessed April 30, 2021.

[15] Sun, Y. and Loparo, K., 2019, July. Information extraction from free text in clinical trials with knowledge-based distant supervision. In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 954-955). IEEE.

[16] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T. and McDermott, M., 2019. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

[17] Fragkou, P.C., Belhadi, D., Peiffer-Smadja, N., Moschopoulos, C.D., Lescure, F.X., Janocha, H., Karofylakis, E., Yazdanpanah, Y., Mentré, F., Skevaki, C. and Laouénan, C., 2020. Review of trials currently testing treatment and prevention of COVID-19. Clinical Microbiology and Infection

[18] Karlsen, A.P.H., Wiberg, S., Laigaard, J., Pedersen, C., Rokamp, K.Z. and Mathiesen, O., 2020. A systematic review of trial registry entries for randomized clinical trials investigating COVID-19 medical prevention and treatment. PloS one, 15(8), p.e0237903.

[19] Alag, S., 2020. Analysis of COVID-19 clinical trials: A data-driven, ontology-based, and natural language processing approach. PloS one, 15(9), p.e0239694.

Address for correspondence

Chunhua Weng, chunhua@columbia.edu. Columbia University, 622 W 168 ST, PH-20 room 407, New York, NY10032, USA.