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/SHTI220151

A Sample Size Extractor for RCT Reports

Fengyang Lin[a], Hao Liu[a], Paul Moon[b], Chunhua Weng[a]

[a] Department of Biomedical Informatics, Columbia University, New York, NY, United States,
[b] College of Physicians and Surgeons: Institute of Human Nutrition, Columbia University, New York, NY, United States

Abstract

Sample size is an important indicator of the power of randomized controlled trials (RCTs). In this paper, we designed a total sample size extractor using a combination of syntactic and machine learning methods, and evaluated it on 300 Covid-19 abstracts (Covid-Set) and 100 generic RCT abstracts (General-Set). To improve the performance, we applied transfer learning from a large public corpus of annotated abstracts. We achieved an average F1 score of 0.73 on the Covid-Set testing set, and 0.60 on the General-Set using exact matches. The F1 scores for loose matches on both datasets were over 0.74. Compared with the state-of-the-art tool, our extractor reports total sample sizes directly and improved F1 scores by at least 4% without transfer learning. We demonstrated that transfer learning improved the sample size extraction accuracy and minimized human labor on annotations.

Keywords:

Sample Size, Randomized Controlled Trial, Natural Language Processing

Introduction

Randomized controlled trials (RCTs) provide reliable medical evidence and have substantial impact on decision-making for patient care. As valuable resources for evidence-based medicine, completed and published RCT studies report enrollment, intervention allocation, follow-up outcome, and data-analysis [1; 2]. Identifying relevant and desired RCT reports becomes increasingly challenging for both clinicians and clinical researchers. This motivates the development of Information Extraction (IE) systems [3] to automatically extract structured and key information from free-text medical literature, e.g., extracting PICO (Patient /Population, Intervention, Comparison, Outcomes) [4] elements from the published RCTs. However, few studies [5] accurately extract sample sizes from PubMed (https://pubmed.ncbi.nlm.nih.gov/) abstracts.

The sample size serves as an indicator of the power, reliability, and efficacy for a RCT study. A small sample size would make a trial less representative and therefore decrease the efficacy of its outcome [1; 6]. For example, the sample size of an interventional RCT study directly impacts the statistical significance of the comparison between intervention group and placebo group, helping clinicians determine whether the cohorts studied are sufficient to support the observed outcome(s). Furthermore, some meta-analysis for RCTs include sample size to investigate the effectiveness of interventions [7; 8]. Some professionals also examined the quality of the sample size calculation and statistical analysis for clinical trials in PubMed, and a latest review has been conducted on Covid-19 research [9].

More recently, the rapid evolution in deep learning and Natural Language Processing (NLP) has substantially advanced the performance of information extraction tools, especially for PICO element extraction [10; 11]. This makes it possible to locate sample size information from a recognized text snippet that contains population element(s). For instance, Trialstreamer [11], which employed a large corpus annotated abstracts including detailed population spans, was a system to extract PICO elements from RCT abstracts. They achieved the state-of-the-art on identifying an integer that is most likely to be the sample size for a study by implementing a neural network-based classification model. However, this approach is constrained by the complexity of population text snippets, which often mix sample sizes with supporting demographics such as sex, age, condition. Given a text snippet *"30 women (age: Mean = 45 years; SD = 11.5)"*, the extractor needs to recognized three numbers and distinguish which number represents age, sample size, or other demographic statistics, respectively. Furthermore, for a RCT study that contains multiple population elements and multiple statistics for sample size, additional checks and classifications are inevitable.

Previous works report total sample size indirectly by extracting all the P-elements or identifying any potential sizes, such as group size or numbers of randomized participants for RCT reports, which require additional postprocessing to identify which number is the total sample size. To overcome this limitation, our sample size extraction method is designed to directly report the total sample size from a RCT abstract.

Our contributions are 3-fold. First, we annotated 400 RCT abstracts from PubMed and extracted the sample sizes in these studies. This sharable resource can be used by future researchers interested in continual development of sample size extractors. Second, we provide a transfer-learning method for this task by pretraining with the large annotated dataset [12]. Third, our extraction method outperforms the state-of-the-art one with a relatively small training dataset.

Methods

Datasets

Firstly, we randomly retrieved 300 out of 316 RCT abstracts in Covid-19 research as of April 13, 2021 from PubMed. The 300 RCT abstracts (denoted as Covid-Set below) were split into three folds for two informatic researchers (PM, FL) to annotate sample sizes as the gold standards for training and testing. For each abstract, the following entities were extracted: total sample size, group/arm size, possible total sample size, possible group/arm size. Specifically, possible sample size or possible group/arm size is labeled when there is another correct sample size that is mentioned in the final analysis, while the "possible" one serves as a reasonable number for the sample size. For each labeled sample size, we marked one accurate total sample size for the corresponding abstract. For publications that no total sample size is reported, we labeled them with NA.

The annotation process started with both annotators labeled the first 100 abstracts and discussed to resolve disagreements. Afterwards, each researcher annotated 100 reports individually. The final labels for total sample size were reviewed by FL to ensure the consistency. We used Brat [13], a web-based tool for annotation. Figure 1 presents an annotation example in Brat. In this paper, our main experiments and evaluations were performed on this set of 300 Covid-19 RCT abstracts.

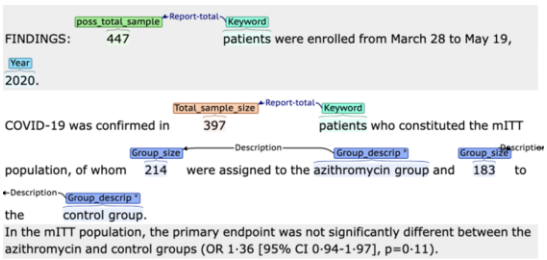

Figure 1 – Example of our annotation interface in brat

Furthermore, to illustrate the generalizability of our method, we randomly collected another set of 100 RCT abstracts from PubMed without applying any selection filter (denoted as General-Set). FL performed annotations on this dataset with the same guidelines. We evaluated our models on this dataset and validated the generalizability of the extractor.

For pretraining, we used an EBM-NLP corpus [12], which contains annotated PICO elements of 5,000 clinical randomized controlled trials. This corpus includes annotated sample size, age, sex, condition for P-elements if identified, and also achieved a high inter-annotator agreement rate. This dataset is used for pretraining to obtain the initial weights of our models.

Workflow Overview

We integrated syntactic and machine learning methods into a tool that automatically extracts the total sample size from a RCT abstract. The workflow of this tool (including model training) can be divided into three steps: 1) preprocessing; 2) feature encoding using designed rules and embeddings; and 3) training multilayer perceptron models that are initialized with randomized weights or learned weights from pretraining. An overview of the above process is shown in Figure 2.

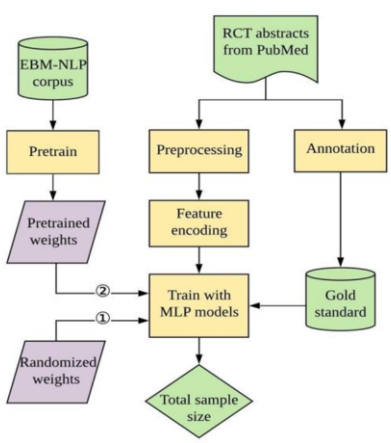

Figure 2 – Overview of our sample size extractor. We provided two options for the initialized weights for models: 1) randomized weights, 2) weights from pretraining on the EBM-NLP corpus. The performance of models initialized with these two options were compared.

Preprocessing

The first step of preprocessing is text normalization, which converting all numeric words to numbers (i.e., converting "sixty-six" to 66). We designed a function to replace all numeric words to integers. In addition to extracting integers and their indices in text, the surrounding contexts for any integer can provide additional features to our extractor. Thus, we parsed the abstracts for tokenization and part-of-speech (PoS) tagging and then collected words and PoS tags for tokens adjacent to the integers (window width t=2, 4 surrounding words in total for each number). The words were embedded into vectors using Word2vec for PubMed, a pre-trained embedding over a large amount of biomedical scientific literature from PubMed [14].

Feature encoding

Inspired by studies [6; 11] on extracting age, sex and sample size from the literature, we manually selected keywords features relevant to sample sizes. Typical keywords are "patients" and "participants" along with a flag is generated to track if there are numbers following these words. The set of words that frequently co-occur with sample sizes include "enrolled", "randomized", "analyzed", "completed", etc. Note that participants could drop out during a clinical trial, hence we put more importance on the final sample size used for analysis when a clinical trial completed, rather than the design sample size at the beginning of a trial. Therefore, we denoted two indicators for the previous keywords, one for randomization population (e.g., "randomized", "enrolled") and another for analysis population (e.g., "assessments", "analysis").

An important task is to distinguish group size from the overall sample size when both are provided in abstracts. We derived features about the numeric characteristics of total sample sizes, including an indicator of whether the number is the biggest one, indicator of whether the number is the sum of any two of the other numbers, and the percentile of the number among all numbers, etc. There is a high probability that the largest number is the total sample size, except for cases where years are mentioned. Therefore, we included an indicator to reveal whether the given number is probably a year.

Model training

The skeleton of our base model for sample size extraction is similar to the one used in Trialstreamer [11]. We adopted the multilayer perceptron (MLP) model and made modifications during the training process. With inputs concatenated from the features above, the model's architecture is composed of four fully-connected dense layers and one output layer with sigmoid activation. The model predicts the probability of a given number is the total sample size among all integers in the abstract. Indexes of numbers with the highest likelihood for each abstract can then be used to map the predicted total sample size. We set a rule that if the highest probability or confidence is lower than 0.2, no total sample size will be reported by the extractor. We used normalization and dropout layer for regularization, and all implementation is conducted with Adam optimizer in default setting in Keras (https://keras.io/). To obtain the initial weights for our model, we used the EBM-NLP dataset for pretraining.

Since our annotated *Covid-Set* is relatively small, we deployed cross-validation to train and evaluated our sample size extractor (referred as "SSE" later in tables). We also implemented the method used by Trialstreamer [11] (referred as "TS" later in tables), and performed assessment on the same dataset for comparison. For the *Covid-Set*, the 300 abstracts were divided into 5 folds, with 240 abstracts as training set and 60 as the testing set. We also divided the *General-Set* into 3 folds for

training and evaluation. In each training iteration, we used 10 epochs with batch size of 32.

Evaluation

To evaluate our model's performance, we calculated the average and best performance on test sets from cross-validation for both the Covid-Set and the General-Set. Additionally, we validated models trained with the Covid-Set on the General-Set. We defined the confusion matrix below based on Baladrón [5]:

- True Positive: Correctly extracted total sample size (i.e., both the extractor and abstract reported a total sample size, and the numbers are the same.)
- True Negative: Abstracts with no sample size reported identified correctly (i.e., both the extractor and abstract did not report any sample size.)
- False Positive: Abstracts for which some sample size was incorrectly estimated (i.e., the extractor reported a sample size, while the reported number is different from the true sample size or no size reported in the abstracts.)
- False Negative: Abstracts reporting a sample size that were incorrectly labeled as not reporting any

Based on the well-defined confusion matrix, we calculated precision, recall, accuracy, and F1 scores for both methods. Furthermore, the analysis is carried out on two levels – Exact match and Loose match. For Exact match, the extraction is considered successful only when the extracted value is the same as the one in gold standards. For Loose match, a predicted sample size within the 10% tolerance percentage of the true sample size is also considered as choices for the correct extracted size. This accommodates situations when a rough extraction is sufficient. It usually happens when a rough sample size is enough for clinicians to obtain certain understanding of the studies. Besides, when the RCT abstract reports both the sample size in design and the actual enrollment size, the two numbers might be similar, but only one is adopted as the true one in exact match.

Results

Descriptive statistics of the annotated datasets

For the Covid-Set, we annotated 300 RCT abstracts in Covid-19 research retrieved from PubMed. A summary of the annotation results regarding annotated sample size entities is shown in Table 1. Among the 300 abstracts labeled, 240 of them directly report a total sample size, while 265 report at least one sample size (any of total sample size, group/arm size, possible total sample size, possible group/arm size). Distribution for total sample sizes in current Covid-19 RCT studies are provided in Figure 3. The median total sample size for the Covid-Set is 128.5 and the mean is 23,981.

Table 1 – Summary of the annotated Covid-Set (P: possible)

# with/of the entity	Sample size entity class				
	Total	Group/ arm	Total (P)	Group/ arm(P)	any
Abstracts	240	154	59	14	265
Annotated entities	240	297	71	33	641

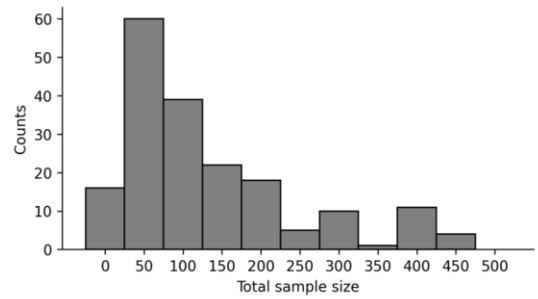

Figure 3 – Distribution of total sample size in Covid-Set

For the *General-Set*, as shown in Table 2, there are 87 out of 100 abstracts directly reporting total sample size and 58 reporting any group size. We labeled 267 sample size entities in total. The median total sample size for the *General-Set* is 255 and the mean is 1,335.

Table 2 – Summary of the annotated General-Set (P: possible)

# with/of the entity	Sample size entity class				
	Total	Group/ arm	Total (P)	Group/ arm(P)	any
Abstracts	87	58	35	4	96
Annotated entities	87	122	50	8	267

Model performance

Detailed model performance for the Covid-Set is shown in Table 3. Our sample size extractor has an average better performance compared with the state-of-the-art method. For Exact match, the best average F1 score is 0.73 and accuracy is 0.62, both obtained by our model with pretrained weights. The model also contributes to the highest accuracy and F1 score among all five folds, which are 0.72 and 0.80 respectively. For Loose match, when initialized with randomized weights, our model performs better than Trialstreamer consistently. Though, after transferring the learned weights, two methods have similar and closer performance. It is reasonable in that both methods are sufficient for finding the possible total sample size. Still, our method is superior to Trialstreamer and improves the four average-level metrics by 5% without pretraining, and by 2% with pretraining. On the other hand, pretraining on EBM-NLP and transferring the learned weights to both models improves the performance, especially on precision. An increase of over 5% can be observed on precision for both methods, and approximately 3% on other metrics. It indicates that pretraining could effectively facilitate the sample size extraction task, while saving efforts on annotation.

The results on the Covid-Set demonstrate that our extraction method reports a more accurate total sample size than the state-of-the-art method. To better illustrate the generalizability of our method, we conducted 3-fold cross-validation on the General-Set. As shown in Table 4, our SSE outperforms Trialstreamer on average, especially for Exact match. Again, models with pretraining weights perform better. The overall performance on the General-Set is lower than Covid-Set, because the General-Set is a smaller dataset and has less training data for each iteration. Additionally, we tested models that trained with the whole Covid-Set on the General-Set. Compared with results from cross-validation, there is an evident increase on precision and accuracy, and the increase of our method is greater than that of Trialstreamer, indicating that our method could achieve a greater improvement if more training data are available.

Table 3 – Model performance on Covid-Set (5-fold cross-validation; TS: Trialstreamer; SSE: Sample Size Extractor)

Model		Exact match				Loose match			
		Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy	F1
TS	**Avg**	0.58	0.75	0.55	0.66	0.70	0.78	0.63	0.74
	Best	0.69	0.82	0.65	0.74	0.79	0.86	0.70	0.80
TS + Pretraining	**Avg**	0.65	0.79	0.61	0.71	**0.76**	0.81	**0.69**	**0.79**
	Best	0.72	0.89	0.72	0.80	0.84	0.90	0.78	0.85
SSE	**Avg**	0.62	0.80	0.59	0.70	0.72	0.82	0.67	0.77
	Best	0.70	0.84	0.65	0.76	0.81	0.86	0.73	0.83
SSE + Pretraining	**Avg**	**0.67**	**0.80**	**0.62**	**0.73**	0.75	**0.82**	0.68	0.78
	Best	0.76	0.89	0.72	0.80	0.90	0.89	0.73	0.84

Table 4 – Model performance on General-Set (3-fold cross-validation if applies; TS: Trialstreamer; SSE: Sample Size Extractor; Trained on CS: Models trained on the Covid-Set and tested on the General-Set)

Model		Exact match				Loose match			
		Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy	F1
TS	**Avg (CV)**	0.41	0.52	0.37	0.46	0.67	0.65	0.53	0.66
	Trained on CS	0.45	0.61	0.41	0.52	0.68	0.71	0.57	0.69
TS + Pretraining	**Avg (CV)**	0.44	0.59	0.39	0.50	0.68	0.70	0.55	0.68
	Trained on CS	0.48	0.64	0.42	0.55	0.74	0.72	0.61	0.73
SSE	**Avg (CV)**	0.47	0.64	0.40	0.52	0.73	0.73	0.58	0.71
	Trained on CS	0.51	0.80	0.50	0.62	**0.75**	***0.86***	0.70	0.80
SSE + Pretraining	**Avg (CV)**	**0.50**	**0.77**	**0.47**	**0.60**	0.69	**0.81**	**0.61**	**0.74**
	Trained on CS	***0.57***	***0.68***	***0.51***	***0.63***	***0.86***	0.76	***0.71***	***0.81***

Examples of outputs on the General-Set for both extractors based on models with pretrained weights is reported in Table 5. We extracted all numbers in three abstracts, converted them into integers and reported both our model and Trialstreamer's predictions of total sample sizes along with confidences. Compared with Trialstreamer, our extractor correctly extracted all three sample sizes. For the case where both methods report the correct sample size, our method are more confident in the result as a higher likelihood is provided.

Table 5 – Examples of outputs for both methods (TS: Trialstreamer; SSE: Sample Size Extractor)

PMID	True total sample size (text)	Model	Extracted size (integer)	Confidence
2403603	"227"	**TS**	**227**	**0.74**
		SSE	**227**	**0.99**
29340676	"240"	**TS**	37	0.99
		SSE	**240**	**0.51**
21343580	"47"	**TS**	50	0.46
		SSE	**47**	**0.50**

Table 6 – Model performance on Covid-Set abstracts with only one or multiple annotated sample size entities (TS: Trialstreamer; SSE: Sample Size Extractor)

# of size entities	Model	Precision	Recall	Accuracy	F1
Only one	**TS**	0.85	0.725	0.66	0.78
	SSE	0.86	0.75	0.68	0.80
Multiple	**TS**	0.59	0.84	0.56	0.70
	SSE	0.61	0.84	0.57	0.71

Furthermore, as presented in Table 6, we noticed that extractors performed worse on abstracts with multiple sample size entities than that of only one annotated sample size entity. There is an obvious drop on precision, accuracy and F1 score of both our method and Trialstreamer. We then conducted *Exact match* evaluation on three kinds of abstracts in *Covid-Set*: 1) abstracts with possible total sample size entity annotated; 2) abstracts with group/arm sample size entity annotated and 3) abstracts with any sample size entity (total sample size, group/arm size, possible total/group/arm size). The results are based on models with pretraining. As shown in Table 7, our sample size extractor could better handle cases where several possible total sample sizes are provided than the state-of-the-art method. The low performance also demonstrates that the "possible" total sample size is the most misleading number for the extractor, which is consistent with our expectation. In addition, the precision and accuracy for abstracts with group/arm size entities or any entities is higher than that of Trialstreamer, validating again that our extractor could better identify the accurate total sample size in circumstances that other cohort sizes are mentioned.

Table 7 – Model performance on Covid-Set abstracts with specific annotated sample size entities (TS: Trialstreamer; SSE: Sample Size Extractor)

Entities	Model	Precision	Recall	Accuracy	F1
possible total	**TS**	0.29	0.59	0.31	0.39
	SSE	0.37	0.79	0.37	0.74
group/ arm	**TS**	0.64	0.87	0.60	0.74
	SSE	0.65	0.87	0.61	0.74
any	**TS**	0.68	0.79	0.59	0.73
	SSE	0.69	0.80	0.61	0.74

Discussion

Implication of the results

For the *Covid-Set*, the average recall is always above 80%. Our model's performance on both *Covid-Set* and *General-Set* proves that our method becomes the new state-of-the-art, indicating the efficiency of the tool and showing significant improvement over Trialstreamer, regardless of pretraining. The accomplishments are achieved by including and separating features regarding sets of keywords, comparing the values of all possible numbers. Evaluation on the three different types of abstracts further proves the advancement of our sample size

extractor. As for pretraining, the increase in all average-level metrics illustrates its advantage, and therefore provides another option for researchers to aid training models and achieving better performance with less cost and efforts on annotation. With better precision and accuracy, our extractor can be included in a comprehensive information extraction system or used to build a database of total sample size for RCT reports, and thereby enable clinicians to filter by sample size when searching for the RCT literature. The new tool paves the way for meta-analysis on sample size and facilitates more evidence-based medicine tasks.

Error analysis

The performance drops when multiple sample sizes are reported in abstracts so that the extractor has difficulty distinguishing the total sample size from other patient counts included in the data analysis rather than counts used in trial-designed phase. For example, given the phrase "*of 491 patients randomly assigned to a group, 423 contributed primary end point data*", more sophisticated computational reasoning is needed to classify 491 as the total sample size and 423 as the effect size. Specifically, on the *Covid-Set*, the precision is 0.86 for abstracts with only one sample size entity and 0.61 for abstracts with total, group, and effect sizes. Our results demonstrate that distinguishing the true total sample size among all size-related entities is still the most challenging task to address, and our extractor outperforms Trialstreamer on this task to achieve the new state-of-the-art.

Limitations and future work

Though not common, some abstracts do not report total sample size directly but report different group/arm sizes separately. Both our current gold standards and extractor could not perform reasoning over the extracted numbers to compute the total sample sizes. For the annotated dataset, since we obtained the true sample size by marking the size from all existing numbers, additional review may be required to calculate the total sample size manually. One possible solution could be to first determine whether total sample size is mentioned in abstracts, followed by a module to perform calculation on all given numbers and recommend feasible sample sizes. Then the extractor predicts the most probable total sample size among these numbers. Moreover, the current extractor only extracts the total sample size. In the future it needs to be extended to extract group size or arm size, and detect and link arm sizes to the arms.

Conclusions

We presented a new sample size extractor with better performance than the state-of-the-art tool on two sets of RCT abstracts. Our extractor can be applied to collect total sample sizes from any RCT abstracts, build a database and pave the way for meta-analysis on sample size. More significantly, a more reliable sample size extractor enables researchers to provide a more accurate and comprehensive description of the study, and design a better IE system upon.

Acknowledgements

This work was supported by the National Library of Medicine grant 5R01LM009886-11.

References

[1] H.O. Stolberg, G. Norman, and I. Trop, Randomized controlled trials, *AJR Am J Roentgenol* **183** (2004), 1539-1544.

[2] K.F. Schulz, D.G. Altman, D. Moher, and C. Group, CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials, *Int J Surg* **9** (2011), 672-677.

[3] D.E. Appelt, Introduction to information extraction, *Ai Communications* **12** (1999), 161-172.

[4] C. Schardt, M.B. Adams, T. Owens, S. Keitz, and P. Fontelo, Utilization of the PICO framework to improve searching PubMed for clinical questions, *BMC medical informatics and decision making* **7** (2007), 1-6.

[5] C. Baladron, A. Santos-Lozano, J.M. Aguiar, A. Lucia, and J. Martin-Hernandez, Tool for filtering PubMed search results by sample size, *J Am Med Inform Assoc* **25** (2018), 774-779.

[6] J. Sargeant, Qualitative Research Part II: Participants, Analysis, and Quality Assurance, *J Grad Med Educ* **4** (2012), 1-3.

[7] K. Suresh and S. Chandrashekara, Sample size estimation and power analysis for clinical research studies*, J Hum Reprod Sci* **5** (2012), 7-13.

[8] A.J. Sutton, N.J. Cooper, D.R. Jones, P.C. Lambert, J.R. Thompson, and K.R. Abrams, Evidence-based sample size calculations based upon updated meta-analysis, *Stat Med* **26** (2007), 2479-2500.

[9] P.H. Lee, The quality of the reported sample size calculation in clinical trials on COVID-19 patients indexed in PubMed, *Eur J Intern Med* **77** (2020), 139-140.

[10] T. Kang, S. Zou, and C. Weng, Pretraining to Recognize PICO Elements from Randomized Controlled Trial Literature, *Stud Health Technol Inform* **264** (2019), 188-192.

[11] I.J. Marshall, B. Nye, J. Kuiper, A. Noel-Storr, R. Marshall, R. Maclean, F. Soboczenski, A. Nenkova, J. Thomas, and B.C. Wallace, Trialstreamer: A living, automatically updated database of clinical trial reports, *J Am Med Inform Assoc* **27** (2020), 1903-1912.

[12] B. Nye, J. Jessy Li, R. Patel, Y. Yang, I.J. Marshall, A. Nenkova, and B.C. Wallace, A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature, *Proc Conf Assoc Comput Linguist Meet 2018* (2018), 197-207.

[13] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J.i. Tsujii, BRAT: a web-based tool for NLP-assisted text annotation, in: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, **2012**, pp. 102-107.

[14] S. Moen and T.S.S. Ananiadou, Distributional semantics resources for biomedical text processing, *Proceedings of LBM* (2013), 39-44.

Address for correspondence

Chunhua Weng, Department of Biomedical Informatics, Columbia University, 622 W 168th Street, PH-20 room 407, New York, NY 10032, USA. E-mail: chunhua@columbia.edu