HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2022 April 01.

Published in final edited form as:

J Biomed Inform. 2021 April ; 116: 103717. doi:10.1016/j.jbi.2021.103717.

Toward assessing clinical trial publications for reporting transparency

Halil Kilicoglu[a,b,*], **Graciela Rosemblat**[b], **Linh Hoang**[a], **Sahil Wadhwa**[c], **Zeshan Peng**[b], **Mario Malički**[d], **Jodi Schneider**[a], **Gerben ter Riet**[e,f]

[a]School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA [b]U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA [c]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, USA [d]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA [e]Urban Vitality Center of Expertise, Faculty of Health, Amsterdam University of Applied Sciences, Amsterdam, the Netherlands [f]Department of Cardiology Heart Center, Amsterdam UMC, University of Amsterdam, the Netherlands

Abstract

Objective: To annotate a corpus of randomized controlled trial (RCT) publications with the checklist items of CONSORT reporting guidelines and using the corpus to develop text mining methods for RCT appraisal.

Methods: We annotated a corpus of 50 RCT articles at the sentence level using 37 fine-grained CONSORT checklist items. A subset (31 articles) was double-annotated and adjudicated, while 19 were annotated by a single annotator and reconciled by another. We calculated inter-annotator agreement at the article and section level using MASI (Measuring Agreement on Set-Valued Items) and at the CONSORT item level using Krippendorff's α. We experimented with two rule-based methods (phrase-based and section header-based) and two supervised learning approaches (support vector machine and BioBERT-based neural network classifiers), for recognizing 17 methodology-related items in the RCT Methods sections.

Results: We created CONSORT-TM consisting of 10,709 sentences, 4,845 (45%) of which were annotated with 5,246 labels. A median of 28 CONSORT items (out of possible 37) were annotated

*Corresponding author at: School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA., halil@illinois.edu (H. Kilicoglu).

CRediT authorship contribution statement

Halil Kilicoglu: Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data curation, Writing - original draft, Writing - review & editing, Supervision, Project administration. **Graciela Rosemblat:** Data curation, Conceptualization, Writing - review & editing. **Linh Hoang:** Methodology, Software, Validation, Investigation, Writing - review & editing. **Sahil Wadhwa:** Methodology, Software, Validation, Investigation. **Zeshan Peng:** Methodology, Software, Validation, Investigation. **Mario Malički:** Data curation, Conceptualization, Writing - review & editing. **Jodi Schneider:** Data curation, Conceptualization, Writing - review & editing. **Gerben ter Riet:** Data curation, Conceptualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2021.103717.

per article. Agreement was moderate at the article and section levels (average MASI: 0.60 and 0.64, respectively). Agreement varied considerably among individual checklist items (Krippendorff's α= 0.06–0.96). The model based on BioBERT performed best overall for recognizing methodology-related items (micro-precision: 0.82, micro-recall: 0.63, micro-F1: 0.71). Combining models using majority vote and label aggregation further improved precision and recall, respectively.

Conclusion: Our annotated corpus, CONSORT-TM, contains more fine-grained information than earlier RCT corpora. Low frequency of some CONSORT items made it difficult to train effective text mining models to recognize them. For the items commonly reported, CONSORT-TM can serve as a testbed for text mining methods that assess RCT transparency, rigor, and reliability, and support methods for peer review and authoring assistance. Minor modifications to the annotation scheme and a larger corpus could facilitate improved text mining models. CONSORT-TM is publicly available at https://github.com/kilicogluh/CONSORT-TM.

Keywords

Reporting guidelines; CONSORT; Corpus annotation; Text mining; Sentence classification

1. Background

Rigor and reproducibility of scientific research has been widely questioned in recent years [1]. The biomedical research enterprise is at the center of these discussions and concerns have been raised about research waste in biomedicine [2]. To address these concerns, various interventions (e.g., grant application requirements [3], transparency and openness guidelines [4], data sharing principles [5], peer review of study protocols [6]) have been proposed.

Incomplete reporting and lack of transparency is a common problem in biomedical publications and can hinder efforts to replicate prior research findings because important methodological details may be missing [7,8]. When key elements such as sample size estimation, randomization/blinding procedures or study limitations are not reported, it can also be difficult to assess the rigor and reliability of a study for evidence synthesis. While transparent reporting alone does not guarantee a rigorous and reliable study, it is an essential step toward assessing a study for such criteria [9]. Conversely, a well-executed but poorly reported study is unlikely to pass a thorough peer review process.

Reporting guidelines have been proposed to improve transparency and accuracy of reporting in publications resulting from biomedical studies. These include CONSORT for randomized controlled trials (RCTs) [10], ARRIVE for pre-clinical animal studies [11], and PRISMA for systematic reviews [12], all developed under the umbrella of the EQUATOR Network [13]. While they have been endorsed by many high-impact medical journals [14], adherence to reporting guidelines remains sub-optimal [15].

RCTs are considered a cornerstone of evidence-based medicine [16] and are placed high in the "evidence pyramid" [17]. To fully exploit their theoretical advantages, they need to be rigorously designed and conducted, and clearly and accurately reported. The CONSORT Statement for RCT reporting, developed in 1996 [10] and updated in 2010 [18,19], consists

of a 25-item checklist and a participant flow diagram. CONSORT has been extended over the years to RCT abstracts [20] as well as to specific types of RCT designs, such as cluster trials [21], and interventions, such as non-pharmacologic treatments [22]. CONSORT is the best known reporting guideline, endorsed by 585 biomedical journals[1] and prominent editorial organizations. Prior research has shown that journal endorsement was correlated with completeness of reporting, while the reporting of key methodological details, such as allocation concealment, was lacking even in endorsing journals [23].

Studies assessing adherence to CONSORT rely on manual analysis of a relatively small number of publications [23]. A text mining tool that can locate statements corresponding to CONSORT checklist items could facilitate adherence assessment at a much larger scale. Such a tool would also have broader utility. For example, it could assist authors in ensuring completeness of their reports, journal editors in enforcing transparency requirements, and peer reviewers, systematic reviewers, and others in critically appraising RCTs for transparency and rigor criteria [24].

Most modern text mining methods need to be trained on large amounts of representative, labeled text to be successful. Developing such corpora based on the CONSORT guidelines is a challenging task, given the large number of checklist items, and the expertise required for identifying statements that correspond to these items in RCT publications. In this study, we aimed to create a corpus of RCT articles annotated with CONSORT checklist items at the sentence level and use it to train and evaluate text mining methods. Herein, we describe the annotation process, corpus statistics, and inter-annotator agreement for the corpus (named CONSORT-TM). We also present baseline experiments in classifying sentences from RCT articles with these items. We limit these experiments to Methods sections and methodology-specific CONSORT items, covering key methodological details most relevant to trial rigor and robustness.

2. Related work

Text mining research on RCT articles has primarily concentrated on annotating and extracting study characteristics relevant for article screening for systematic reviews and evidence synthesis [25,26].

Much attention has been paid to PICO elements (Population, Intervention, Comparator, and Outcome), used in evidence-based medicine [16] to capture the most salient aspects of clinical intervention studies [27–30]. PICO elements were annotated at the noun phrase level [27,29] as well as the sentence level [27,28]. Semi-automatic methods [28] and crowdsourcing [29] have also been used to generate training data. For example, EBM-NLP corpus [29] consists of 5000 abstracts, annotated at the text span level through crowdsourcing. In their corpus, some PICO elements are further subcategorized at more granular levels (e.g., Physical Health as a subcategory of Outcomes). Text mining methods used to extract PICO characteristics ran the gamut from early knowledge-based and traditional machine learning methods [27] to more recent semi-supervised [28] and neural

[1] http://www.consort-statement.org/about-consort/endorsers, retrieved on 11/02/2020.

network models, including recurrent neural networks (RNN) such as Long Short Term Memory (LSTM) [29,30]. PICO variants, such as PIBOSO (B: background, S: study design, O: other), have also been considered. A corpus of 1000 abstracts annotated at the sentence level with PIBOSO elements has been published (PIBOSO-NICTA) [31], and various machine learning models (conditional random fields, neural network models) trained on this corpus have been reported [31–34].

Other research looked beyond PICO and its variants. Kiritchenko et al. [35] annotated 21 elements from 132 full-text articles, including eligibility criteria, sample size, and drug dosage. To automatically recognize these elements, they used a two-stage pipeline which consisted of machine learning-based sentence classification followed by phrase matching based on regular expressions. Hsu et al. [36] focused on information related to statistical analysis in full-text articles (hypothesis, statistical method, interpretation) and used rule-based methods to identify these elements in a dataset of 42 full-text articles (on non-small-cell lung carcinoma) and map them to a structured representation. Marshall et al. [37] identified risk-of-bias statements in RCT publications and categorized the studies as high or low risk with respect to risk categories, including sequence generation and allocation concealment. Their dataset was semi-automatically generated from the Cochrane Database of Systematic Reviews, and they used support vector machine (SVM) classifiers to jointly learn the risk-of-bias levels and the supporting statements in the article [37]. We have previously developed methods to automatically recognize limitation statements in clinical publications using a manually annotated corpus of 1257 sentences [38].

Annotation and extraction of key statements has also been considered for other publication types, such as pre-clinical animal studies and case reports [39–42]. One work that is particularly relevant to ours is SciScore [42], a tool which assesses life sciences articles for rigor and transparency, by extracting characteristics including subject's sex, sample size calculation, institutional review board statements, anti-bodies and cell lines, and calculating a score based on them. They manually annotated several datasets to train named entity recognizers; however, these datasets are not publicly available, to the best of our knowledge.

Our work extends earlier research in several ways. Owing to our focus on CONSORT, we annotated a greater number of study characteristics, some of which have not been annotated in any publicly available dataset, to our knowledge. As a result, our dataset is more suitable for training text mining methods that address transparency more comprehensively. We also believe that methods trained on CONSORT- TM can be useful for a broader set of tasks other than systematic review screening and authoring. In addition to traditional rule-based and supervised learning approaches, we also experimented with neural network classifiers based on contextualized language models, specifically BioBERT [43].

3. Materials and methods

3.1. Corpus annotation

We manually annotated 50 articles from 11 journals with 37 fine- grained CONSORT checklist items (the complete checklist and the list of 11 journals are provided in Supplementary file 1)[2]. We used a modified version of Cochrane's sensitivity and precision-

maximizing query for RCTs as the search strategy[3] to sample RCT articles for annotation. We limited the search to articles published in 2011 or later, since the most recent CONSORT statement was published in 2010. Our search resulted in 563 articles (search date: July 16, 2018). We randomly selected a convenience sample of 50 articles for annotation.

We downloaded PubMed Central (PMC) XML files of the articles and processed them with an in-house sentence splitter and section recognizer to identify the sections and the sentences in each section. In addition to the title, abstract, and the full text of the article, we also extracted its back matter, as some CONSORT items are often stated in this section (e. g., Funding (25)).

The annotation task consisted in labeling sentences of the articles with relevant CONSORT items. First, two authors (HK and GR) labeled sentences in one article using the CONSORT explanation and elaboration article [19] as their guide. These annotations were then discussed and adjudicated, and were provided as an example annotated document to all annotators. They also formed the basis of the preliminary annotation guidelines. Next, six annotators double-annotated 30 articles independently (10 articles per annotator). The annotators are experts in meta-research (with medical degrees), clinical trial methodology, biomedical informatics, text mining, and linguistics, all well-versed in scholarly communication, specifically biomedical literature. The annotators were expected to read the guidelines and comment on them before beginning annotation. The guidelines were iteratively updated throughout the study, as needed (the guideline document is provided in Supplementary file 1). One annotator labeled a single article due to time constraints, their remaining articles were assigned to other annotators, while still ensuring double annotation of each article. Each of the 30 articles were then adjudicated by one of the two authors (HK or GR). Next, another 19 articles were annotated by a single author (GR), whose annotations were inspected and corrected by the first author (HK).

We used a custom, web-based annotation tool for labeling articles. The tool allowed the annotator to navigate the article using tabs (corresponding to article sections) and select labels for each sentence from a drop-down list. A link to the annotation guidelines was provided, as well as a link to the PDF of the article for better contextualization and access to tables and figures. The user was afforded the ability to check which items they had and had not annotated in the article (Item Check button). A screenshot of the annotation interface is shown in Fig. 1. Annotation adjudication was performed using a modified version of the same tool.

3.2. Inter-annotator agreement

We calculated inter-annotator agreement for 30 double-annotated articles. Since each sentence could be labeled with multiple items, we chose MASI measure (Measuring Agreement on Set-Valued Items) [44] to calculate agreement. MASI is a distance metric for comparing two sets and incorporates the Jaccard index. It penalizes more the case in which

[2] In this paper, we capitalize CONSORT checklist item names and indicate the corresponding item number in CONSORT guidelines in parentheses (e.g., Trial Design (3a)).
[3] https://work.cochrane.org/pubmed.

two compared sets have disjoint elements than the case in which one set subsumes the other. We calculated pairwise inter-annotator agreement using MASI at the article and section levels. We also calculated Krippendorff's α [45] for each CONSORT item. This measure was chosen since the annotation involved more than two annotators and the annotation data was incomplete (i.e., not all sentences were labeled by all annotators). We excluded CONSORT items annotated at the article level (1a and 1b) from agreement calculations. We also excluded some sentences (title, abstract, and section headers), which the annotators were instructed not to annotate, from these calculations.

3.3. Text mining methods

We cast the problem of associating sentences with CONSORT items as a sentence-level, multi-label classification task. In this study, for text mining, we specifically focused on the Methods section of RCT articles and the 17 fine-grained items associated with this section in CONSORT guidelines (items 3a to 12b), as they are highly relevant for assessing rigor and reliability of a clinical trial.

3.3.1. Rule-based methods—Our rule-based methods were based on two observations:

- When articles are organized by subsections, the subsection headers can provide clues about the items discussed in those sections (e.g., a subsection with the header *Participants* is likely to contain sentences discussing Eligibility Criteria (4a)).
- Certain words/phrases in a sentence are clues for particular CONSORT items (e.g., *block size* is likely to indicate a sentence on Sequence Generation (8a)).

To identify such patterns, we collected the list of all Methods subsection headers from a separate, larger set of RCT articles from PMC (about 18K articles with a clinicaltrials.gov registry number), ranked them by their frequency, and mapped the most frequent ones (or their relevant substrings) to CONSORT items. For rare checklist items (e.g., Changes to Outcomes (6b)), we also examined the long tail to identify at least a few relevant headers. An example mapping relates the string *concealment* to the CONSORT item Allocation Concealment (9). We assign this label to the sentences from the sections with the following headers: *Allocation concealment mechanism*, *Concealment of allocation*, *Concealment of group allocation to participants*. This method uses 48 mappings for 17 Methods-specific items. For phrase-based classification, we generated a set of predictive phrases, including *power to detect* for Sample Size Determination (7a) and *masked to treatment* for Blinding Procedure (11a), from the subsections with the headers corresponding to CONSORT items (a total of 232 phrase mappings). If a sentence contains phrases associated with a particular CONSORT item, we label the sentence with that item.

3.3.2. Supervised machine learning—Two other methods were based on supervised machine learning. One approach involved a support vector machine (SVM) classifier that uses as features *tf-idf* representation of the sentence in addition to the enclosing subsection header. The section header was prepended to the sentence and included in *tf-idf* calculation. We excluded common English words using the NLTK stopword list and used the

LIBLINEAR SVM implementation in the scikit-learn package [4]. C regularization parameter was set to 10 after a grid search. The classifier was embedded into a one-vs-rest classifier to enable prediction of multiple labels for each sentence.

The other approach involved a neural network classifier based on BioBERT [43], a variant of the BERT model [46] pre-trained on the biomedical literature. BERT is a bidirectional Transformer [47] trained on language modeling tasks over massive datasets in an unsupervised manner. The BERT encoder produces a vector of hidden states from text input, which can then be fine-tuned for a supervised task, such as sentence classification. BERT has been shown to yield state-of-the-art results on many NLP tasks in recent years, even with small datasets, motivating its use in this study. BioBERT was pre-trained on PubMed abstracts and PMC full-text articles in addition to original BERT training data. Sentence text and its subsection header were fed as input to the BioBERT encoder, whose output was then used to train the final sigmoid layer for multi-label classification. We used the simpletransformers package to implement multi-label text classification[5]. The following hyperparameters were used for model training and evaluation: batch size (4), learning rate (3e-5), number of epochs (30), optimizer (Adam), dropout (0.1).

4. Results

4.1. Corpus annotation

The descriptive statistics of CONSORT-TM are given in Table 1. The corpus contains over 10K sentences, 45% of which were annotated (4845 sentences). The total number of annotations is 5246, indicating that about 6.5% of the annotated sentences were annotated with multiple items. Annotation density (annotations per sentence) is 0.48 (1.08, if only the sentences with at least one annotation are considered).

The highest number of items associated with a sentence is 5. For instance, the sentence below was labeled with the items Trial Design (3a), Sequence Generation (8a), Allocation Concealment (9), Randomization Implementation (10), and Similarity of Interventions (11b).

1. *Patients were randomly assigned, using a computer-generated randomization schedule, from a central location utilizing an interactive voice response system with blinded medication kit number allocation in a 2:1 ratio to identical-appearing tablets of HZT-501 (800 mg ibuprofen and 26.6 mg famotidine) or ibuprofen (800 mg) thrice daily for 24 weeks*.

We found that out of 37 fine-grained items, a median of 28 were annotated per article. The most complete article included 35 CONSORT items, missing discussion relating to Changes to Trial Design (3b) and Interim Analyses/Stopping Guidelines (7b). Since both can be considered contingent items, we considered this article the only one in CONSORT-TM that was fully compliant with the CONSORT guidelines.

[4] https://scikit-learn.org/stable.
[5] https://github.com/ThilinaRajapakse/simpletransformers.

Table 2 shows the number of articles annotated with each checklist item and the average number of sentences per article annotated for the item. The results show that 80% of the articles identified themselves as RCTs in the title (n = 40) and almost all had structured abstracts (n = 49). The major PICO characteristics captured as CONSORT checklist items (Eligibility Criteria (4a), Interventions (5), and Outcomes (6a)) were well-reported (n = 49, n = 50, n = 50, respectively). Similarly, all articles discussed Trial Design (3a) and Statistical Methods for Outcome Comparison (12a) and reported Baseline Data (15) and Outcome Results (17a). Items that indicate some modification to the trial after its launch were infrequently reported. These include Changes to Trial Design (3b) (n = 4), Changes to Outcomes (6b) (n = 5), and Trial Stopping (14b) (n =6). Protocol Access (24) was the least reported item (n =7), among the rest. Randomization and masking-related items (8a to 11a), which were often found to be reported inadequately in earlier studies, were moderately reported in our collection (60%−80%), except Allocation Concealment (9) (n = 19, 38%).

When annotated at all, a checklist item is labeled over 3.47 sentences on average per article (range: 1.13–14.38). While most are annotated in a few sentences (median: 2.19, IQR: 1.33–4.31), a few items are annotated in more than 10 sentences per article (Outcomes (6a), Outcome Results (17a), and Interpretation (22)). We find that items annotated over many sentences cover several aspects of the RCT or have been defined somewhat broadly in CONSORT. For example, the Outcomes category (6a) covers not only primary and secondary outcome measures, but also how and when they were assessed, leading to a large number of sentences being labeled with this item. Similarly, Interpretation (22) is defined as “interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence,” which can encompass the majority of sentences in Discussion sections.

Annotators were instructed to label CONSORT items in the sections with which they are typically associated (excluding those in Other category in CONSORT, such as Registration (23), which could be annotated in any section). However, CONSORT items were not always reported in their associated sections by the authors. This most commonly occurred with Periods of Recruitment and Follow-up (14a) (n = 28), Numbers Analyzed (16) (n = 15), Participant Flow (13a) (n = 12), and Data Collection Setting (4b) (n = 11). The first three were often reported in Methods sections, instead of the Results sections, and the latter vice versa.

4.2. Inter-annotator agreement

Inter-annotator agreement at the article level was moderate (mean MASI = 0.60, median = 0.63) (Table 3). An interim analysis after 10 articles were double-annotated and adjudicated showed a median MASI of 0.57, indicating that inter-annotator agreement improved over the course of the study. Agreement was highest in back matter sections (Acknowledgements, etc.), where Registration (23) or Funding (25) were often reported (section mean MASI = 0.89). While the agreement in Introduction sections was low in the interim analysis (mean MASI = 0.50), it improved after guidelines about annotating Background (2a) were clarified (MASI = 0.67). Agreement in Methods and Discussion sections were similar (mean MASI of 0.59 and 0.58, respectively), with more spread in agreement in Discussion sections. The agreement in Results sections was lowest (mean MASI = 0.50). The agreement of each

annotator with other annotators was largely similar (mean annotator MASI = 0.60, range: 0.58–0.64). The annotator who annotated a single article was excluded from this calculation.

At the checklist item level (Fig. 2), agreement was highest for Registration (23) (Krippendorff's α=0.96). There was also high agreement for Sample Size Determination (7a) (α=0.85), followed by Eligibility Criteria (4a) (α=0.79) and Objectives (2b) (α=0.71). Agreement was lowest for Numbers Analyzed (16) (α=0.06), Generalizability (21) (α=0.14), and Binary Outcome Results (17b) (α=0.15). Average α was 0.47.

4.3. Text mining results

The results of baseline experiments with four methods are shown in Table 4. The phrase-based method largely outperformed the section header-based method, especially for rare items (e.g., Changes to Outcomes (6b)), which are unlikely to be discussed in dedicated subsections. In contrast, the section header-based method performed better for common items that are often discussed in dedicated subsections (Outcomes). Both of these methods were mostly outperformed by the supervised learning algorithms, with the BioBERT-based model performing best overall in all categories (0.82 precision, 0.63 recall, 0.72 F_1 score, 0.812 AUC). This model performed particularly well for items with larger number of annotations (Interventions, Outcomes, Statistical Methods for Outcomes). However, it performed poorly for rare items yielding no correct predictions for several items, which led to its lower macro-averaged performance, compared to macro-averaged performance of the phrase-based method and the linear SVM classifier. On the whole, we obtained best performance for Sample Size Determination. Model performance differences on individual items (calculated using McNemar's test) were largely statistically significant at 95% confidence level, excluding the infrequent items, as shown in Table 4.

The results based on model combinations are provided in Table 5. We only provide the micro- and macro-averaged results for these combinations. Majority vote yields the best precision overall (0.78 micro, 0.74 macro) among combination methods, while the aggregation of all method predictions yields the best recall (0.87 micro, 0.71 macro). Linear SVM and BioBERT model combination (SVM + BERT) improves upon the best base model by about 2 F_1 points (0.72 to 0.74).

5. Discussion

5.1. Annotation

We obtained moderate agreement in annotating fine-grained CONSORT checklist items on average. However, agreement varied considerably between different items. It was tempting to annotate at coarser granularity to achieve higher agreement. We did not do this, as our primary goal was to evaluate CONSORT in its entirety as annotation target and ultimately we would like to address all aspects of RCT reporting. This clearly made the annotation task more difficult and reduced agreement.

We had highest agreement on well-defined items, such as Objectives (2b), Sample Size Determination (7a), and Registration (23). Agreement was lowest on Numbers Analyzed (16), often only reported in tables/figures. While the annotators were instructed to annotate

the captions in these cases, this was not done consistently. Some categories with low agreement are those that are easy to confuse with others. For example, it is challenging to determine whether an outcome result sentence should be labeled as Outcome Result (17a), Binary Outcome Result (17b), or Ancillary Analyses (18) (α's of 0.41, 0.15, and 0.23), as this requires keeping track of all different outcomes and analyses discussed in an article while annotating, a cognitively demanding task. We found that it was easy to over-annotate some broadly conceived items (Background, Interventions, Outcomes, Interpretation). We instructed annotators to limit Background annotations to the two most representative sentences as these sentences generally did not directly relate to the RCT study under consideration. We did not do so for other items to avoid missing important study characteristics, although it may be reasonable to limit the annotation of broadly conceived items to the most representative sentences.

Low agreement on some items may raise questions about the reliability of the annotations and their use for text mining, as α coefficient lower than 0.67 is sometimes considered unreliable. We note that the sentences on which agreement was measured were further adjudicated prior to being used for text mining and the rest were examined by two annotators; therefore, we expect that the input labels for text mining methods should be reasonably reliable. Of course, it would be ideal for all annotators to discuss all disagreements to reach a consensus; however, given the complexity of the task, this was not deemed feasible in this study. It is also worth noting that α, originally developed for content analysis, emphasizes replicability of annotations, whereas the focus is generally on usefulness in developing natural language processing corpora [48]. Lastly, reported inter-annotator agreement in biomedical corpora focusing on similar phenomena is comparable to ours. For example, in the PIBOSO-NICTA corpus [31], Cohen's κ was measured as 0.71, 0.63, 0.61, and 0.41 for Outcomes, Population, Intervention, and Study Design, respectively. We calculated κ for the corresponding CONSORT items (6a, 4a, 5, 3a) and obtained 0.56, 0.82, 0.53, and 0.64, respectively. Note also that their annotation focused on abstracts only, which are arguably much easier to annotate than full-text articles.

Nevertheless, low agreement on some items suggests that using CONSORT checklist items as annotation target as-is may not be the most appropriate strategy. For alternative categorizations, we investigated whether higher inter-annotator agreement could be achieved by collapsing related items into a single category. Higher agreement was achieved for the following combinations:

- Randomization Type (8b) and Randomization Implementation (10): combined α=0.50 vs. 0.48 and 0.35, respectively.
- Statistical Methods for Outcomes (12a) and Statistical Methods for Other Analyses (12b): combined α=0.67 vs. 0.53 and 0.28, respectively.
- Participant Flow (13a) and Participant Loss/Exclusion (13b): combined α= 0.51 vs. 0.44 and 0.45, respectively.
- Outcome Result (17a), Binary Outcome Result (17b), and Ancillary Analyses (18): combined α= 0.56 vs. 0.41, 0.15, and 0.23, respectively.

Furthermore, we observed that the most frequent confusion by the BioBERT-based model was predicting Statistical Methods for Outcomes (12a) instead of Statistical Methods for Other Analyses (12b). From an annotation and text mining perspective, it may be beneficial to combine these categories, as this would lead to more consistent annotations and likely better text mining performance; however, this needs to be weighed against whether the resulting categorization would still serve the use cases under consideration.

On a related note, we also considered merging several infrequent and contingent items into their more frequent siblings. These combinations were:

- Changes to Trial Design (3b) into Trial Design (3a)
- Changes to Outcomes (6b) into Outcomes (6a)
- Trial Stopping (14b) into Recruitment/Follow-Up (14a)

This did not affect inter-annotator agreement significantly, since there were only a small number of annotations for the first items. However, from a text mining standpoint, it may be also be advantageous to merge these categories.

We also noted that some important trial characteristics are not captured by CONSORT guidelines. One example is the lack of a checklist item corresponding to research ethics and consent statements, which are often reported in publications. While it is pointed out in the CONSORT explanation and elaboration article [19] that this is by design, we believe it would be important to include this as an additional category.

While we did not formally measure the time it takes to annotate an article, we found anecdotally that it took about 2–3 hours of concentrated effort. Some variation is to be expected based on annotator's knowledge of the subject matter discussed in the article and clinical trial methodology. Despite the annotation challenges, we believe that CONSORT-TM can be useful as a benchmark for automated systems that aim to measure CONSORT adherence or extract study characteristics of RCTs. Considering that it involves 37 categories, CONSORT-TM can also serve as a challenging biomedical sentence classification corpus.

5.2. Text mining methods

Supervised classification methods had a clear advantage for common checklist items (e.g., Interventions, Outcomes), while only the phrase- based method performed relatively well for infrequent items. This confirms the need for large quantities of labeled data for training effective supervised learning models. We found that approaches that combine predictions from different base models can also improve classification performance. An effective approach could be to rely on the phrase-based method for predicting rare items and combine those with predictions from the BioBERT-based model.

We used standard settings for supervised learning, and it may be possible to achieve better performance with more advanced features or modeling approaches. In the case of SVM classification, we experimented with semantic features derived from MetaMap [49] (entities and their semantic types extracted from sentences). Semantic types features slightly

improved results (although not statistically significantly; results not shown), whereas concept features caused a minor degradation. In another approach, we can cast the problem as a sequence labeling task, leveraging the fact that discussion of items often follows a predictable sequence (e.g., Methods sections generally begin with Study Design sentences).

In training the BioBERT-based model, a simple sigmoid layer was used on top of the BERT encoder for classification, which can be substituted by more layers or a more complex neural architecture, such as convolutional or recurrent neural network (CNN or RNN) for higher classification performance. Note, however, that the BioBERT model is already much more complex than the other methods reported here, and takes orders of magnitude longer to train compared to the SVM classifier (hours vs. seconds). Improvements due to additional layers or architectural features may not be sufficiently large to justify the added complexity.

Overall, our preliminary results were encouraging, although it is clear that there remains much room for improvement. Performance for several items may be acceptable for practical use (e.g., Eligibility Criteria, Sample Size Determination), whereas more work is needed for others.

A promising direction may be to use weak supervision to automatically annotate a large number of clinical trial publications using simple heuristics, such as the phrase-based or section header-based methods and then use the resulting (somewhat noisy) data to train more effective classifiers. We have shown that this improves the performance of recognizing sample size and power calculation statements [50], and we plan to extend it to recognition of other CONSORT items as well. Data augmentation [51] can also be used to generate a larger set of (synthetic) examples for infrequent items, which may also benefit machine learning performance.

5.3. Limitations

Our study has several limitations. First, our dataset consists of a small number of articles from 11 journals, which may not be representative of all RCT articles. While we collected more than 5K CONSORT annotations from this set, making this a moderate-sized corpus amenable to data-driven approaches, we acknowledge that text mining approaches would benefit from a larger corpus, considering the large number of categories. Second, our approach focuses solely on textual elements. Much important information about RCTs is in tables, figures, and increasingly in supplementary material. We attempted to address the former two by caption annotation, but we have not tried to incorporate supplementary material or other trial resources. Third, we only annotated whether a sentence discussed a particular checklist item. We did not annotate whether the item is adequately discussed or whether the study fulfills rigor criterion for that item. For instance, a sentence indicating that no blinding was performed was still annotated with the item Blinding Procedure (11a). While this would work well in a human-in-the-loop system where an expert assesses rigor based on sentences highlighted by our methods, it cannot automatically determine whether the study is sufficiently rigorous. Finally, as discussed, our classification methods were preliminary and, for the most part, their accuracy needs to be further improved for practical use.

6. Conclusion

We presented CONSORT-TM, a corpus annotated with CONSORT checklist items, and studied baseline sentence classification methods as well as their combinations to recognize a subset of these items. By adopting all the checklist items as our annotation target, we created a corpus with more granular clinical trial information compared to earlier similar efforts [28,29,31,35], one which can facilitate more comprehensive analysis of rigor and transparency of RCT articles. Because CONSORT has been developed and refined by a consortium of trialists, methodologists, and journal editors and reflects their recommendations and practical information needs, we believe that CONSORT-TM can support automated approaches to RCT assessment as well as authoring assistance. On the other hand, our experience showed some challenges in annotating against CONSORT, which leads us to suggest some refinements that may inform follow-up studies in using reporting guidelines for annotation and text mining.

In future work, we plan to extend our methods to section other than Methods and explore more sophisticated modeling approaches. Another direction is to extract relevant mention-level information and mapping it to standardized terms/identifiers in controlled vocabularies, such as the Ontology for Biomedical Investigations (OBI) [52] (e.g., instead of labeling a sentence as a Study Design sentence, extracting *double-blind study* as the study design). Such mapping could enable large-scale interrogation of clinical trial reports and reveal long-term trends. The principles learned with CONSORT can also be applied to annotating corpora targeting other reporting guidelines, such as ARRIVE for pre-clinical animal studies [11].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Tony Tse for his contribution to guideline development and annotation.

Funding

This work was supported in part by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

[1]. Baker M, 1,500 scientists lift the lid on reproducibility, Nature 533 (2016) 452–454, 10.1038/533452a. [PubMed: 27225100]

[2]. Chalmers I, Glasziou P, Avoidable waste in the production and reporting of research evidence, The Lancet 374 (9683) (2009) 86–89, 10.1016/s0140-6736(09)60329-9.

[3]. Collins FS, Tabak LA, Policy: NIH plans to enhance reproducibility, Nature 505 (7485) (2014) 612–613, 10.1038/505612a. [PubMed: 24482835]

[4]. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon T, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C,

Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T, Promoting an open research culture, Science 348 (6242) (2015) 1422–1425, 10.1126/science.aab2374. [PubMed: 26113702]

[5]. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B, The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018, 10.1038/sdata.2016.18. [PubMed: 26978244]

[6]. Chambers CD, Tzavella L, Registered reports: Past, present and future (2 2020). doi:10.31222/osf.io/43298. URL osf.io/preprints/metaarxiv/43298.

[7]. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP, Reproducible research practices and transparency across the biomedical literature, PLoS Biol. 14 (1) (2016) e1002333. [PubMed: 26726926]

[8]. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCall JM, Moxley RT, Narasimhan K, Noble LJ, Perrin S, Porter JD, Steward O, Unger E, Utz U, Silberberg SD, A call for transparent reporting to optimize the predictive value of preclinical research, Nature 490 (7419) (2012) 187–191, 10.1038/nature11556. [PubMed: 23060188]

[9]. Editorial Nature, Checklists work to improve science, Nature (556) (2018) 273–274. doi:10.1038/d41586-018-04590-7.

[10]. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, et al., Improving the quality of reporting of randomized controlled trials: the CONSORT statement, JAMA 276 (8) (1996) 637–639. [PubMed: 8773637]

[11]. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG, Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research, PLoS Biol. 8 (6) (2010) e1000412, 10.1371/journal.pbio.1000412. [PubMed: 20613859]

[12]. Moher D, Liberati A, Tetzlaff J, Altman DG, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, BMJ 339 (2009), 10.1136/bmj.b2535.

[13]. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG, Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network, BMC Med. 8 (1) (2010) 24, 10.1186/1741-7015-8-24. [PubMed: 20420659]

[14]. Shamseer L, Hopewell S, Altman DG, Moher D, Schulz KF, Update on the endorsement of CONSORT by high impact factor journals: a survey of journal ”Instructions to Authors” in 2014, Trials 17 (1) (2016) 301. [PubMed: 27343072]

[15]. Samaan Z, Mbuagbaw L, Kosa D, Debono VB, Dillenburg R, Zhang S, Fruci V, Dennis B, Bawor M, Thabane L, A systematic scoping review of adherence to reporting guidelines in health care literature, J. Multidiscip. Healthcare 6 (2013) 169.

[16]. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS, Evidence based medicine: what it is and what it isn’t, BMJ 312 (7023) (1996) 71–72, 10.1136/bmj.312.7023.71. [PubMed: 8555924]

[17]. Murad MH, Asi N, Alsawas M, Alahdab F, New evidence pyramid, BMJ Evidence-Based Med. 21 (4) (2016) 125–127.

[18]. Schulz KF, Altman DG, Moher D, CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials, BMJ 340 (2010) c332, 10.1136/bmj.c332. [PubMed: 20332509]

[19]. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG, CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials, BMJ 340 (2010), 10.1136/bmj.c869.

[20]. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, Schulz KF, CONSORT for reporting randomised trials in journal and conference abstracts, The Lancet 371 (9609) (2008) 281–283.

[21]. Campbell MK, Piaggio G, Elbourne DR, Altman DG, CONSORT 2010 statement: extension to cluster randomised trials, BMJ 345 (2012) e5661. [PubMed: 22951546]

[22]. Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P, CONSORT statement for randomized trials of nonpharmacologic treatments: a 2017 update and a CONSORT extension for nonpharmacologic trial abstracts, Ann. Intern. Med 167 (1) (2017) 40–47. [PubMed: 28630973]

[23]. Turner L, Shamseer L, Altman D, Weeks L, Peters J, Kober T, Dias S, Schulz K, Plint A, Moher D, Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals, Cochrane Database System. Rev. (11) (2012), 10.1002/14651858.MR000030.pub2.

[24]. Kilicoglu H, Biomedical text mining for research rigor and integrity: tasks, challenges, directions, Briefings Bioinformat. 19 (6) (2017) 1400–1414.

[25]. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S, Using text mining for study identification in systematic reviews: A systematic review of current approaches, System. Rev. 4 (1) (2015) 5, 10.1186/2046-4053-4-5.

[26]. Jonnalagadda SR, Goyal P, Huffman MD, Automating data extraction in systematic reviews: a systematic review, System. Rev. 4 (1) (2015) 78.

[27]. Demner-Fushman D, Lin J, Answering clinical questions with knowledge-based and statistical techniques, Comput. Linguist. 33 (1) (2007) 63–103, 10.1162/coli.2007.33.1.63.

[28]. Wallace BC, Kuiper J, Sharma A, Zhu M, Marshall IJ, Extracting PICO Sentences from Clinical Trial Reports Using Supervised Distant Supervision, J. Machine Learn. Res 17 (132) (2016) 1–25.

[29]. Nye B, Li JJ, Patel R, Yang Y, Marshall I, Nenkova A, Wallace B, A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 197–207. doi:10.18653/v1/P18-1019. URL https://www.aclweb.org/anthology/P18-1019.

[30]. Brockmeier AJ, Ju M, Przybyła P, Ananiadou S, Improving reference prioritisation with PICO recognition, BMC Med. Inform. Decis. Mak 19 (1) (2019) 256.

[31]. Kim SN, Martínez D, Cavedon L, Yencken L, Automatic classification of sentences to support Evidence Based Medicine, BMC Bioinformat. 12 (S-2) (2011) S5, 10.1186/1471-2105-12-S2-S5.

[32]. Hassanzadeh H, Groza T, Hunter J, Identifying scientific artefacts in biomedical literature: The Evidence Based Medicine use case, J. Biomed. Inform 49 (2014) 159–170, 10.1016/j.jbi.2014.02.006. [PubMed: 24530879]

[33]. Dernoncourt F, Lee JY, Szolovits P, Neural networks for joint sentence classification in medical paper abstracts, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 694–700. URL https://www.aclweb.org/anthology/E17-2110.

[34]. Jin D, Szolovits P, Hierarchical neural networks for sequential sentence classification in medical scientific abstracts, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3100–3109, https://doi.org/10.18653/v1/D18-1349 . URL https://doi.org/10.18653/v1/D18-1349https://www.aclweb.org/anthology/D18-1349. URL https://www.aclweb.org/anthology/D18-1349 .

[35]. Kiritchenko S, de Bruijn B, Carini S, Martin J, Sim I, ExaCT: automatic extraction of clinical trial characteristics from journal publications, BMC Med. Inform. Decis. Mak 10 (1) (2010) 56, 10.1186/1472-6947-10-56.

[36]. Hsu W, Speier W, Taira RK, Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature, in: AMIA Annual Symposium Proceedings, vol. 2012, American Medical Informatics Association, 2012, p. 350.

[37]. Marshall IJ, Kuiper J, Wallace BC, RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials, J. Am. Med. Inform. Assoc (2015) 193–201, 10.1093/jamia/ocv044. [PubMed: 26104742]

[38]. Kilicoglu H, Rosemblat G, Malički M, ter Riet G, Automatic recognition of self-acknowledged limitations in clinical research literature, J. Am. Med. Inform. Assoc 25 (7) (2018) 855–861. [PubMed: 29718377]

[39]. Névéol A, Lu Z, Automatic integration of drug indications from multiple health resources., in: Veinot TC, Çatalyürek Ümit V., Luo G, Andrade H, Smalheiser NR (Eds.), IHI, 2010, pp. 666–673.

[40]. Zeiss CJ, Shin D, Vander Wyk B, Beck AP, Zatz N, Sneiderman CA, Kilicoglu H, Menagerie: A text-mining tool to support animal-human translation in neurodegeneration research, PLoS One 14 (12) (2019) e0226176. [PubMed: 31846471]

[41]. Smalheiser NR, Luo M, Addepalli S, Cui X, A manual corpus of annotated main findings of clinical case reports, Database 2019 (2019).

[42]. Menke J, Roelandse M, Ozyurt B, Martone M, Bandrowski A, Rigor and transparency index, a new metric of quality for assessing biological and medical science methods, BioRxiv (2020), 10.1101/2020.01.15.908111.

[43]. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240. [PubMed: 31501885]

[44]. Passonneau R, Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006.

[45]. Krippendorff K, Content analysis: An Introduction to its Methodology, Sage Publications, Beverly Hills, CA, 1980.

[46]. Devlin J, Chang M-W, Lee K, Toutanova K, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, 10.18653/v1/N19-1423.

[47]. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[48]. Artstein R, Poesio M, Inter-coder agreement for computational linguistics, Comput. Linguist. 34 (4) (2008) 555–596.

[49]. Aronson AR, Lang F-M, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Informat. Assoc (JAMIA) 17 (3) (2010) 229–236.

[50]. Kilicoglu H, Hoang L, Wadhwa S, Identifying Sample Size Characteristics in Randomized Controlled Trial Publications, in: AMIA Annual Symposium Proceedings vol. 2020, American Medical Informatics Association, 2020.

[51]. Wei J, Zou K, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, in, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6383–6389.

[52]. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, Clancy K, Courtot M, Derom D, Dumontier M, et al., The Ontology for Biomedical Investigations, PloS One 11 (4) (2016).

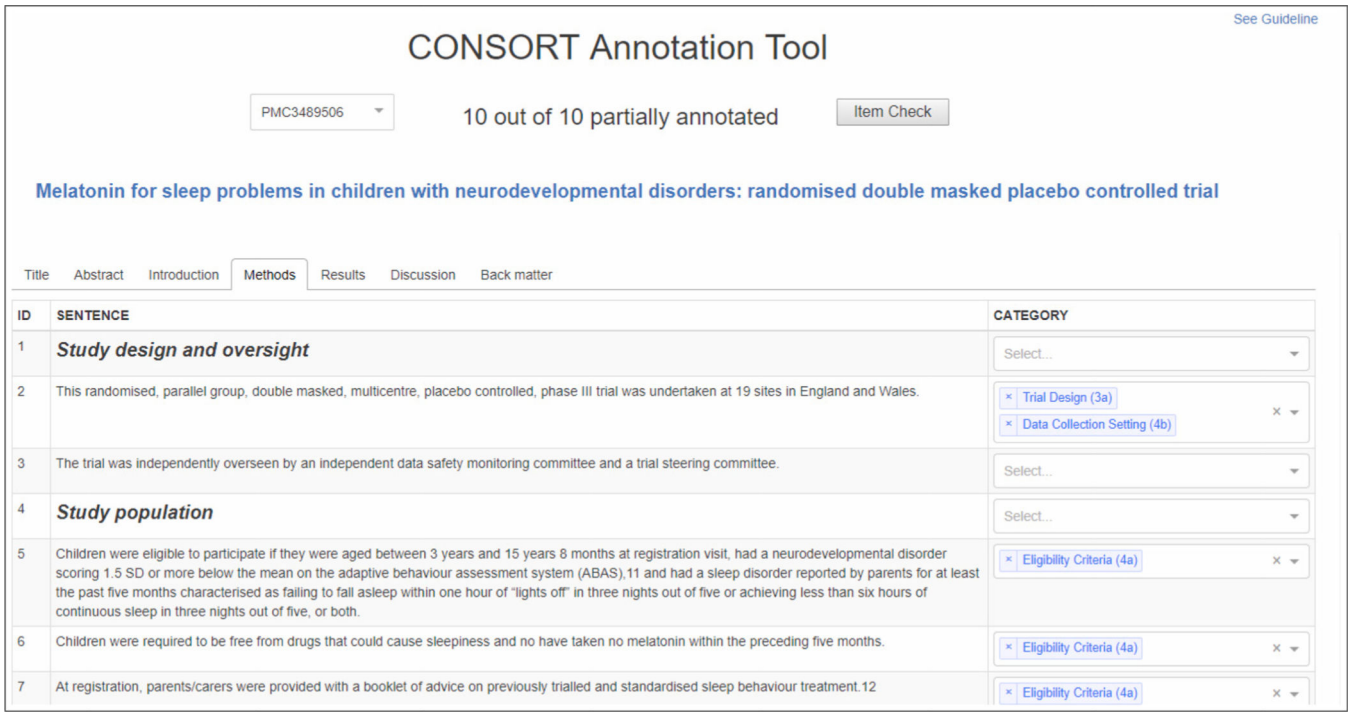

Fig. 1.
CONSORT-TM annotation interface.

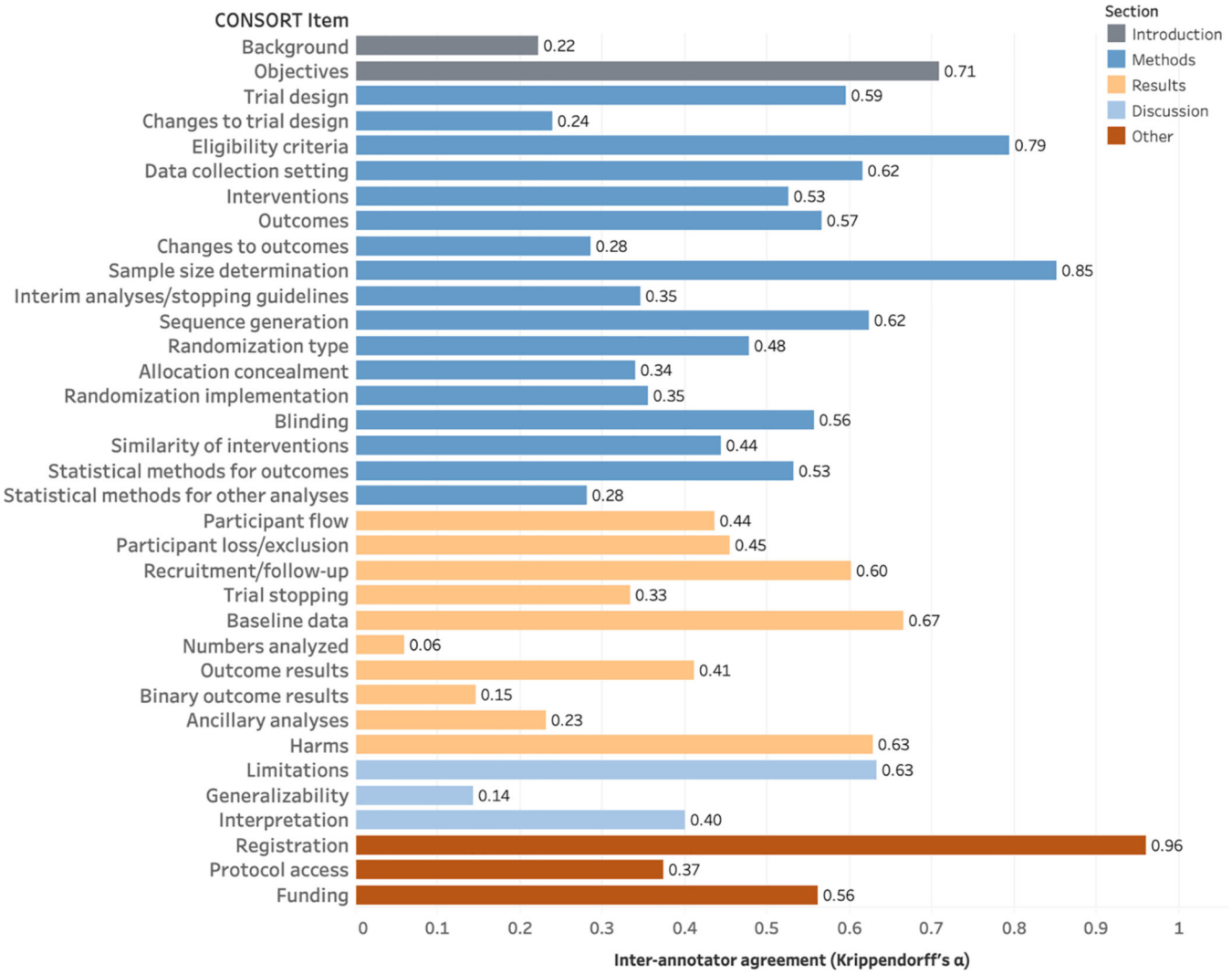

Fig. 2.

Inter-annotator agreement at the CONSORT item level, calculated using Krippendorff's α.Items are color coded by their associated sections, as shown in the legend.

Table 1

Descriptive statistics regarding 50 manually annotated RCT articles in CONSORT-TM.

	Total No.	Mean (±SD)	Median (IQR)
Sentences	10709	214.2 (±42.0)	208 (185.8–237.3)
Tokens	293977	5879.5 (±1544.8)	5197.5 (4864.3–6186.5)
Annotated sentences	4845	96.9 (±21.1)	92.5 (80–109.8)
Annotations	5246	104.9 (±22.3)	101 (92–119)
Checklist items per article		27.5 (±4.5)	28 (25–31)

Table 2

Descriptive statistics regarding the annotation of CONSORT checklist items in CONSORT-TM. SD: standard deviation.

CONSORT Item	No. of articles	Avg. number of sentences per article (±SD)	Range
Title Randomized (1a)	40	0.80 (±0.40)	0–1
Structured Abstract (1b)	49	0.98 (±0.14)	0–1
Background (2a)	50	2.04 (±0.20)	2–3
Objectives (2b)	50	1.44 (±0.61)	1–3
Trial Design (3a)	49	1.48 (±0.95)	0–5
Changes to Trial Design (3b)	4	0.20 (±0.70)	0–3
Eligibility Criteria (4a)	49	3.20 (±1.56)	0–8
Data Collection Setting (4b)	40	0.96 (±0.60)	0–2
Interventions (5)	50	5.40 (±3.17)	1–15
Outcomes (6a)	50	13.22 (±7.24)	3–38
Changes to Outcomes (6b)	5	0.12 (±0.39)	0–2
Sample Size Determination (7a)	46	2.26 (±1.65)	0–8
Interim Analyses/ Stopping Guidelines (7b)	11	0.42 (±1.01)	0–5
Sequence Generation (8a)	38	0.86 (±0.61)	0–3
Randomization Type (8b)	39	1.00 (±0.70)	0–3
Allocation Concealment (9)	19	0.44 (±0.64)	0–3
Randomization Implementation (10)	30	1.14 (±1.51)	0–8
Blinding (11a)	40	1.18 (±1.38)	0–9
Similarity of Interventions (11b)	15	0.36 (±0.60)	0–2
Statistical Methods for Outcomes (12a)	50	5.44 (±2.60)	1–13
Statistical Methods for Other Analyses (12b)	29	1.44 (±1.67)	0–6
Participant Flow (13a)	45	2.52 (±1.39)	0–6
Participant Loss/Exclusion (13b)	43	2.32 (±1.65)	0–7
Periods of Recruitment/ Follow-Up (14a)	42	1.04 (±0.67)	0–3
Trial Stopping (14b)	6	0.18 (±0.63)	0–4
Baseline Data (15)	50	3.80 (±2.24)	1–12
Numbers Analyzed (16)	47	2.06 (±1.43)	0–7
Outcome Results (17a)	50	14.38 (±8.89)	2–48
Binary Outcome Results (17b)	32	3.94 (±5.59)	0–26
Ancillary Analyses (18)	37	4.50 (±4.46)	0–16
Harms (19)	45	3.88 (±3.90)	0–17
Limitations (20)	44	4.16 (±2.97)	0–11
Generalizability (21)	29	0.88 (±0.98)	0–4
Interpretation (22)	50	13.00 (±6.74)	3–40
Registration (23)	45	1.20 (±0.61)	0–2
Protocol Access (24)	7	0.16 (±0.42)	0–2
Funding (25)	50	2.52 (±1.82)	1–10

Table 3

Inter-annotator agreement calculated by MASI formulation. SD: standard deviation; IQR: inter-quartile range.

	Mean MASI (±SD)	Median MASI (IQR)
Article	0.60 (±0.10)	0.63 (0.53–0.67)
Introduction	0.67 (±0.24)	0.75 (0.67–0.80)
Methods	0.59 (±0.12)	0.60 (0.53–0.66)
Results	0.50 (±0.16)	0.50 (0.44–0.59)
Discussion	0.58 (±0.15)	0.59 (0.48–0.71)
Other	0.89 (±0.14)	0.95 (0.81–1.00)
Annotator 1	0.64 (±0.07)	0.66 (0.58–0.68)
Annotator 2	0.59 (±0.10)	0.56 (0.53–0.67)
Annotator 3	0.62 (±0.10)	0.66 (0.61–0.67)
Annotator 4	0.60 (±0.10)	0.63 (0.53–0.65)
Annotator 5	0.58 (±0.12)	0.56 (0.52–0.66)

Table 4

Baseline experiment results. The results for phrase-based and section header-based methods were obtained from all 50 articles, and the results for linear SVM and BioBERT-based models were obtained using 5-fold cross-validation. Best results for each CONSORT item are in bold. 3a: Trial Design; 3b: Changes to Trial Design; 4a: Eligibility Criteria; 4b: Data Collection Setting; 5: Interventions; 6a: Outcomes; 6b: Changes to Outcomes; 7a: Sample Size Determination; 7b: Interim Analyses/ Stopping Guidelines; 8a: Sequence Generation; 8b: Randomization Type; 9: Allocation Concealment; 10: Randomization Implementation; 11a: Blinding Procedure; 11b: Similarity of Interventions; 12a: Statistical Methods for Outcomes; 12b: Statistical Methods for Other Analyses; Micro: Micro-averaging; Macro: Macro-averaging. AUC: Area Under Receiver Operator Characteristic (ROC) Curve. In the last column, each letter indicates that the results of one method is statistically significantly different from those of another method at 95% confidence level, as measured by McNemar's test (a: phrase-based vs. section header-based; b: phrase-based vs. linear SVM; c: phrase-based vs. BioBERT; d: section-header based vs. linear SVM; e: section-header based vs. BioBERT; f: linear SVM vs. BioBERT).

	Phrase-based			**Section header-based**			**Linear SVM**			**BioBERT**			**Statistical significance ($p < 0.05$)**
Item	**Prec.**	**Recall**	**F_1**	**Prec.**	**Recall**	**F_1**	**Prec.**	**Recall**	**F_1**	**Prec.**	**Recall**	**F_1**	
3a	0.71	0.55	0.62	0.14	0.75	0.23	0.70	0.58	0.62	**0.93**	0.49	0.63	abcdef
3b	**1.00**	**0.20**	**0.33**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
4a	0.62	0.68	0.65	0.40	**0.88**	0.54	0.87	0.60	0.70	0.90	0.82	**0.85**	abcef
4b	0.34	0.44	0.38	0.03	0.03	0.03	0.88	**0.46**	**0.59**	**0.80**	0.24	0.36	ef
5	0.48	0.49	0.48	0.66	0.35	0.46	0.66	0.50	0.56	**0.76**	**0.69**	**0.72**	abcdef
6a	0.57	0.34	0.42	**0.86**	0.48	0.62	0.74	0.64	0.69	0.84	**0.78**	**0.81**	abcde
6b	**0.50**	**0.17**	**0.25**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
7a	0.85	0.65	0.74	0.51	0.39	0.44	**0.93**	0.70	0.79	0.88	**0.80**	**0.84**	abcdef
7b	0.18	**0.88**	0.29	0.00	0.00	0.00	**0.80**	0.64	**0.70**	0.00	0.00	0.00	abcdf
8a	0.71	0.63	0.67	0.00	0.00	0.00	**0.92**	**0.64**	**0.74**	0.86	0.26	0.38	abcdef
8b	0.55	**0.71**	**0.62**	0.15	0.53	0.23	0.67	0.46	0.54	**0.71**	0.29	0.38	abcdef
9	**0.86**	**0.27**	**0.41**	0.00	0.00	0.00	0.28	0.19	0.22	0.00	0.00	0.00	bdf
10	0.63	0.18	0.27	0.24	**0.72**	0.35	0.68	0.25	**0.36**	**0.72**	0.15	0.24	abcde
11a	0.53	**0.58**	0.55	0.21	0.28	0.24	**0.84**	0.45	**0.58**	0.77	0.29	0.42	abcef
11b	**0.36**	**0.56**	**0.43**	0.00	0.00	0.00	0.20	0.13	0.16	0.00	0.00	0.00	
12a	0.52	0.52	0.52	0.38	**0.99**	0.55	0.72	0.64	0.67	**0.75**	0.76	**0.75**	abcf
12b	**0.35**	**0.24**	**0.28**	0.00	0.00	0.00	0.32	0.13	0.17	0.05	0.03	0.04	adf
Micro	0.54	0.46	0.50	0.40	0.52	0.45	0.74	0.56	0.64	**0.82**	**0.63**	**0.72**	
Macro	0.57	**0.47**	0.47	0.21	0.32	0.22	**0.60**	0.41	**0.48**	0.52	0.33	0.38	

	Phrase-based			Section header-based			Linear SVM			BioBERT			Statistical significance ($p < 0.05$)
Item	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	
AUC	-			-			0.774			**0.812**			

Table 5

Results of combining four base models. The results for the base models are also provided for comparison. Best base model performances as well as best combination performances are in bold. PHR: phrase-based method; SECT: section header-based method; SVM: linear SVM model; BERT: BioBERT-based model.

Combination	Micro-averaging			Macro-averaging		
	Prec.	Recall	F_1	Prec.	Recall	F_1
PHR	0.54	0.46	0.50	0.57	**0.47**	0.47
SECT	0.40	0.52	0.45	0.21	0.32	0.22
SVM	0.74	0.56	0.64	**0.60**	0.41	**0.48**
BERT	**0.82**	**0.63**	**0.72**	0.52	0.33	0.38
Majority Vote	**0.78**	0.68	0.73	**0.74**	0.46	0.52
PHR + SEC	0.40	0.72	0.51	0.45	0.63	0.43
PHR + SVM	0.57	0.71	0.63	0.57	0.60	0.54
PHR + BERT	0.61	0.76	0.68	0.59	0.59	0.54
SEC + SVM	0.45	0.74	0.56	0.46	0.54	0.42
SEC + BERT	0.47	0.75	0.58	0.30	0.45	0.31
SVM + BERT	0.73	0.74	**0.74**	0.67	0.50	0.54
PHR + SEC + SVM	0.41	0.82	0.55	0.44	0.69	0.46
PHR + SEC + BERT	0.42	0.84	0.56	0.45	0.67	0.45
PHR + SVM + BERT	0.58	0.81	0.68	0.57	0.65	**0.56**
SEC + SVM + BERT	0.46	0.82	0.59	0.47	0.58	0.44
PHR + SEC + SVM + BERT	0.42	**0.87**	0.56	0.44	**0.71**	0.47