Corpus Creation and Detection of CONSORT Elements in Randomized Controlled Trials

Joshua Fisher, MSc; Anita de Waard, MSc; Conor O'Donnell, MSc
Research Collaborations Unit, Elsevier

Introduction

Elsevier's leading medical journal The Lancet requires authors to submit a completed CONSORT checklist upon submission of their randomized control trial (RCT) focused manuscript.

The current workflow consists of authors submitting this checklist, indicating what page number each element can be found. Editors then check these assertions for validity and report back to the author.

While checklists aim to simplify the structuring of clinical trials, they create an unintentional burden on authors and editors.

We describe work on a tool to automatically support the identification of elements of the CONSORT guidelines.

The tool aims to support Lancet editors in assessing compliance with these guidelines.

Objectives

The objectives of this project is two-fold:

Creation of suitable dataset, consisting of RCT's submitted to The Lancet and annotated with CONSORT checklist elements

Development of NLP based classification model to identify CONSORT elements within unseen RCT manuscripts

Agreement Analysis

Checklist	CONSORT Element	Average Agreement
Abstracts Average 0.720	Funding	1.000
	Registration	0.939
	Objective	0.835
	Conclusions	0.818
	Trial Design	0.710
	Setting	0.682
	Eligibility criteria	0.680
	Results Outcome	0.673
	Randomization	0.657
	Blinding	0.544
	Harms	0.541
	Interventions	0.502
	Numbers Analyzed	0.473
	Recruitment	0.464
	Methods Outcome	0.441
	Numbers Randomized	0.425
Methods Average 0.549	Changes to Trial Design (3b)	0.787
	Statistical Methods for Outcomes (12a)	0.645
	Who was Blinded and How (11a)	0.629
	Interim Analysis and Stopping Guidelines (7b)	0.589
	Methods for Additional Analysis (12b)	0.576
	Eligibility Criteria (4a)	0.554
	Similarity of Interventions (11b)	0.553
	How Sample Size was Determined (7a)	0.519
	Type of Randomization (8b)	0.513
	Implementation of Random Allocation Sequence (9)	0.496
	Trial Design (3a)	0.462
	Settings and Locations (4b)	0.459
	Interventions by Group (5)	0.459
	Who Generated Random Allocation Sequence (10)	0.428
	Primary & Secondary Outcomes (6a)	0.323
	Generation of Random Allocation Sequence (8a)	0.283
	Changes to Trial Outcomes (6b)	0.226

Materials & Methods

Dataset Annotation

- 168 abstracts and 368 methods sections were extracted from RCT's submitted to The Lancet between 2010 and the present
- Three subject matter experts were recruited to label the full span of each sentence with corresponding CONSORT elements using the Brat annotating tool
- To test inter-annotator agreement, Measuring Agreement on Set-Valued Items (MASI) was used to calculate distance between sets of tags (CONSORT elements) assigned to a sentence

Modeling

- Multi-label classification problem
- Various models including Multilabel k Nearest Neighbors (MLkNN), Support Vector Classifier (SVC), and SciBERT were trained and compared
- Model performance was calculated using precision, recall, and F1 score

Results

- The annotators had substantial agreement (0.72) in the abstract set and moderate agreement in the methods data (0.549)
- The fine-tuned SciBERT model was most reliable in predicting most CONSORT elements
- Elements corresponding to numbers analyzed, changes in trials design and outcomes were most difficult to accurately identify

Conclusions

- We have created a robust dataset of annotated sentences that capture critical information described in RCTs
- Using SciBERT, we were able to identify these elements automatically with reasonable precision and recall
- Further testing is required to improve items with lower scores
- This tool has potential to reduce editorial effort, improve guideline compliance, and reduce reviewing time
- Potential to extend to other guidelines such as ARRIVE or PRISMA

References

- Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials
- Hopewell, S. et al. (2008) "Consort for reporting randomised trials in journal and conference abstracts," The Lancet, 371(9609), pp. 281–283. Available at: https://doi.org/10.1016/s0140-6736(07)61835-2.

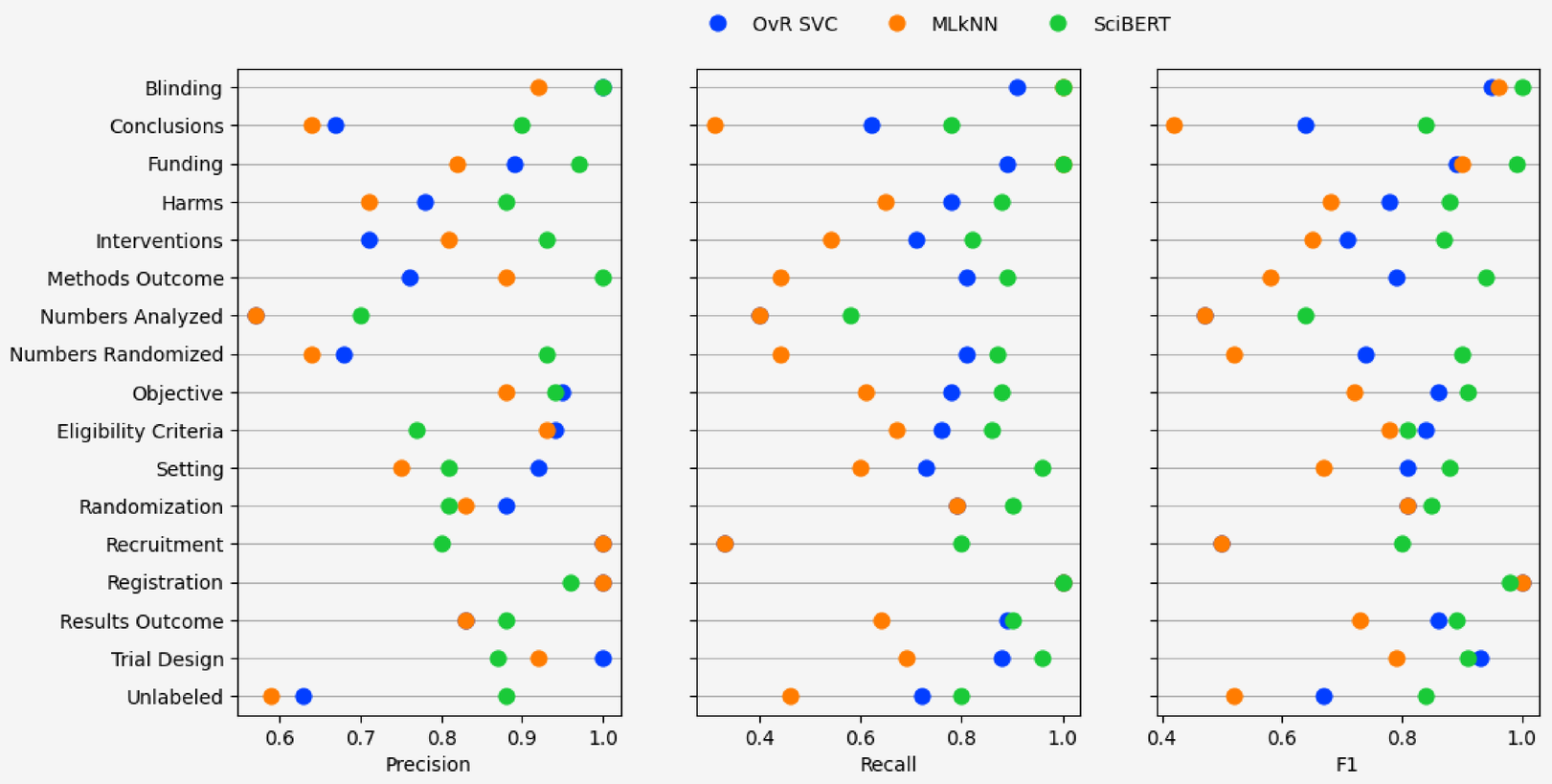
Contact Info

J.fisher@elsevier.com

Data Scientist

Research Collaborations Unit, Elsevier

Abstracts



Methods Sections