Annotating Clinical Trial Publications to Assess CONSORT Adherence: A Feasibility Study

Halil Kilicoglu[1], Graciela Rosemblat[1], Zeshan Peng[1], Mario Malički[2,3], Jodi Schneider[4], Gerben ter Riet[2,3]

[1]U.S. National Library of Medicine

[2]Amsterdam University Medical Center

[3]Amsterdam University of Applied Sciences

[4]University of Illinois at Urbana-Champaign

Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.

Reporting Guidelines

- Promote transparent, complete and accurate reporting
- EQUATOR Network
  - CONSORT, ARRIVE, STROBE, PRISMA
- Improve reporting quality
  - May be easier to reproduce
- Adherence remains inadequate

CONSORT Statement

- **CON**solidated **S**tandards **O**f **R**eporting **T**rials
- Guidelines for parallel group randomized controlled trials
- 25-item checklist and flow diagram
- Endorsed by over 600 journals
  - Lancet, BMJ, NEJM, etc.
- Extensions
  - Abstracts
  - Cluster randomized trials
  - Non-inferiority or equivalence trials

CONSORT Checklist Examples

Checklist Item	Category	Example Sentence
Objective (2b)	Introduction	*We studied the effects of metformin in obese children aged 6–12 years who were believed to be at particular risk because they manifested a significant degree of insulin resistance.*
Allocation concealment (9)	Methods	*The pharmacy produced identical, sequentially numbered, randomly assigned boxes of study medication, containing either magnesium sulphate or placebo.*
Outcome results (17a)	Results	*No difference between bosentan and placebo treatments was observed in the time to healing of the cardinal ulcer (HR 0.91 (95% CI 0.61 to 1.35), p=0.63, figure 3).*
Limitations (20)	Discussion	*The main limitation of our trial is the small sample size of patients with bacteraemia, in whom results suggest an important advantage for vancomycin.*
Protocol access (24)	Other	*The trial protocol has been published previously.*[11]

Automating Adherence Assessment

- Text-mining techniques
  - Locate key statements for checklist items in a manuscript/publication
  - Give alerts in their absence
- Benefits for journal editors, peer reviewers, authors, systematic reviewers
- Commercial/academic software for some items
  - Penelope.ai, StatReviewer, RobotReviewer, ExaCT

Automating Adherence Assessment

- Text-mining techniques
  - Locate key statements for checklist items in a manuscript/publication
  - Give alerts in their absence
- Benefits for journal editors, peer reviewers, authors, systematic reviewers
- Commercial/academic software for some items
  - Penelope.ai, StatReviewer, RobotReviewer, ExaCT

- Labeled data needed to train and evaluate text-mining tools

Objective

- Annotate sentences from RCT articles with the relevant CONSORT checklist items
- Develop baseline text-mining methods to automatically recognize these items

Article Selection

- Cochrane RCT search strategy maximizing sensitivity and precision
  - Exclude meta-analyses, systematic reviews
  - 2011 to present
  - 11 journals (9 CONSORT-endorsing)
- 563 articles retrieved
- 50 articles selected

Annotation

- Sentence-level, multi-label annotation
  - 25 checklist items ⟶ 37 fine-grained categories
- 6 annotators
  - Experts in text mining/informatics, linguistics, meta-research, and clinical trials
- 50 articles annotated
  - 1 exploratory annotation
  - 30 double-annotated and adjudicated
  - 19 single-annotated and corrected
- Annotation instructions provided
- Web-based annotation/adjudication tool used

Corpus Statistics

- 50 articles, 10779 sentences

	Total	Mean (Range)	Median (IQR)
Annotations	5679	113.6 (66-197)	110.5 (93.8-126.5)
Annotated sentences	4845	96.9 (61-158)	92.5 (80.0-109.8)
Items per article		27.5 (15-35)	28 (25-31)

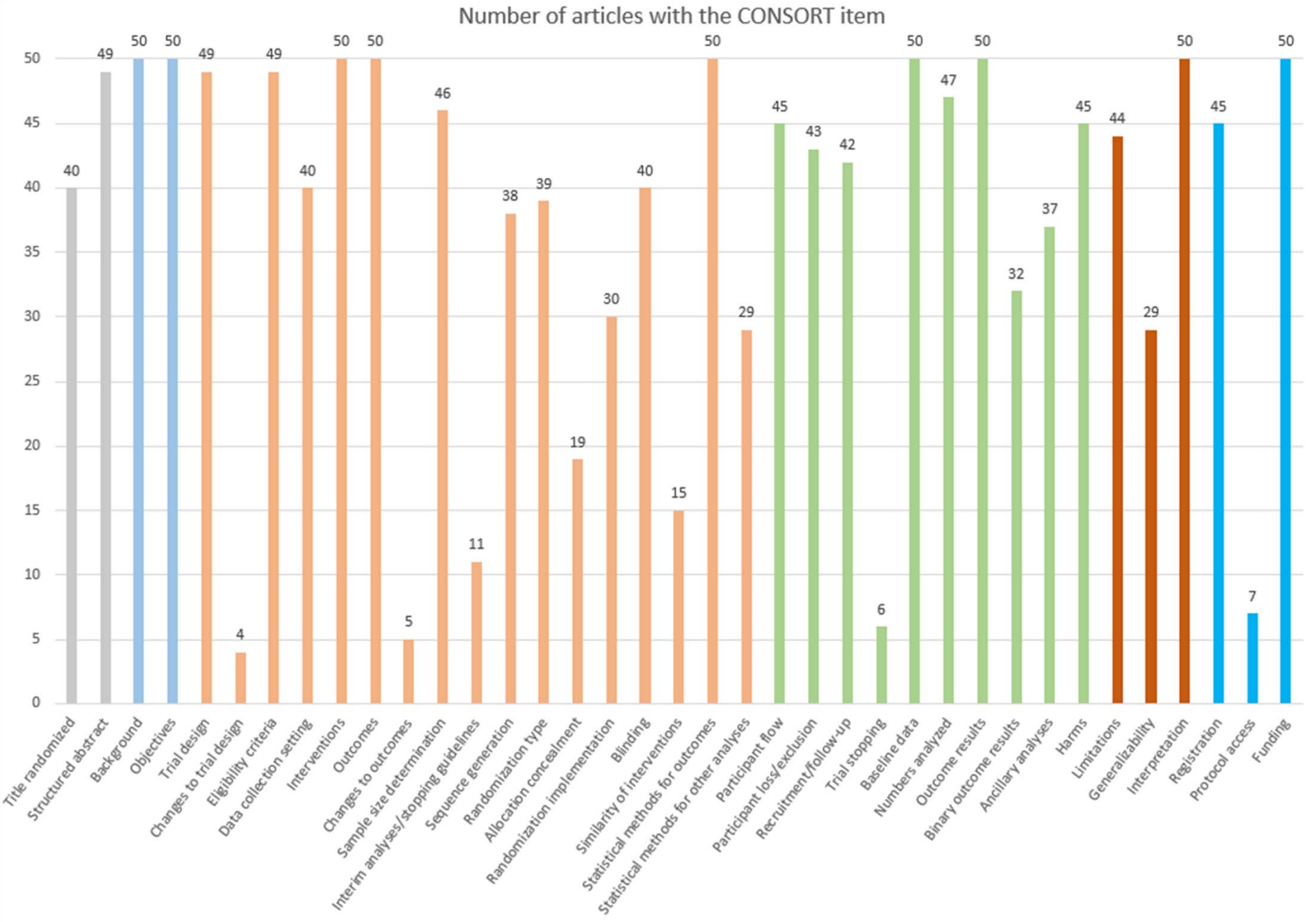
Corpus Statistics

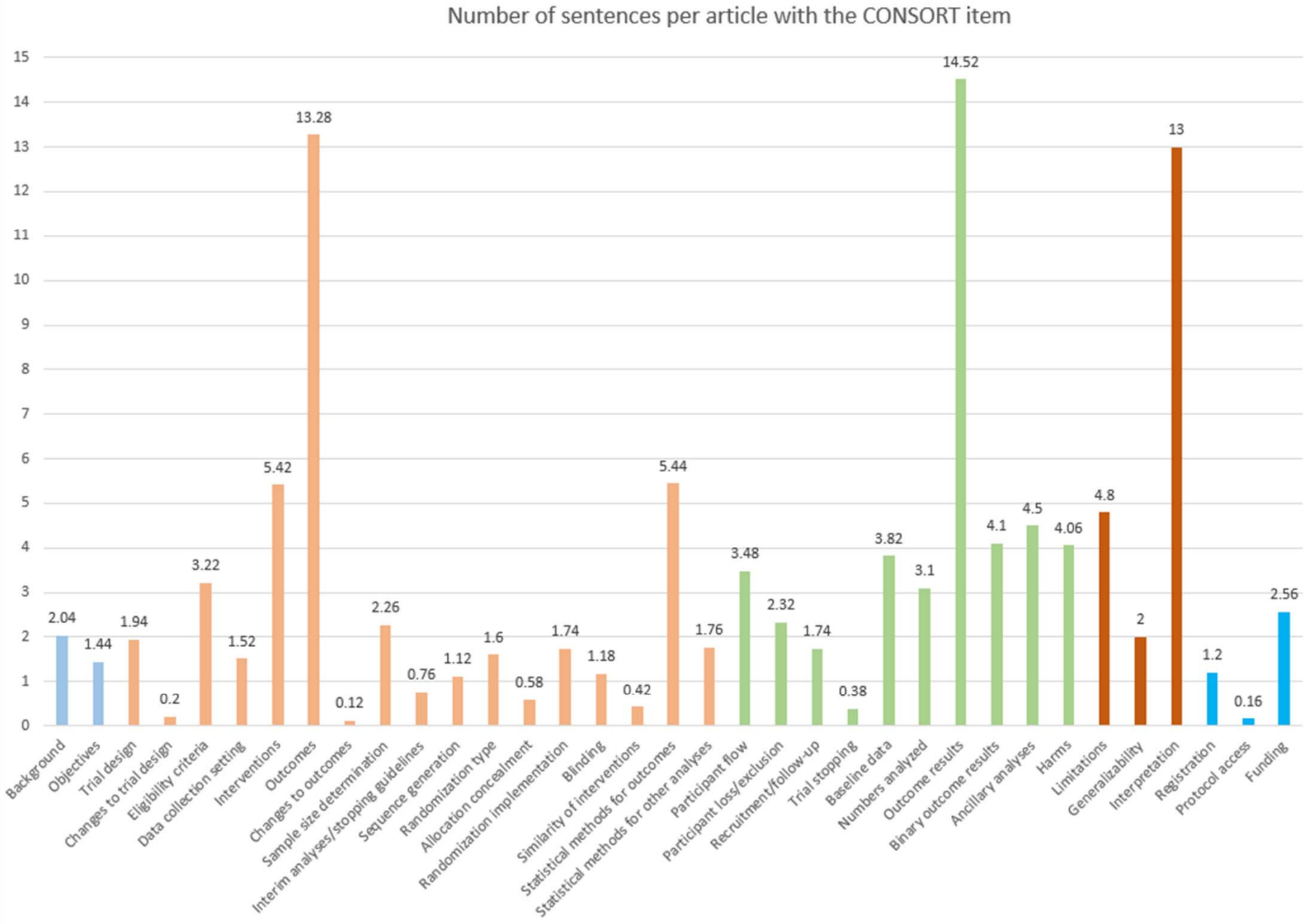
- 50 articles, 10779 sentences

	Total	Mean (Range)	Median (IQR)
Annotations	5679	113.6 (66-197)	110.5 (93.8-126.5)
Annotated sentences	4845	96.9 (61-158)	92.5 (80.0-109.8)
Items per article		27.5 (15-35)	28 (25-31)

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 (800mg ibuprofen and 26.6mg famotidine) or ibuprofen (800mg) thrice daily for 24 weeks.

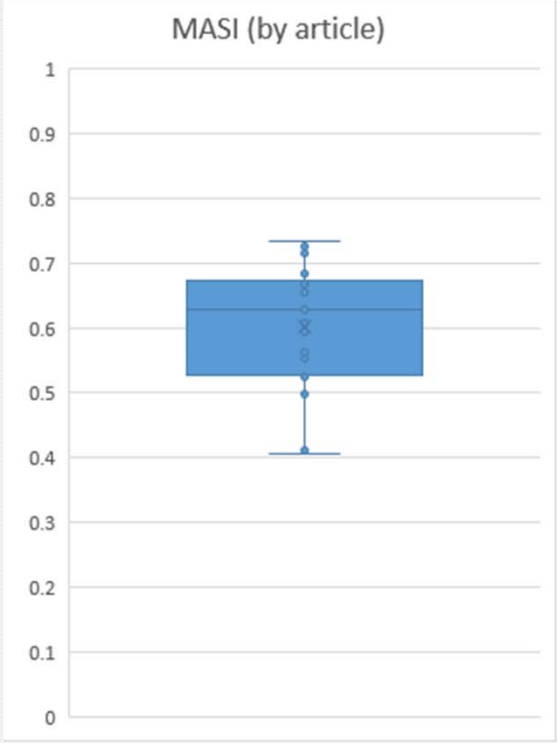
- Trial design, Sequence generation, Allocation concealment, Randomization implementation, Similarity of interventions

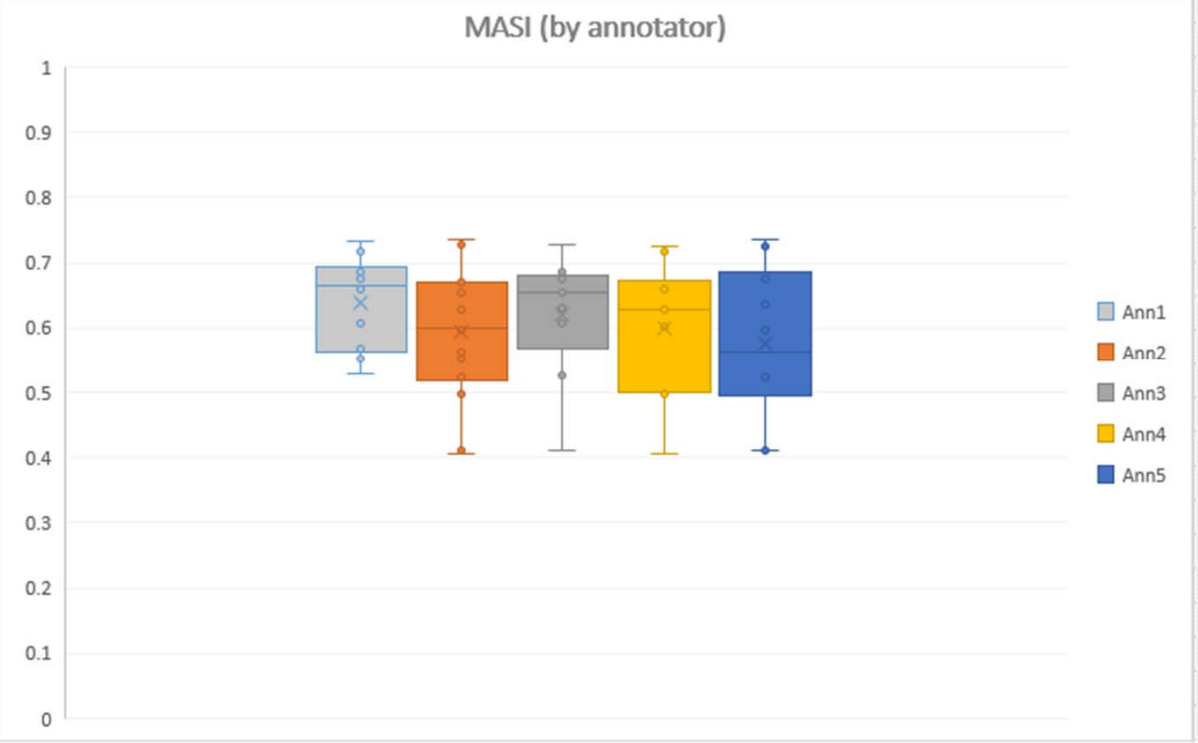

Number of articles with the CONSORT item
0
5
10
15
20
25
30
35
40
45
50
40
49
50
50
49
4
49
40
50
50
5
46
11
38
39
19
30
40
15
50
29
45
43
42
6
50
47
50
32
37
45
44
29
50
45
7
50
Title randomized
Structured abstract
Background
Objectives
Trial design
Changes to trial design
Eligibility criteria
Data collection setting
Interventions
Outcomes
Changes to outcomes
Sample size determination
Interim analyses/stopping guidelines
Sequence generation
Randomization type
Allocation concealment
Randomization implementation
Blinding
Similarity of interventions
Statistical methods for outcomes
Statistical methods for other analyses
Participant flow
Participant loss/exclusion
Recruitment/follow-up
Trial stopping
Baseline data
Numbers analyzed
Outcome results
Binary outcome results
Ancillary analyses
Harms
Limitations
Generalizability
Interpretation
Registration
Protocol access
Funding

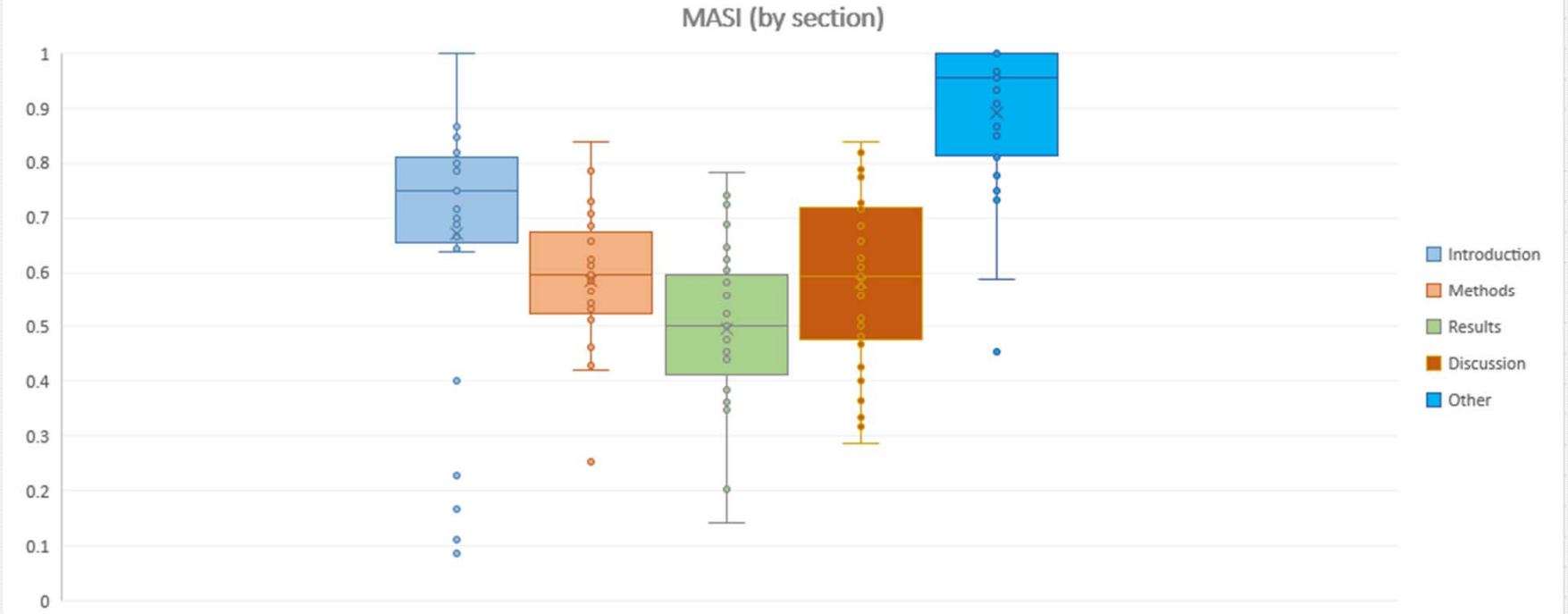

Number of sentences per article with the CONSORT item
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
2.04
1.44
1.94
0.2
3.22
1.52
5.42
13.28
0.12
2.26
0.76
1.12
1.6
0.58
1.74
1.18
0.42
5.44
1.76
3.48
2.32
1.74
0.38
3.82
3.1
14.52
4.1
4.5
4.06
4.8
2
13
1.2
0.16
2.56
Background
Objectives
Trial design
Changes to trial design
Eligibility criteria
Data collection setting
Interventions
Outcomes
Changes to outcomes
Sample size determination
Interim analyses/stopping guidelines
Sequence generation
Randomization type
Allocation concealment
Randomization implementation
Blinding
Similarity of interventions
Statistical methods for outcomes
Statistical methods for other analyses
Participant flow
Participant loss/exclusion
Recruitment/follow-up
Trial stopping
Baseline data
Numbers analyzed
Outcome results
Binary outcome results
Ancillary analyses
Harms
Limitations
Generalizability
Interpretation
Registration
Protocol access
Funding

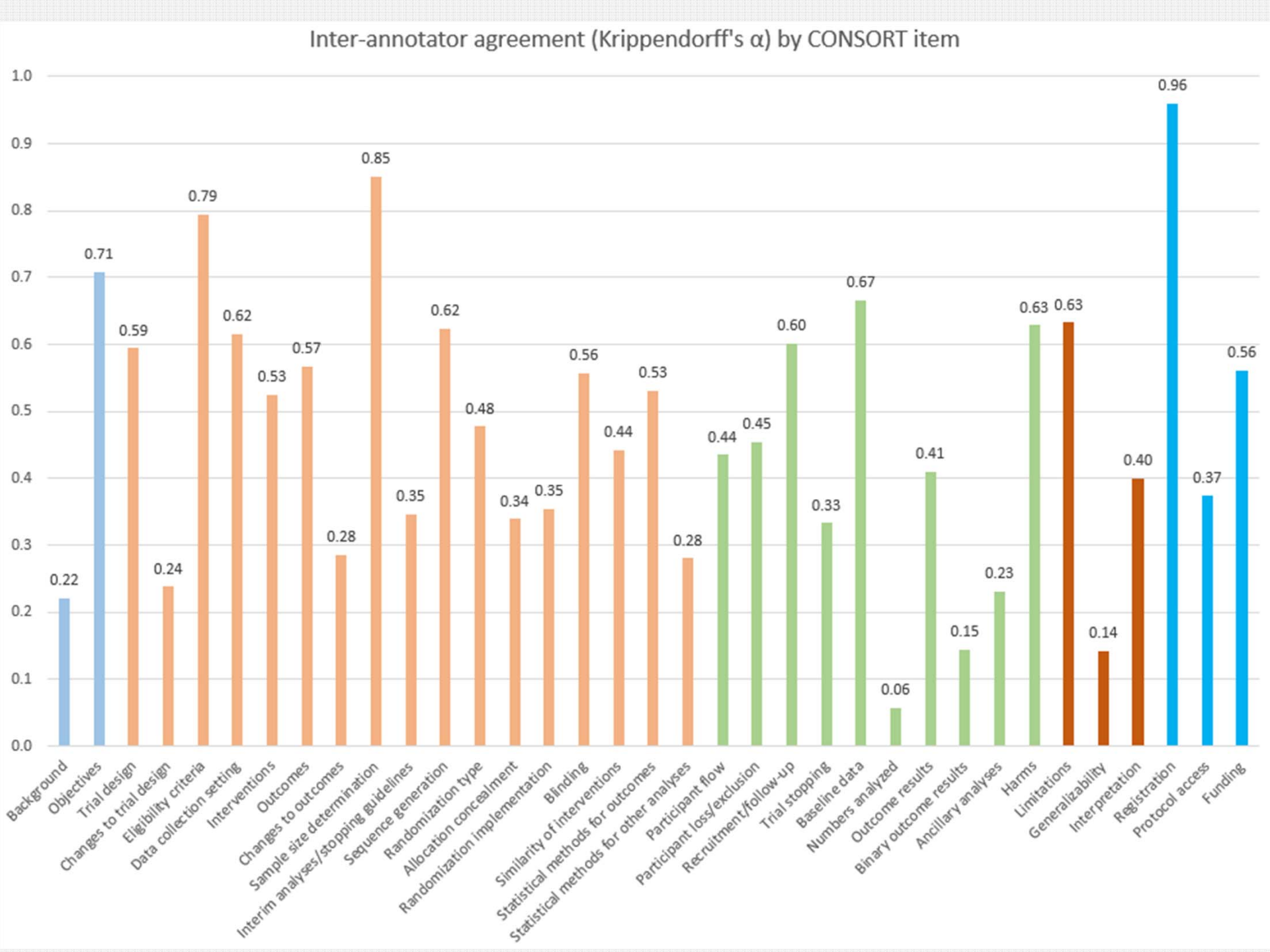
Inter-annotator agreement

- Measured over 30 double-annotated articles
- MASI (Measuring Agreement on Set-Valued Items)
  - Range [0,1]
  - Combines Jaccard index and higher penalty for disjoint items
  - Agreement at the article and section level
- Krippendorff's α
  - Range [0,1]
  - Agreement at the CONSORT item level
- Excluded from agreement calculation
  - Titles, section/subsection headers, authors' contributions, etc.


MASI (by article)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0


MASI (by annotator)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Ann1
Ann2
Ann3
Ann4
Ann5


MASI (by section)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Introduction
Methods
Results
Discussion
Other

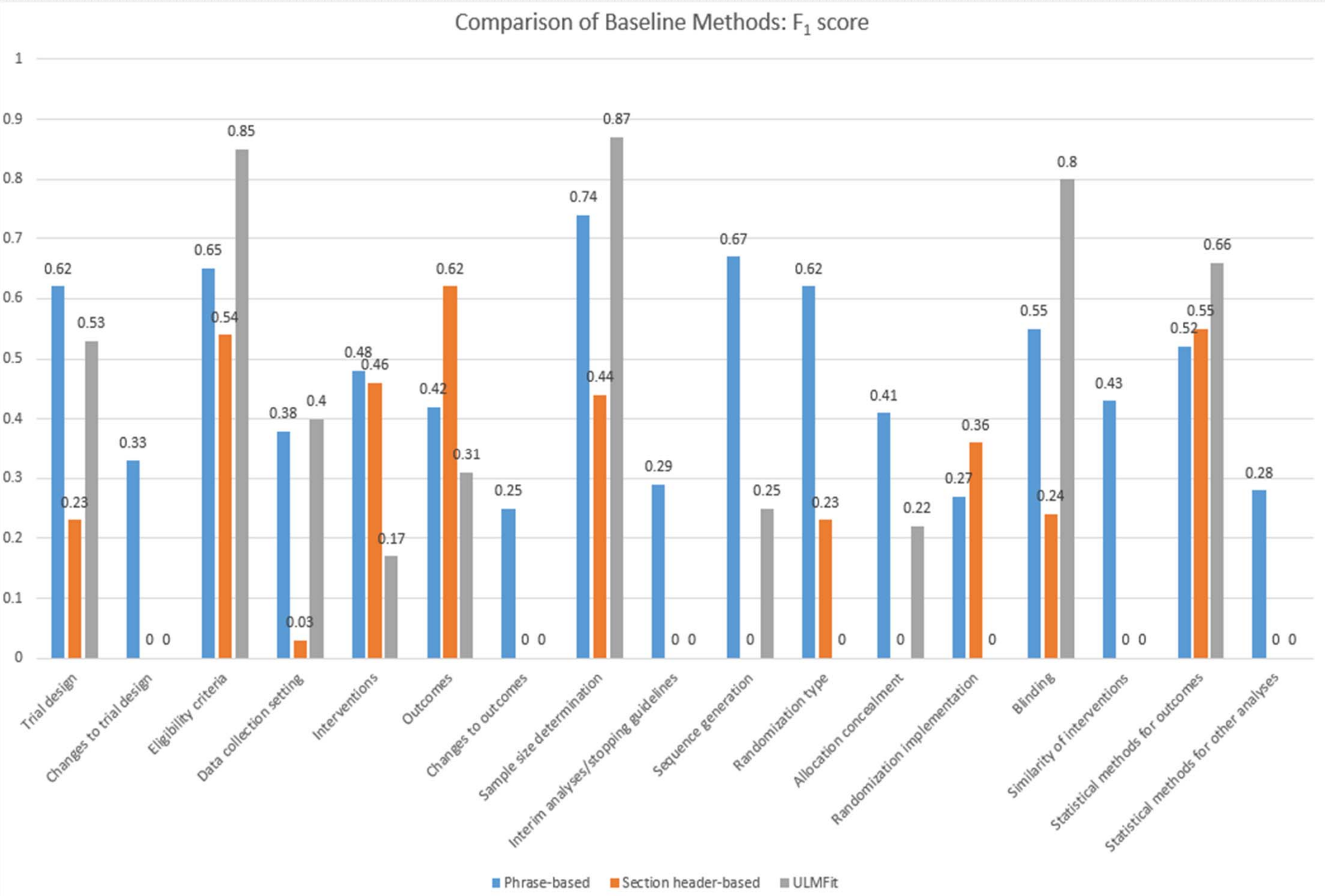

Inter-annotator agreement (Krippendorff's α) by CONSORT item
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.22
0.71
0.59
0.24
0.79
0.62
0.53
0.57
0.28
0.85
0.35
0.62
0.48
0.34
0.35
0.56
0.44
0.53
0.28
0.44
0.45
0.60
0.33
0.67
0.06
0.41
0.15
0.23
0.63
0.63
0.14
0.40
0.96
0.37
0.56
Background
Objectives
Trial design
Changes to trial design
Eligibility criteria
Data collection setting
Interventions
Outcomes
Changes to outcomes
Sample size determination
Interim analyses/stopping guidelines
Sequence generation
Randomization type
Allocation concealment
Randomization implementation
Blinding
Similarity of interventions
Statistical methods for outcomes
Statistical methods for other analyses
Participant flow
Participant loss/exclusion
Recruitment/follow-up
Trial stopping
Baseline data
Numbers analyzed
Outcome results
Binary outcome results
Ancillary analyses
Harms
Limitations
Generalizability
Interpretation
Registration
Protocol access
Funding

Baseline Classification Experiments

- Applied to Methods sections and Methods-specific items
- Automatic analysis of frequent section headers and phrases
- Section header-based classification
  - *"change" … "plan"* ⟶ Changes to trial design
- Phrase-based classification
  - *"masked to treatment"* ⟶ Blinding procedure
- ULMFit (Universal Language Model Fine-Tuning)
  - Deep neural network-based method
  - Training data (with some noise) automatically generated with section header heuristics

Evaluation Results

- Item-level evaluation
  - Macro-precision (p), macro-recall (r), macro-F1 (f)
  - Phrase-based (p: 0.57, r: 0.47, f: 0.47)
  - Section header-based (p: 0.21, r: 0.32, f: 0.22)
  - ULMFit (p: 0.40, r: 0.28, f: 0.30)
- Article-level evaluation
  - CONSORT item present in the article or not?
  - Phrase-based (p: 0.88, r: 0.79, f: 0.84)


Comparison of Baseline Methods: F_1 score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.62
0.23
0.53
0.33
0 0
0.65
0.54
0.85
0.38
0.03
0.4
0.48
0.46
0.17
0.42
0.62
0.31
0.25
0 0
0.74
0.44
0.87
0.29
0 0
0.67
0
0.25
0.62
0.23
0
0.41
0
0.22
0.27
0.36
0
0.55
0.24
0.8
0.43
0 0
0.52
0.55
0.66
0.28
0 0
Trial design
Changes to trial design
Eligibility criteria
Data collection setting
Interventions
Outcomes
Changes to outcomes
Sample size determination
Interim analyses/stopping guidelines
Sequence generation
Randomization type
Allocation concealment
Randomization implementation
Blinding
Similarity of interventions
Statistical methods for outcomes
Statistical methods for other analyses
Phrase-based
Section header-based
ULMFit

Conclusion

- Cognitively challenging annotation task
  - Large number of fine-grained categories (37)
- Inter-annotator agreement varied significantly for items (α range: 0.06-0.96)
  - Broad (Background, Interpretation)
  - Similar (Outcome result, Binary outcome result, Ancillary analysis)
- The manually annotated corpus can be used as a benchmark
- Phrase-based baseline method yields moderate results