ILLINOIS
School of Information Sciences

Natural language processing to promote transparency of clinical publications

Halil Kılıçoğlu
University of Illinois at Urbana-Champaign

Agenda

- Natural language processing for enhancing research transparency
- Recent work
  - Assessing clinical trial publications for reporting guideline adherence
  - Supporting meta-research investigations with natural language processing

Agenda

- Natural language processing for enhancing research transparency
- Recent work
  - Assessing clinical trial publications for reporting guideline adherence
  - Supporting meta-research investigations with natural language processing

Kilicoglu, H. “Biomedical text mining for research rigor and integrity: tasks, challenges, directions.” *Briefings in Bioinformatics*, 2018. 19(6): 1400-1414.

“Reproducibility Crisis”

- Causes
  - Poor experimental design and oversight
  - Publication bias for positive, statistically significant results
  - Novelty over reproducibility
  - “Publish or perish”
- Interventions
  - Standards and guidelines for reporting, data and code sharing
  - Cyberinfrastructure support for reproducibility

Research Transparency

The reporting of experimental materials and methods in a manner that provides enough information for others to independently assess and/or reproduce experimental findings

Antin PB, Baldwin TO, Freeze HH, Haywood JR, Simon SI. Enhancing research reproducibility: recommendations from the Federation of American Societies for Experimental Biology. FASEB. 2016.

Disentangling the Terminology

Rigor[1]	Robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results
Integrity[2]	Honest/verifiable methods, accurate reporting of results with adherence to rules, guidelines, and professional norms

[1] https://grants.nih.gov/reproducibility/index.htm
[2] https://grants.nih.gov/grants/research_integrity/whatis.htm

Disentangling the Terminology

Reproducibility	Same methods and data as in the original research
Replicability	Same methods as in the original research applied to newly collected data
Translatability	Different experimental design and data to reach the same conclusion

National Academies of Sciences, Engineering, and Medicine. Reproducibility and replicability in science. National Academies Press; 2019.

Biomedical Language Processing (bioNLP)

- Transform text into computable representations
  - Scientific publications, clinical notes, drug labels, etc.
  - Support knowledge discovery and clinical decision making
- Tools/techniques adapted from open-domain NLP
- Knowledge representation/ontologies, corpus annotation
- Tasks
  - Text classification, entity recognition, acronym/abbreviation resolution, relation extraction

BioNLP and Research Transparency

- Textual artifacts are core to biomedical communication lifecycle
  - Manual analysis is time-consuming
- NLP provides scalability
  - Scrutinize reports of conducted research
  - Manage published literature more effectively to improve quality of proposed research
- It can complement efforts in standardization and guideline development

Agenda

- Natural language processing for enhancing research rigor, integrity, and transparency
- Recent work
  - Assessing clinical trial publications for reporting guideline adherence
  - Supporting meta-research investigations with natural language processing

Kilicoglu, H, Rosemblat G, Peng Z, Malički M, Schneider J, ter Riet, G. "Annotating Clinical Trial Publications to Assess CONSORT Adherence: A Feasibility Study." *World Conference on Research Integrity (WCRI 2019).*

Study Summary

- Goal
  - Develop text-mining methods to automatically recognize the CONSORT checklist items in randomized controlled trial reports (RCTs)
- Approach
  - An annotation study
  - Comparison of baseline rule-based and weakly supervised machine learning methods

Reporting Guidelines

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

CONSORT Statement

- **CON**solidated **S**tandards **O**f **R**eporting **T**rials
- Reporting guidelines for parallel group RCTs
- 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	Section	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

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 sampled
  - PubMed Central XML download
  - Sentence splitting
  - Section extraction

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

Corpus Statistics

- 50 articles, 10779 sentences

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

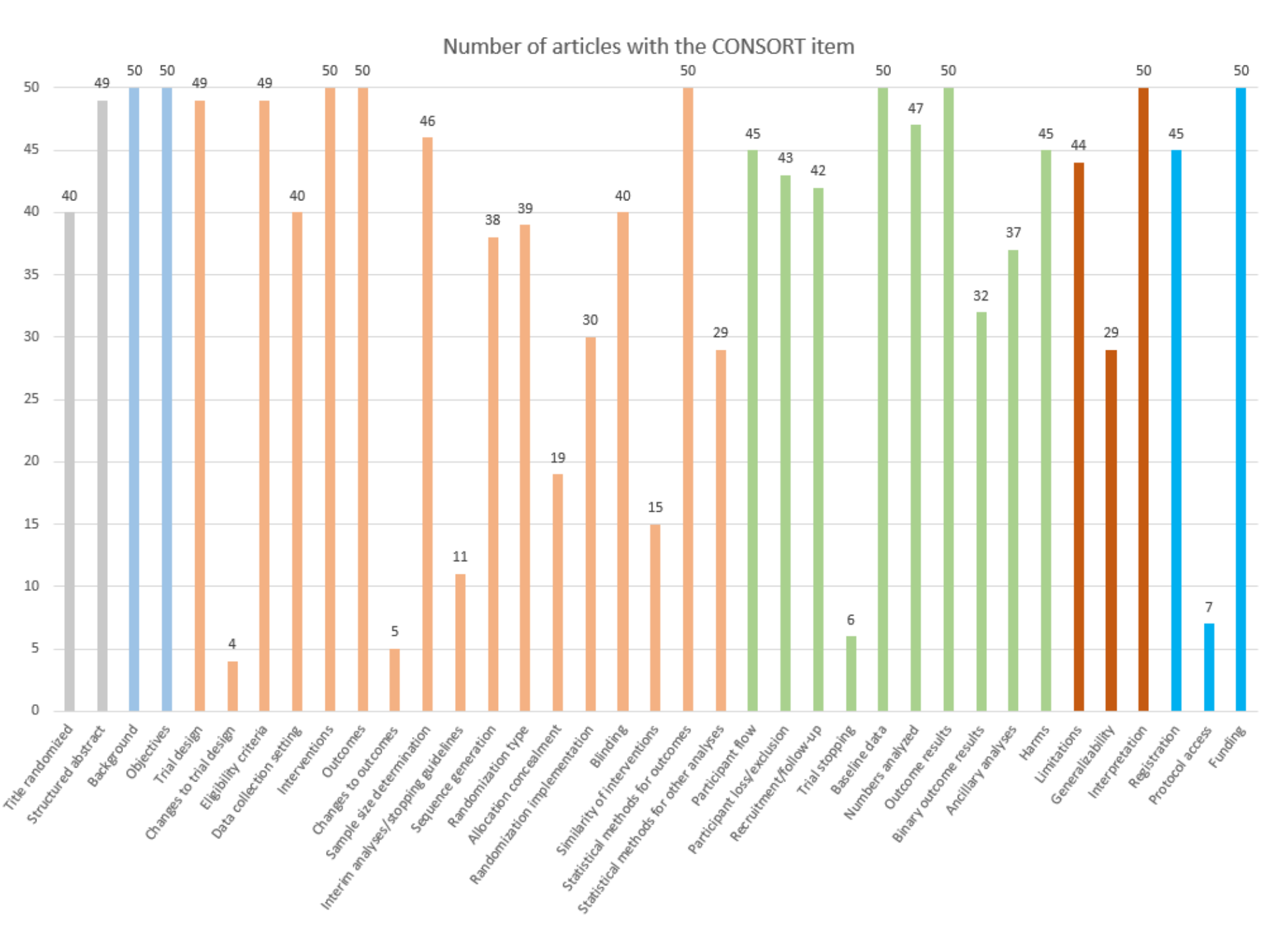
Corpus Statistics

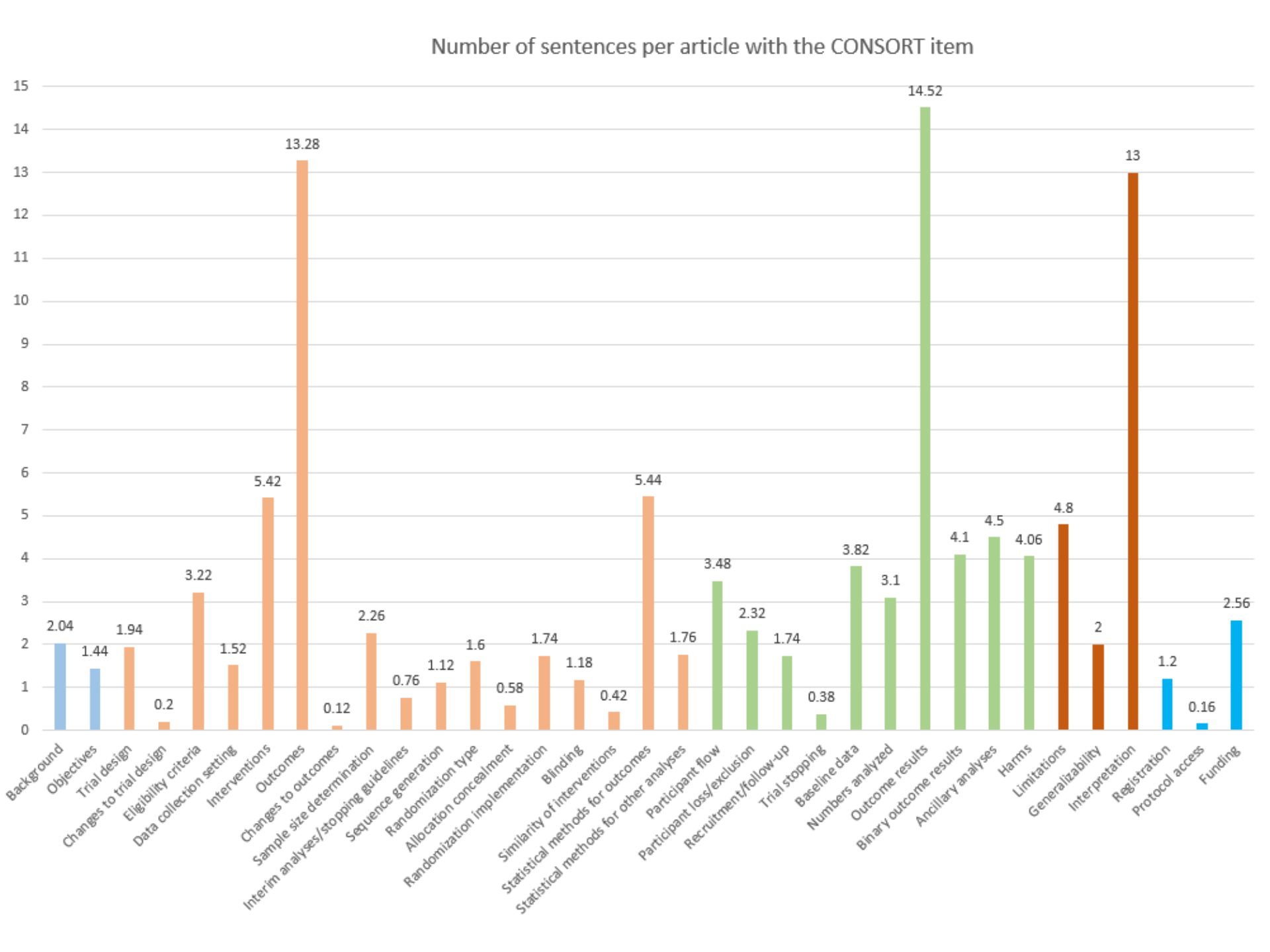
- 50 articles, 10779 sentences

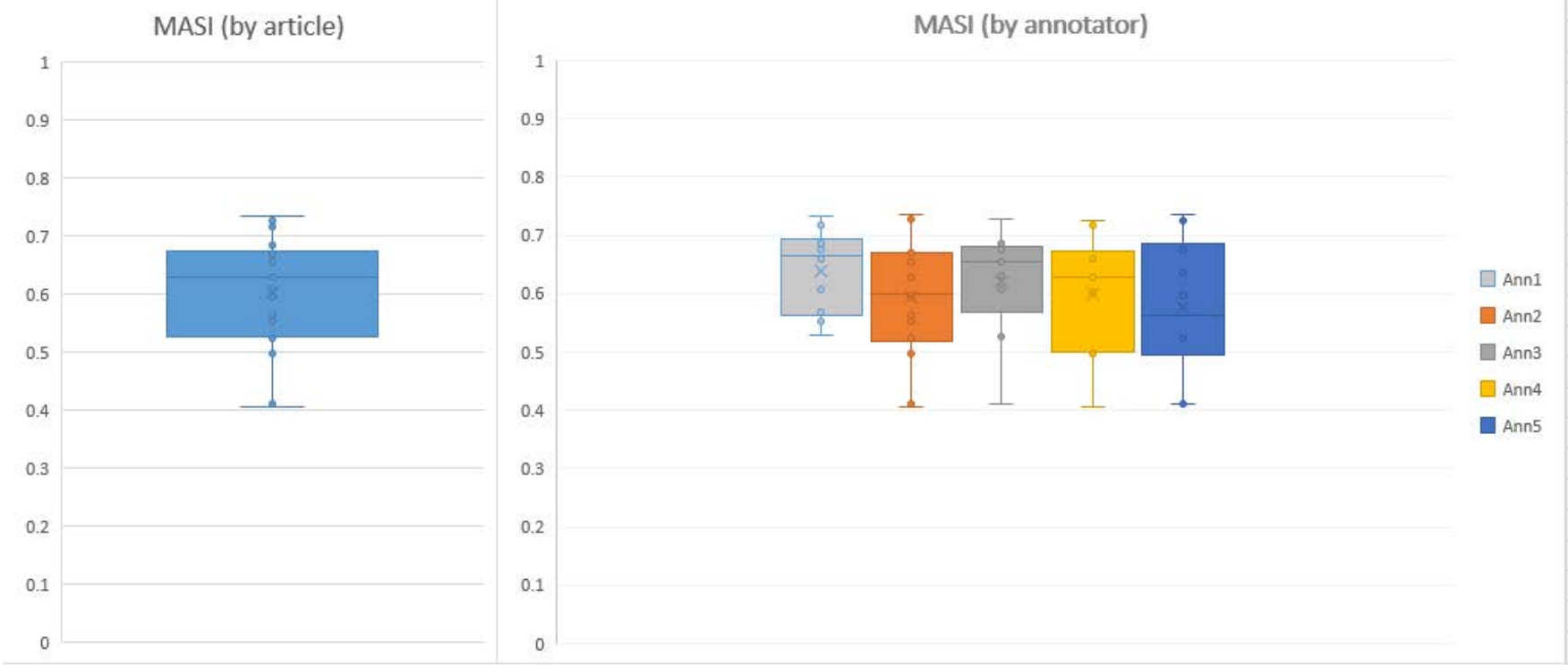
	Total	Mean (Range)	Median (IQR)
Annotated sentences	4845	96.9 (61-158)	92.5 (80.0-109.8)
Annotations	5679	113.6 (66-197)	110.5 (93.8-126.5)
Items per article		27.5 (15-35)	28 (25-31)

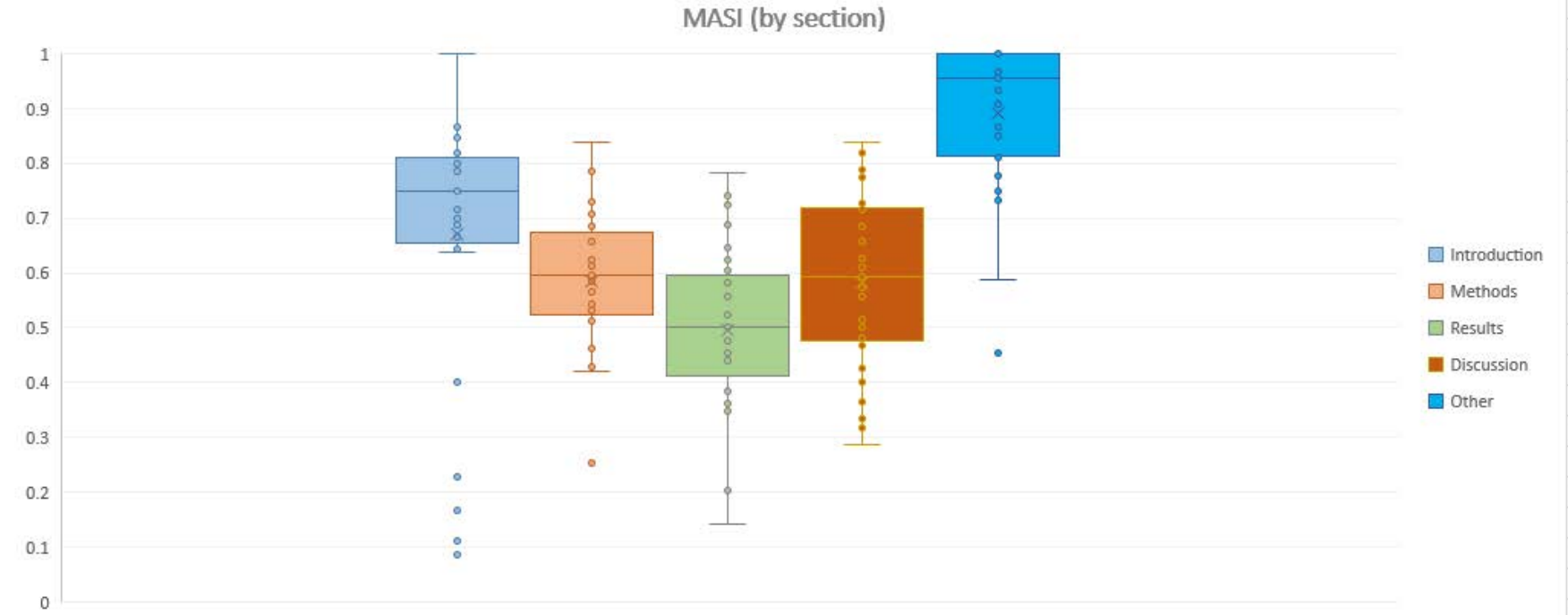
Patients were randomly assigned, using a computer-generated randomization schedule, from a central location utilizing an interactive voice response system with blinded medication kit number allocation in a 2:1 ratio to identical-appearing tablets of HZT-501 (800mg ibuprofen and 26.6mg famotidine) or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

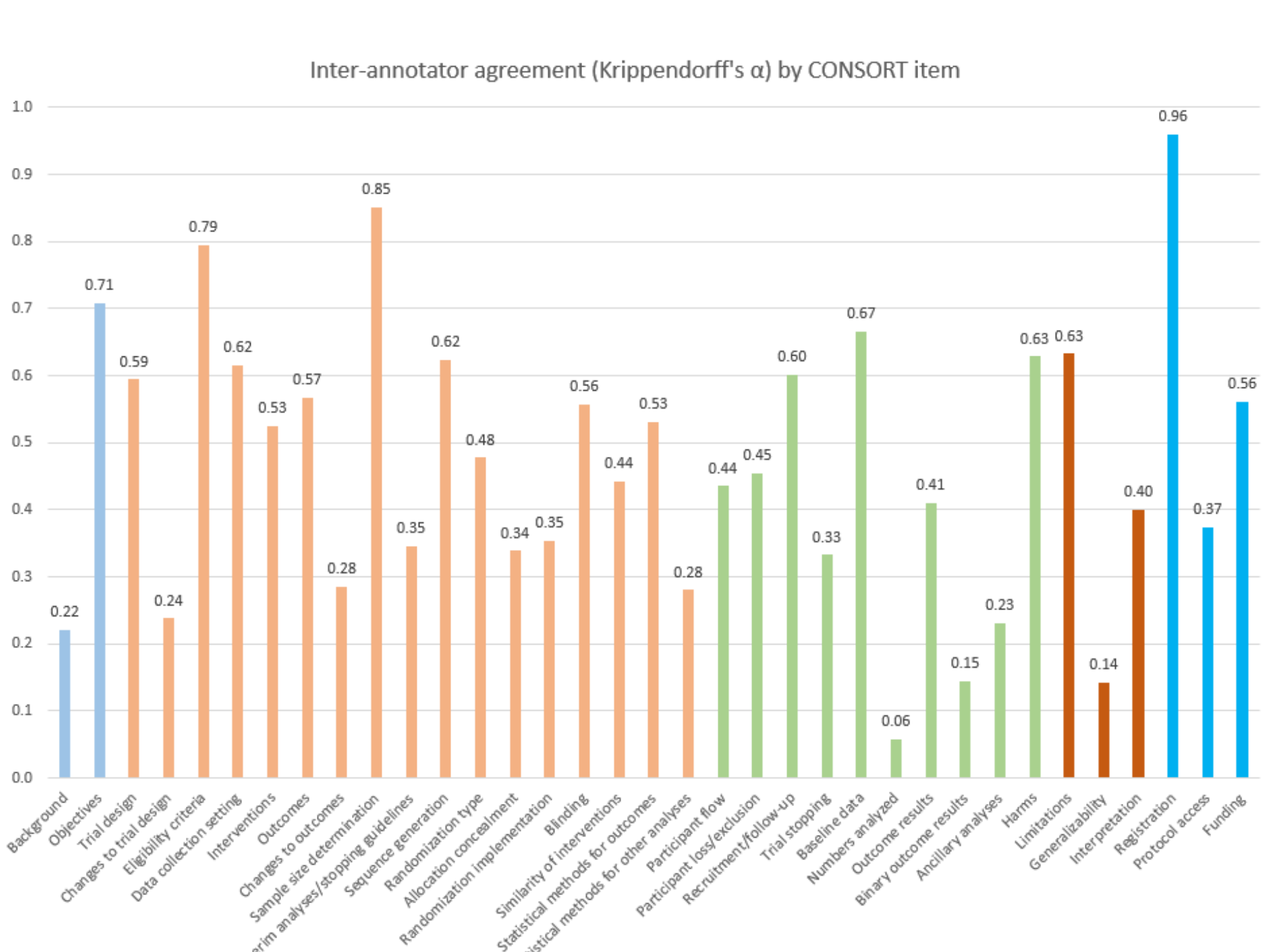
- Trial design, Sequence generation, Allocation concealment, Randomization implementation, Similarity of interventions


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


MASI (by article)
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

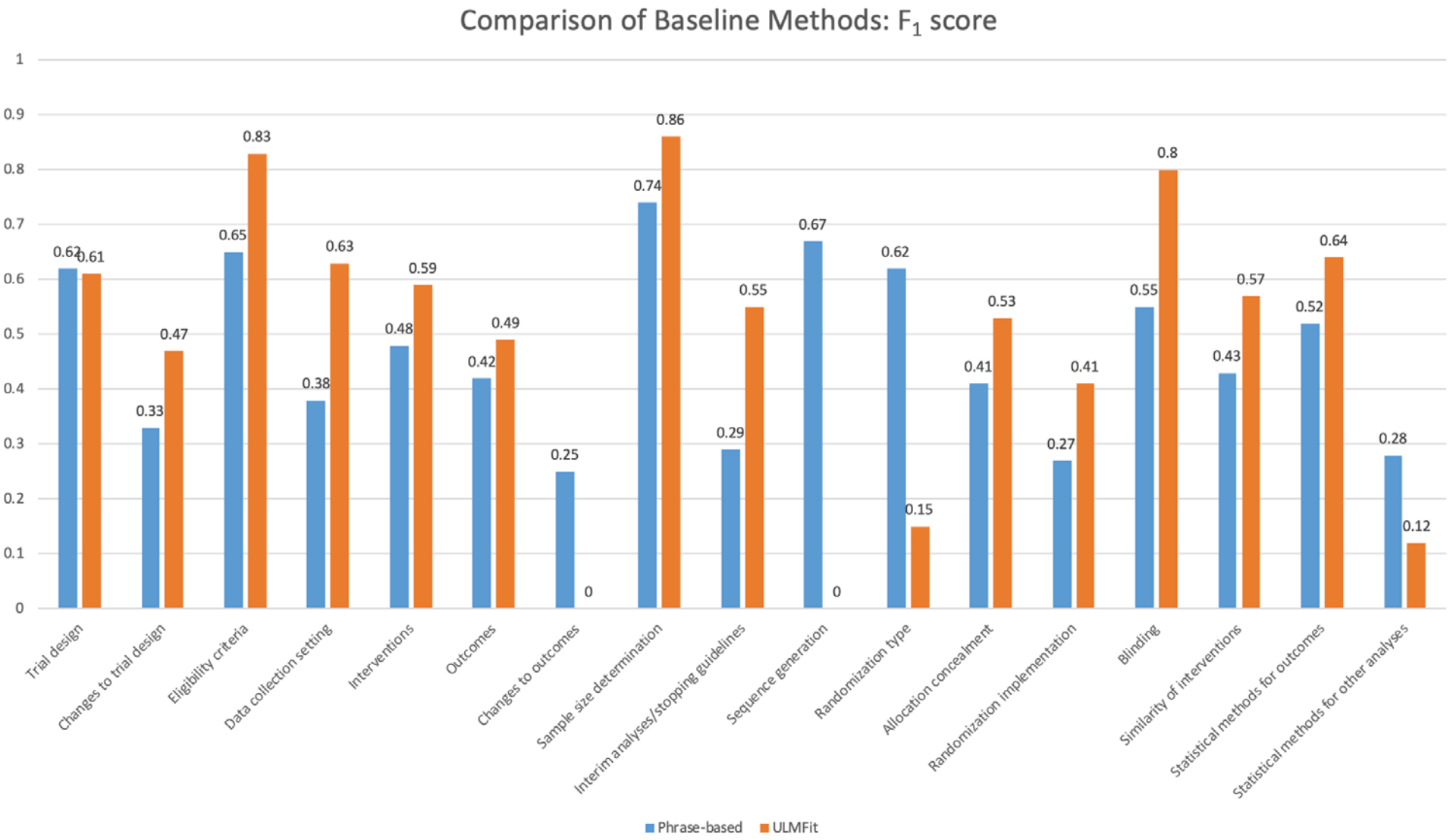
- Applied to Methods sections and Methods-specific items
  - Trial design, Sample size determination, Blinding procedure, etc.
- Rule-based methods
  - Automatic analysis of frequent subsection headers and phrases
  - ~15K unlabeled clinical trial publications
  - Phrase-based classification
    - *"masked to treatment"* ⟶ Blinding procedur
  - Subsection header-based classification
    - *"change" … "plan"* ⟶ Changes to trial design

Baseline Classification Experiments

- ULMFit (Universal Language Model Fine-Tuning)
  - Model pretrained on a large corpus, Wikitext-103 (103M tokens)
  - Fine-tuned on the task corpus
- Training data (with some noise) automatically generated
  - Phrase-based classification
  - Subsection header-based classification
  - RobotReviewer predictions (for Eligibility criteria, Interventions, Outcomes)
  - Nearest neighbor based on sentence vector similarity
- Validation and testing with the manually annotated dataset

Preliminary Evaluation Results

- Checklist item-level evaluation
  - Macro-precision (p), macro-recall (r), macro-F1 (f)
  - Phrase-based (p: 0.54, r: 0.48, f: 0.46)
  - ULMFit (p: 0.51, r: 0.54, f: 0.49)
- Article-level evaluation
  - CONSORT item present in the article or not?
  - Phrase-based (p: 0.88, r: 0.80, f: 0.84)
  - ULMFit (p:0.87, r: 0.83, f: 0.85)


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
0.61
0.33
0.47
0.65
0.83
0.38
0.63
0.48
0.59
0.42
0.49
0.25
0
0.74
0.86
0.29
0.55
0.67
0
0.62
0.15
0.41
0.53
0.27
0.41
0.55
0.8
0.43
0.57
0.52
0.64
0.28
0.12
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
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 (Interpretation)
  - Similar (Outcome result, Binary outcome result, Ancillary analyses)
- The manually annotated corpus can be used as a benchmark
- Simple phrase-based method yields moderate results
- ULMFit has similar performance

Agenda

- Natural language processing for enhancing research rigor, integrity, and transparency
- Recent work
  - Assessing clinical trial publications for reporting guideline adherence
  - Supporting meta-research investigations with natural language processing

Peer Review and Limitation Reporting

- *Spin*: reporting practices that distort the interpretation of the results of a study [Chiu et al., 2017]
  - Failure to acknowledge the limitations of the study
  - Inappropriate overstatement of claims
- Hypotheses
  - Compared to subsequent publications, discussion sections of submitted manuscripts
    - Discuss fewer limitations
    - Make stronger claims

Keserlioglu K, Kilicoglu H, ter Riet, G. "Impact of peer review on discussion of study limitations and strength of claims in randomized controlled trial reports: a before and after study." *Research Integrity and Peer Review, 2019(4):19.*

Approach

- NLP methods applied to discussion sections of manuscript/publication pairs
  - Recognize limitation sentences (91.5% accuracy) [Kilicoglu et al., 2018]
  - Identify speculative sentences ("hedging") and measure their level of speculative-ness ("normalized hedging score") (93% accuracy) [Kilicoglu and Bergler, 2009]
- 446 RCT reports from BMJ Open and 27 BMC journals
  - Open peer review

Results

- Limitation reporting increases by 56% in peer review
  - But the difference is small in absolute terms

    (2.48 vs. 3.87 sentences)
  - Greater impact on manuscripts with zero or low number of limitation sentences
- No support for the hypothesis that the peer review leads to increased hedging of claims
  - Authors are asked to both temper their statements and hedge less, resulting in minimal changes

Translatability of Animal Studies

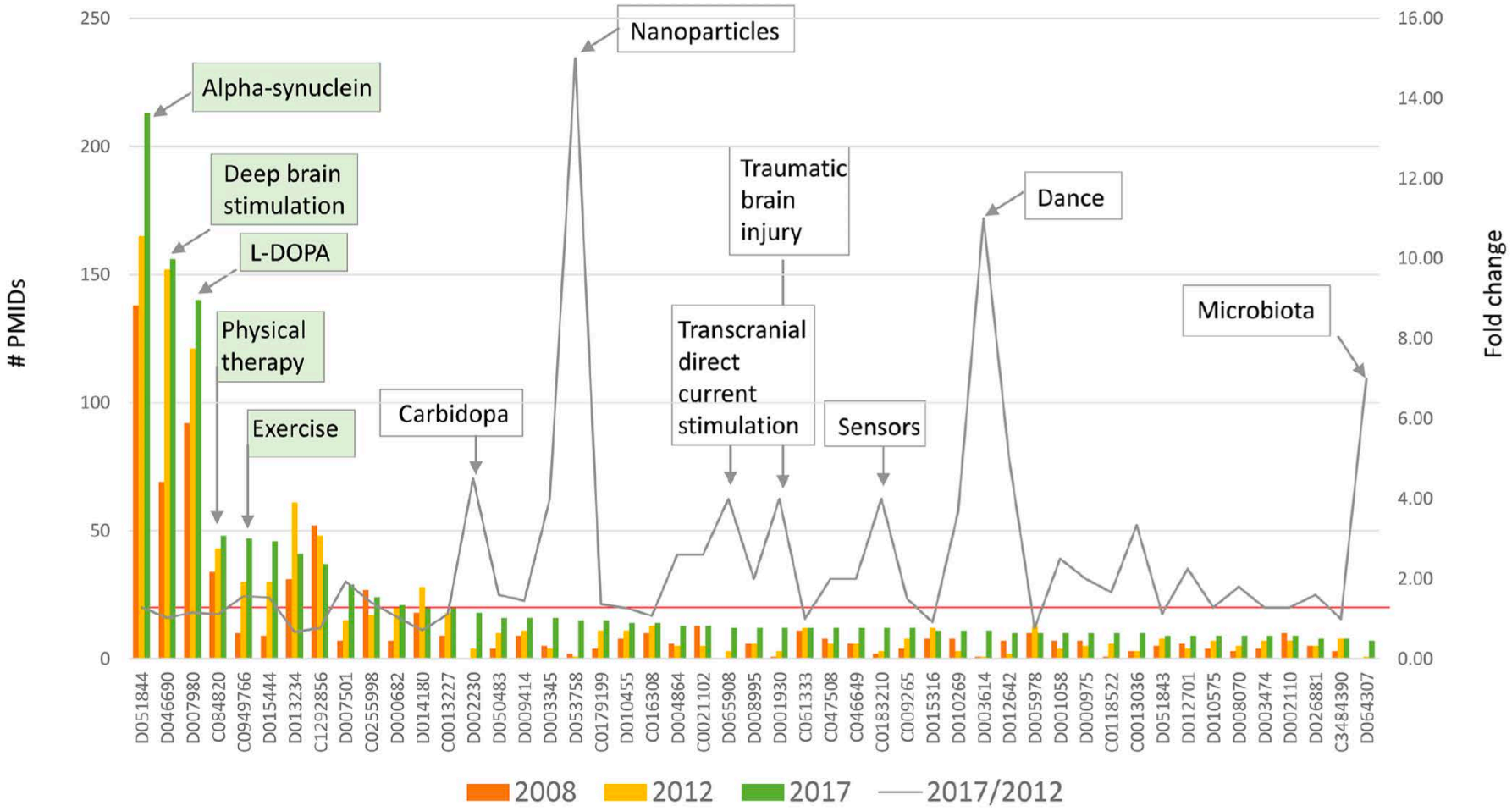
- No disease-altering therapies for the most common neurodegenerative diseases are available
  - Promising results in animal models for the same conditions
- Assess therapeutic approaches in the context of
  - Species, models, molecular targets, outcomes
  - Large-scale analysis to reveal patterns of animal use

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." *Under review.*

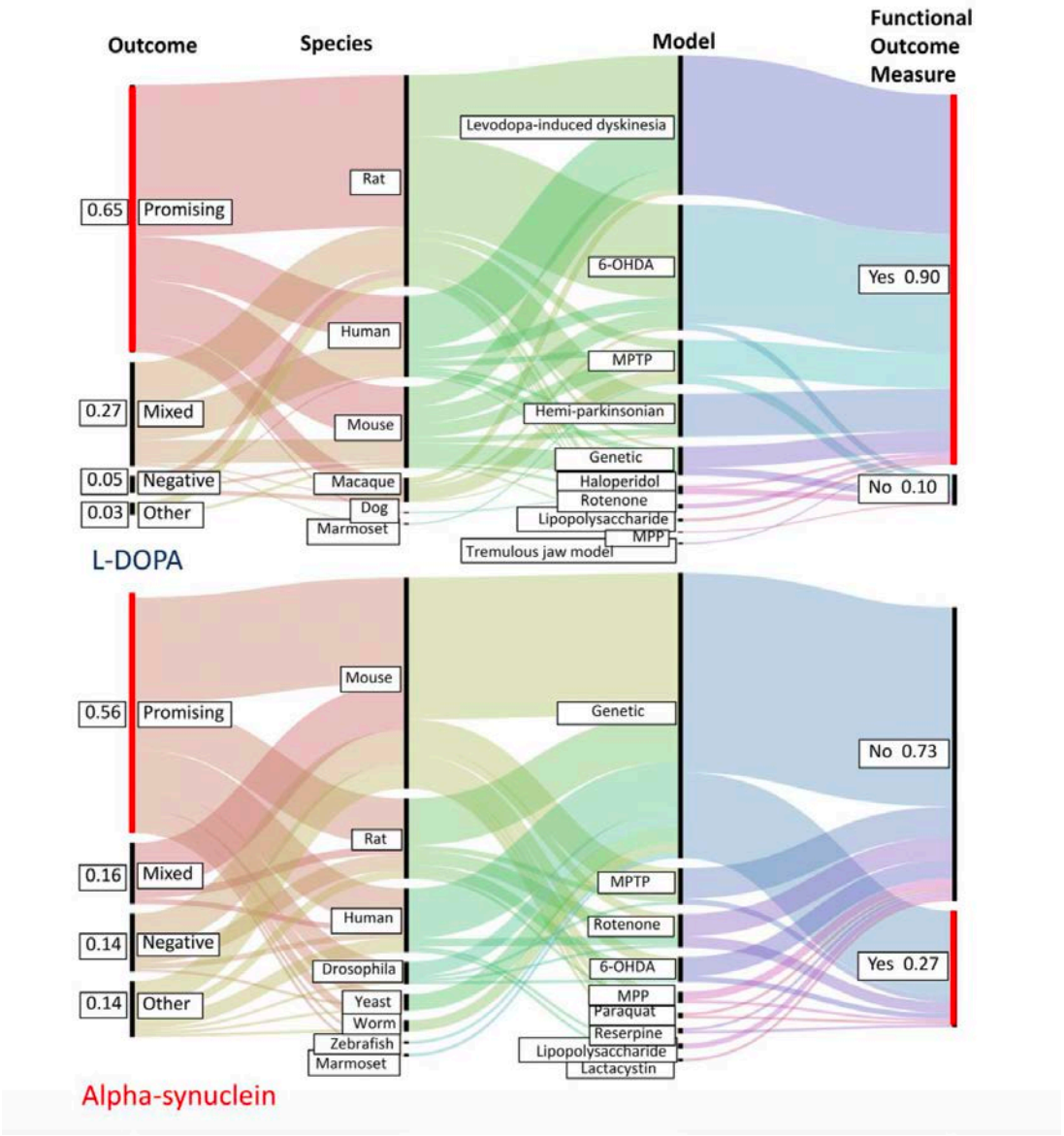
Case Study: Parkinson's Disease

- Text mining approach to extract
  - Interventions, species, models, molecular targets, study outcome, functional outcome measures
  - Off-the-shelf NLP tools augmented with rules, term lists, and supervised machine learning
- Annotation and intrinsic evaluation on 504 abstracts
- Large-scale trend analysis on 3-year data
  - ~15K abstracts (2008, 2012, 2017)
  - e.g., Species=MOUSE and Model=GENETICALLY ALTERED MODEL and StudyOutcome=PROMISING

Therapeutic Intervention Trends



Integrating Information from Modules



Concluding Remarks

- NLP/text mining can aid in assessing and improving research practices
- But many challenges lie ahead
  - Availability of full-text articles and other textual artifacts
    - Restrictions on text mining
  - Annotated corpora
    - Difficulty of manual annotation
  - Information in modalities other than narrative text
    - Tables, figures, supplementary data
  - Accuracy of methods