Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230210

First Steps Towards a Risk of Bias Corpus of Randomized Controlled Trials

Anjani DHRANGADHARIYA[a,b,1] ,Roger HILFIKER [c], Martin SATTELMAYER[d], Katia GIACOMINO[d], Rahel CALIESCH[d], Simone ELSIG[d], Nona NADERI[e,f] and Henning MÜLLER[a,b]

[a] *Informatics Institute, HES-SO Valais-Wallis, Sierre, Switzerland*
[b] *University of Geneva (UNIGE), Geneva, Switzerland*
[c] *IUFRS, University of Lausanne, Lausanne, Switzerland*
[d] *School of Health Sciences, HES-SO Valais-Wallis, Leukerbad, Switzerland*
[e] *Geneva School of Business Administration, HES-SO Geneva, Switzerland*
[f] *SIB Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland*
ORCiD ID: Anjani Dhrangadhariya https://orcid.org/0000-0003-1691-1338

Abstract. Risk of bias (RoB) assessment of randomized clinical trials (RCTs) is vital to conducting systematic reviews. Manual RoB assessment for hundreds of RCTs is a cognitively demanding, lengthy process and is prone to subjective judgment. Supervised machine learning (ML) can help to accelerate this process but requires a hand-labelled corpus. There are currently no RoB annotation guidelines for randomized clinical trials or annotated corpora. In this pilot project, we test the practicality of directly using the revised Cochrane RoB 2.0 guidelines for developing an RoB annotated corpus using a novel multi-level annotation scheme. We report inter-annotator agreement among four annotators who used Cochrane RoB 2.0 guidelines. The agreement ranges between 0% for some bias classes and 76% for others. Finally, we discuss the shortcomings of this direct translation of annotation guidelines and scheme and suggest approaches to improve them to obtain an RoB annotated corpus suitable for ML.

Keywords. risk of bias, annotation, systematic reviews, corpus, automation

1. Introduction

Systematic reviews (SRs) synthesized from randomized controlled trials (RCTs) are the highest quality evidence in the evidence hierarchy and are used by doctors to make diagnostic and treatment decisions. In theory, an RCT accurately measures the treatment effect on patient outcomes but can be biased in practice due to flawed study design, execution, analysis, or outcome reporting [1]. Biases in RCTs cannot be measured, but risk bias can be assessed. So, the reviewers must rigorously look for possible biases before incorporating them into SRs. Published RCTs are exponentially increasing[2], making manual assessment a protracted process. Machine learning (ML) can help

[1] Corresponding Author: Anjani Dhrangadhariya, Informatics Institute, HES-SO Valais-Wallis, Technopole 3, 3960 Sierre, Switzerland; E-mail: anjani.dhrangadhariya@hevs.ch.

[2] https://pubmed.ncbi.nlm.nih.gov/?term=randomized%20controlled%20trial&filter=pubt.randomizedcontrolledtrial

accelerate this process by directly pointing the reviewers to the parts of the text relevant to identifying RoB, leading to quickly judging the trial quality. Automation will thereby accelerate the process of writing SRs, which is tedious and time-consuming. Both Marshall *et al.* and Millard *et al.* attempted automated RoB assessment, albeit using proprietary, pay-walled data [2,3]. Recently, Wang *et al.* released a hand-labelled RoB corpus for preclinical animal studies, not RCTs [4]. RoB assessment of RCTs is a knowledge-heavy task where even highly trained experts are prone to subjective judgments. Developing such a corpus entails creating a clear-cut annotation scheme and guidelines. As neither exists, we focus on two primary concerns: 1) To test whether the widely used revised Cochrane's RoB 2.0 tool for RCTs (RoB 2.0) could be used as RoB annotation guidelines to develop a corpus that could be used for training ML models. 2) To develop and test an RoB annotation scheme that closely mimics the RoB 2.0 [5,6].

2. Methods

2.1. Formulating annotation scheme

RoB 2.0 tool divides biases into five risk domains which further decompose into several signalling questions (SQ), each corresponding to different parts of the trial design. Each signalling question prompts the reviewer to look for a piece(s) of factual evidence in the RCT and, depending on the amount of evidence found to respond with one of the five response options: “Yes”, “Probably yes”, “No”, “Probably no”, or “No information”. E.g., to respond to the SQ “Was the allocation sequence random?”, the reviewers need to identify whether a proper methodology was used for random participant allocation, and only if a proper methodology is identified the reviewer responds to this question as “Yes”, and otherwise “No”. We formulated an annotation scheme (see Figure 1 where each SQ is an entity. Each entity has five entity labels corresponding to the five response options to that question. Entities represent the factual evidence from the RCTs, and the entity labels incorporate the reviewer's risk judgment.

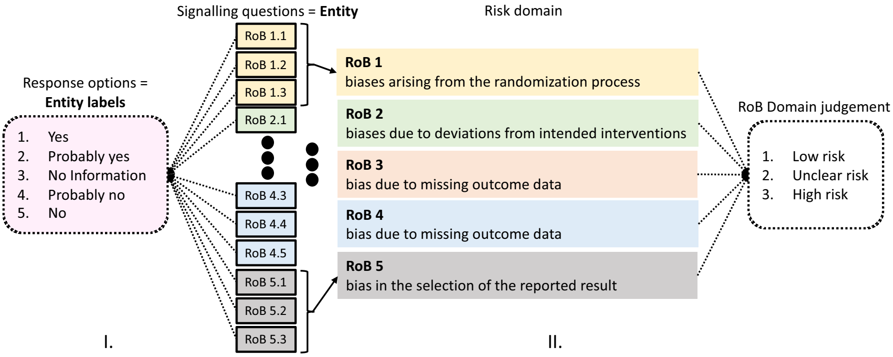

Figure 1. Annotation scheme. I. SQ level: each SQ (RoB 1.1, 1.2, ...) is an entity that could take either of five response options (entity labels). SQ response judgements for individual risk domains (RoB 1-5) could be combined to arrive at risk domain judgement. Note: We do not address risk domain judgments in this work.

2.2. Preliminary annotation guidelines

Full-text RCTs were annotated using the RoB assessment instructions from RoB 2.0.[3] The author, with Natural Language Processing expertise, developed the generic annotation guidelines with four physiotherapists experienced in bias assessment to ensure consistency. Complete sentence(s) or phrase(s) were annotated depending on the text parts relevant to answering an SQ. All the text information pertinent to answering a question was marked, even if the information was found in different parts of the full text. Table or figure captions relevant to answering were marked. If the information was not found in the captions, it was marked within the table contents. If a table or figure reference answered the question, it was annotated.

2.3. Pilot annotation

R.H., M.S., K.G., and R.C. consented to annotate and did the pilot annotation on a corpus of ten RCTs sampled in the following manner. An Entrez[4] search using the search query "`(randomized[title] or randomized[title]) and (rehabilitation or (physical therapy))`" was performed ten times to retrieve studies from one-year timespans, each between 2000 - 2019. Each query was restricted to retrieve 1000 documents, of which ten were randomly chosen for each period. We took the first possible study of the ten sampled studies with a freely available PDF (Portable Document Format). R.H. and M.S. are professors and associate professors, and K.G. and R.C. are doctoral researchers with experience conducting RoB ratings in several SRs. Tagtog[5], a commercial tool, was used for annotating PDFs. The task was to annotate text relevant to answering each signalling question entity and choose a judgment response option entity label. We report the pairwise, token-level F1 that disregards out-of-the-span (unannotated) tokens, which is the ideal measure of annotation reliability for the token-level annotations. [7] F1 is reported for entity IAA_{sq} and entity label $\text{IAA}_{\text{response}}$ annotations. IAA_{sq} and $\text{IAA}_{\text{response}}$ measure the reliability of the RoB 2.0 guidelines for selecting the same parts of the text to answer SQs.

3. Results

The pilot annotation resulted in 902 labels corresponding to the SQs and their response options. Table 1 reports pairwise IAA_{sq} and $\text{IAA}_{\text{response}}$ averaged over all the annotator pairs at the SQ response option level. Individual pairwise IAA_{sq} range between 0% (poor) and 75% (substantial), with most values falling under the poor category and very few under the substantial agreement. SQs RoB 1.1, 1.2, 1.3, 2.6, and 3.1 fared well regarding the average pairwise agreement between all pairs, but none of these categories had a substantial agreement. Questions 2.1, 2.3, 2.4, 2.5, 2.7, 3.4, 4.4, 4.5, and the entire domain 5 fared extremely poorly or with no agreement or annotation. The $\text{IAA}_{\text{response}}$ scores are considerably lower (to zero) than IAA_{sq}, hinting that annotators choose the

[3] https://drive.google.com/file/d/19R9savfPdCHC8XLz2iiMvL_71lPJERWK/view

[4] The Entrez Global Query Cross-Database Search System is a federated search engine or web portal that allows users to search PubMed database.

[5] https://www.tagtog.com/

same text to answer an SQ but assign different response options to the selected text. The $\text{IAA}_{\text{response}}$ scores remain variable across the risk domains, with 52.63% of the total scores being zero and no annotation for about 22% of the total scores.

Table 1. Left: Table lists IAA_{sq} between the six annotator pairs (P1-P6)[6] for the RoB SQs. Substantial (≥61) agreements are in bold. Right: Table lists IAA_{sq} averaged over the six annotator pairs for the SQs at the entity label level ($\text{IAA}_{\text{response}}$). Note Y = Yes, PY = Probably Yes, NI = No Information, N = No and PN = Probably No, Avg. = Average. "-" shows that one of the annotators did not annotate any text for a particular SQ.

SQ	P1	P2	P3	P4	P5	P6	Avg.	Y	PY	NI	PN	N
1.1	23.1	24.5	52.2	57	48	21.5	37.7	21.8	7.1	0	-	-
1.2	**66.1**	50.3	72.8	50.7	46	50.5	56.1	4.9	11.5	10.2	0	-
1.3	**69.5**	20.5	16.1	31.6	59.9	53.5	41.8	-	-	41.8	11.4	9.9
2.1	1	1.4	0	9.1	19.1	0	5.1	8.2	0	-	3	0
2.2	18.3	7.3	11.1	0	23	7.4	11.2	3.6	0	0	0	0
2.3	20.6	5.5	13.4	0	0	0	6.6	-	0	-	1	0
2.4	0	-	-	0	0	-	0	-	0	-	0	-
2.5	0	0	0	0	0	-	0	0	0	-	0	-
2.6	**75.3**	**68.9**	19.3	**63.9**	12.9	19.6	43.3	39.4	0	0	0	3.6
2.7	0	6.6	0	0	0	0	1.1	0	0	-	0	0
3.1	45.8	23.6	32.2	43.4	22.9	14.8	30.4	47.6	0.6	-	1.3	3.3
3.2	1.4	0	0	3.3	7.4	0.9	2.2	0	0	-	0	0
3.3	0	0	0	16.4	0	0	2.8	-	0	31.4	0	0
3.4	-	0	-	0	0	0	0	0	0	0	0	0
4.1	4	6.6	14.2	25.6	22.3	6.3	13.2	-	-	-	0.8	12
4.2	1.8	0	0.4	0	40.1	0	7.1	-	-	-	0.3	0
4.3	7.6	13.9	5	10.5	39.5	8.4	14.2	0	0	0	13.1	20.5
4.4	0	0	0	0	0	0	0	0	0	-	0	0
4.5	0	0	0	0	0	0	0	0	0	-	0	-
5.1	0	0	0	0	0	4.2	0.7	0	0	0	0	0
5.2	23.9	0	0	0	0	2.4	4.4	-	0	0	0	0
5.3	0.2	0	0	0.4	8.1	42	8.4	-	0	0.6	0	0

4. Discussion

We analyzed annotations over all annotator pairs and RoB classes identifying four types of annotation disagreements. A **polarity disagreement** arises when two annotators choose the same chunk of text to answer an SQ but choose polar opposite entity labels ("Yes" or "Probably yes" vs "No" or "Probably no" vs "No information"). In one of the documents, all four annotators chose the same text evidence ("71 allocated routine services, 67 allocated intervention service, ...") to answer SQ 3.1. However, three of the four annotators responded to this question with "Yes", but one chose "Probably no". This SQ asks whether the outcomes data were available for all, or nearly all, participants randomized but does not clarify the exact cut-off for how many participant dropouts increase the risk. Therefore, the annotators make subjective response judgments depending upon what exact percentage of participant dropout is considered valid in their experience. A **degree disagreement** causes low $\text{IAA}_{\text{response}}$ and arises because some annotators are lenient in judging risk while others are sceptical. The lenient ones select

[6] P1 = R.H. and K.G., P2 = R.H. and M.S., P3 = R.H. and R.C., P4 = K.G. and M.S., P5 = K.G. and R.C., P6 = M.S. and R.C.

definitive "Yes" or "No" for responding to an SQ, while the sceptical ones choose "Probably yes" or "Probably no". A practical and rationally justified solution is to merge the response options "Probably yes" with "Yes" and "Probably no" with "No" to reduce the complexity of the task and increase IAA without altering the final risk judgment for this risk domain. [6] A low IAA is also caused by our annotation guidelines not limiting the annotators to selecting either the phrase vs a sentence(s) vs a paragraph for answering the question leading to a **text span disagreement**. RoB 2.0 tool led to some annotators using and annotating very condensed information to come to a response. In contrast, others used an entire paragraph to reach the same response for an SQ leading to a low token-level IAA. This problem requires mending the annotation guidelines to precisely instruct authors to select the complete information they used to decide or the minimum necessary information to decide on an SQ. Another method is automatically extending the more condensed annotations to the broadest ones. In our guideline improvement, we restrict the annotation to marking the full sentence(s) where the relevant information is found. Sometimes annotators came to a response judgment for an SQ but used different parts of the RCT text leading to **disparate document section disagreement**. Such disagreements emanate because RoB 2.0 do not instruct the annotators about what part of the RCT to annotate and what part to not annotate for a particular SQ. We noticed many SQs remained unanswered because the annotators did not understand what part of the text to annotate, even after following the RoB 2.0 guidelines.

5. Conclusion

In conclusion, the revised Cochrane RoB 2.0 guidelines cannot be directly used as RoB corpus annotation guidelines. It is imperative to develop clear-cut guidelines to instruct the annotators in signalling question and response judgment decisions. The multi-level annotation schema also needs improvement, as discussed. We are using the insights from this pilot annotation to develop detailed, crisp guidelines and obtain consistent annotations. The annotated dataset is available on Zenodo (DOI: 10.5281/zenodo.7698941).

References

[1] Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness research. BJOG: an international journal of obstetrics and gynaecology. 2018 Dec; 125(13): p. 1716.

[2] Marshall IJ, Kuiper J, Wallace BC. Automating risk of bias assessment for clinical trials. IEEE journal of biomedical and health informatics. 2015; 19(4): p. 1406--1412.

[3] Millard LA, Flach PA, Higgins JP. Machine learning to assist risk-of-bias assessments in systematic reviews. International journal of epidemiology. 2016; 45(1): p. 266--277.

[4] Wang Q, Liao J, Lapata M, Macleod M. Risk of bias assessment in preclinical literature using natural language processing. Research synthesis methods. 2022; 13(3): p. 368--380.

[5] Lansbury L, Lim B, Baskaran V, Lim WS. Co-infections in people with COVID-19: a systematic review and meta-analysis. Journal of Infection. 2020; 81(2): p. 266--275.

[6] Sterne JA, Savović J, Page M, Elbers R, Blencowe N, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019 Aug; 366.

[7] Deleger L, Li Q, Lingren T, Kaiser M, Molnar K, Stoutenborough L, et al. Building gold standard corpora for medical natural language processing tasks. In AMIA Annual Symposium Proceedings; 2014. p. 144.