Research Integrity Indicators in the Age of Artificial Intelligence

Leslie D. McIntosh[*], Simon Porter[**], Cynthia Hudson Vitale[***]

[*] *leslie@digital-science.com*
https://orcid.org/0000-0002-3507-7468
Vice President, Research Integrity, Digital Science, UK

[**] *s.porter@digital-science.com*
https://orcid.org/0000-0002-6151-8423
Vice President, Research Futures, Digital Science, UK

[***] *cvitale@arl.org*
https://orcid.org/0000-0001-5581-5678
Director of Science Policy, Association of Research Libraries, USA

Abstract

Generative artificial intelligence (AI) and large language models significantly change how disciplines and communities analyze and report research. Leveraging these new tools, such as ChatGPT or Bard, authors can easily generate text and analyses for research articles. As a result, we have already witnessed several instances in which generative AI was used to write a paper or manuscript which unknowingly contained fake citations or false information. The scholarly community needs new indicators to signal, assess, and evaluate manuscripts and research quality to fortify public trust in research. This paper proposes a set of indicators for research integrity that encompasses the much-needed transparency for generative AI. We have then used AI to train algorithms to detect these indicators and applied them to 33 million full-text research publications. We can now see the key indicators as metrics to understand where various fields of research are in communicating and signalling trust.

1. Introduction

Large language models (LLM), such as ChatGPT or Bard, are a type of generative artificial intelligence (AI) that use machine learning to generate text based on the statistical likelihood or frequency of those words in a dataset. Unlike web search engines that return websites as results, generative AI pulls together text in response to a prompt. In using web-based sources as their training dataset, LLMs produce texts that often seem to be accurate, but may not be given that the text is produced based on this likelihood-model algorithm.

Questions and concerns about the use of text authored by generative AI tools in scholarly publishing surfaced almost immediately after the emergence of ChatGPT and Bard with problems highlighting the difficulty in identifying AI-generated text. According to research conducted by

Dr. Catherine Goa, abstracts created by ChatGPT were submitted to academic reviewers, who could only spot ChatGPT generated abstracts 68% of the time. The reviewers also incorrectly identified 14% of real abstracts as being AI generated (Paul, 2023). .

Yet, leveraging the power of AI-backed by an ontology may quell the fears of mistaking faked research with quality communication. Somewhat relatedly, across the scientific publishing landscape there exist many checklists, guidelines, and best practices - or, more broadly, indicators - for the responsible reporting of research - as a set of research trust indicators. The EQUATOR network (Equator Network, n.d.), which tracks and makes discoverable many of these guidelines, includes over 500 different entries.

These reporting guidelines are a critical indicator to ensure research is adequately reported to verify its integrity and potential reproducibility. In many ways, these guidelines and best practices represent a knowledge system for responsible reporting (MIT, 2020).

In the case of research reporting two primary problems have emerged: i) Its knowledge system has developed without an ontology or set of key indicators. Thus, normalising and standardising classes, subclasses, and relationships, through an ontology is a critical first step in alleviating much confusion and burden on users - while also ensuring the quality of the scientific reporting (Jones, 1998). and, ii) Due to the sheer volume of publications even before generative AI, an automated approach to detecting the key indicators is needed.

In our work, we have synthesized the primary reporting indicators based on community engagement to define trust markers of publications. Then we trained, tested, and validated algorithms to automatically detect text within articles (McIntosh, 2023).

2. Methods

Developing Research Integrity Indicators

Based on previous research and building from the Repeatability Assessment Tool (RepeAT) Framework (McIntosh, Juehne & Vitale, 2017). The RepeAT framework was developed through a multi-phase process that involved coding and extracting recommendations and practices for improving reproducibility from publications and reports across the biomedical and statistical sciences, field testing the instrument, and refining variables. This Framework surfaced 5 key classes for assessing reproducibility and 119 subclasses that could be used to evaluate the quality of scientific reporting. While frameworks such as this are thorough, the criticalness of each variable for quality reporting was not evaluated. Further, this early research and framework focused heavily on the reproducibility of research re-using electronic health records.

Starting with the RepeAT Framework, the research team then conducted a comparative analysis of these variables and seven reporting guidelines and one publisher-based research reporting tool. Table 1 includes a list of all guidelines included in the comparative analysis.

Table 1: Reporting Guidelines and Checklist

Reporting Guidelines/Checklist	Study Design
MDAR	Establishes a minimum set of requirements in transparent reporting applicable to studies in the life sciences

ARRIVE 2.0	Animal pre-clinical research
CONSORT 2010	Clinical trials, Experimental studies
STARD 2015	Clinical trials, Diagnostic and prognostic studies, Experimental studies, Observational studies
SPIRIT 2013	Clinical trials, Experimental studies, Study protocols
STROBE	Observational studies
PRISMA 2020	Systematic reviews, Meta-analyses, Reviews, HTA, Overviews
AGREE	Clinical practice guidelines
SQUIRE	Quality improvement studies

The goal of this phase in the research was to determine the key elements or critical elements of reporting - now called trust markers. To do this, the research team mapped indicators across these reporting guidelines and collated all sections of the reporting guidelines into a full set of classes.

Trust Marker Algorithm Development

Trust marker development begins by reviewing a corpus of relevant articles (initially around 100) to develop a sense of common practices and locations for reporting the indicator. These definitions form the basis of a data dictionary to create annotation guidelines, which in turn provides accurate and consistent training data for our models.

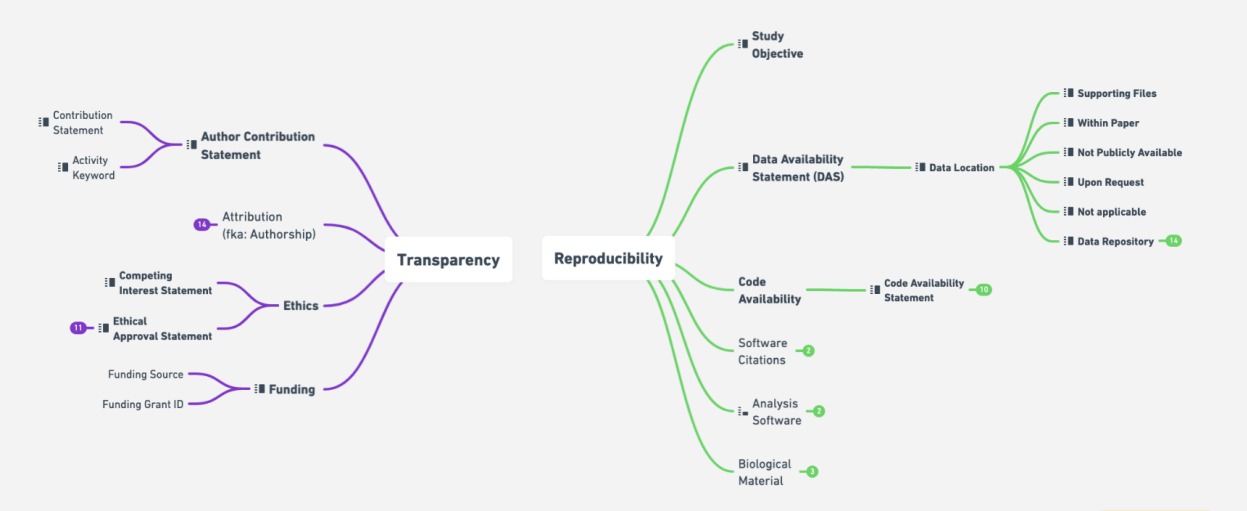
Once a corpus of relevant papers have been identified we use a tool called Prodigy (https://prodi.gy/) to present the publication text extracted to annotate. Prodigy is an annotation and training tool that allows for rapid NLP model development. The system uses active learning, encouraging human annotators to annotate the documents the current model is least sure about. As a human answers the prompts in Prodigy, the model is updated. After the first round of annotations, we will train alpha models. Our experience indicates that the first round of annotations often requires us to clarify or refine the data dictionary described above. A second round of annotations with a new corpus of approximately 500 articles will then be completed.

The algorithms produced are then applied to the full-text publications within Dimensions.

3. Results

Compiling section names and variables across the RepeAT Framework, the seven reporting guidelines, and the publisher research reporting tool, resulted in over 178 subclasses and 17 classes. These classes and subclasses were then compared across the nine reporting guidelines and the one publisher checklist. These classifications were then organised into three categories of article quality: authorship and attribution, transparency, and reproducibility. The taxonomy of trust markers are shown in Figure 1.

Figure 1: Key Indicators as Trust Markers



The transparency category represents necessary subclasses to support transparency and fidelity of critical research reporting practices. It creates a set of indicators to represent good standards of practice for research communications, and may be external to the manuscript, such as the publishing of a protocol or registration.

The reproducibility category is centered around the elements of a paper which may facilitate a future researcher's ability to achieve the same results when replicating the original study. The presence or lack of these elements and the subclasses below does not definitively determine the reliability of the authors, merely presents a likelihood.

Trust Marker Algorithm Implementation

When these trust marker algorithms are applied to 33 million full-text journal articles located in the Dimensions database, adoption practices of the trust markers across fields of research become more apparent. As shown in Figure 2, based on the adoption percentages for the calendar year 2021, fields of research are assigned policy implementation bands. For Fields in band 1, there is already well-established practice of reporting the trust markers. It would be reasonable to work towards 100% compliance for all papers. For band 2, there is awareness of the trust marker, but more training is required to shift practice. For band 3, low awareness is assumed and significant education or training is needed in those fields.

Figure 2: Field of Research and Adoption Practice

Field of research	Funding statement	Competing interests	Author contributions	Data availability	Ethics approval
Biomedical and Clinical Sciences	1	1	1	1	1
Health Sciences	1	1	1	1	1
Psychology	1	1	1	1	2
Biological Sciences	1	1	1	1	2
Environmental Sciences	1	1	2	2	2
Economics	1	1	2	2	2
Agricultural, Veterinary and Food Sciences	1	1	2	2	2
Chemical Sciences	1	1	2	2	2
Earth Sciences	1	1	2	2	2
Engineering	1	1	2	2	2
Built Environment and Design	1	1	2	2	
Physical Sciences	1	2	2	2	2
Human Society	1	2	2	2	2
Information and Computing Sciences	1	2	2	2	2
Mathematical Sciences	1	2	2	2	2
Commerce, Management, Tourism and Services	2	2	2	3	2
Education	2	2	3	3	1
Philosophy and Religious Studies	2	2	3	3	2
Creative Arts and Writing	2	2	3	3	
History, Heritage and Archaeology	2	2	3	3	
Language, Communication and Culture	2	2	3	3	
Law And Legal Studies	2	2	3	3	

Table: Dimensions Research Integrity • Source: Dimensions

4. Conclusion

The scientific community and publishing industry needs more streamlined methods and approaches to ensuring the responsible reporting of research in manuscripts and assessment of manuscripts that may leverage generative AI. While certain study types and research participant types may require specificity, there are many general indicators that can be standardised and normalised across the research communications ecosystem.

All quality indicators can and are being developed to be automatically extracted from publications (McIntosh, 2021). What they offer is a faster means to check the quality of the vital piece (manuscript) of scholarly communication. They provide more than a metric of attention. They provide a signal of trust.

References

Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37–46.

Equator network. (n.d.). Retrieved April 23, 2022, from https://www.equator-network.org/reporting-guidelines/.

Jones, D., Bench-Capon, T., & Visser, P. (1998). Methodologies for ontology development.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–74.

MIT, Knowledge Meta-Networking for Decision and Strategy. (2020). Retrieved April 30, 2022, from https://gssd.mit.edu/knowledge-system.

McIntosh, L.D., Juehne, A., Vitale, C.R.H. *et al.* Repeat: a framework to assess empirical reproducibility in biomedical research. *BMC Med Res Methodol* 17, 143 (2017). https://doi.org/10.1186/s12874-017-0377-6

McIntosh, L.D., Automating Quality checks in the publishing process. *Transforming Scholarly Publishing with Blockchain Technologies and AI.* (2021). https://doi.org/10.4018/978-1-7998-5589-7.ch013

McIntosh, L.D., Whittam, R., Porter, S., Vitale, CR.H., Kidambi, M.; Science, Digital (2023): Dimensions Research Integrity White Paper. Digital Science. Report. https://doi.org/10.6084/m9.figshare.21997385.v2

Paul, M., When ChatGPT writes scientific abstracts, can it fool study reviewers? (2023). Retrieved April 21, 2023, from https://news.northwestern.edu/stories/2023/01/chatgpt-writes-convincing-fake-scientific-abstracts-that-fool-reviewers-in-study/.

Open science practices

Data and analysis for the checklist are available on our GitHub webpage: https://github.com/CBMIWU/Research_Reproducibility.

Competing interests

Leslie McIntosh is an employee of Digital Science and the founder of Ripeta, now part of Digital Science. Simon Porter is an employee of Digital Science. Cynthia Hudson Vitale is co-founder and advisor for Ripeta.

Funding information

This research has received no outside funding.