Ripeta Approach and Criteria Definitions

This document outlines Ripeta's workflow and defines the reproducibility criteria that we use to evaluate manuscripts. Because Ripeta frequently updates its processes and reproducibility criteria, this document represents our workflow at a given moment in time. We will regularly publish updated versions of this document to reflect any changes that we implement.

Analyzing Manuscripts

Ripeta analyzes manuscripts using the Ripeta application, which leverages natural language processing (NLP) to identify and extract key pieces of text from scientific articles. Ripeta has developed several NLP models, each tuned to a specific reproducibility criterion. Trained to read like humans, these NLP models scan articles for seed phrases and terms that indicate the presence of their respective reproducibility criteria. The NLP models can process large numbers of articles very rapidly, making them powerful tools to hold researchers accountable.

Before the NLP models scan their first manuscripts, however, the Ripeta team must do significant pre-processing work. First, the team must understand and define each of the reproducibility criteria. Writing these definitions shapes every other aspect of our work, yet it is also our most subjective task; lenient definitions will yield dramatically different results than strict ones. Thus, the Ripeta team has given particular care to this process, analyzing current reporting practices and author guidelines and comparing existing definitions of the criteria to a set of common variables. These analyses allow the team to frame and prioritize the Ripeta reproducibility criteria. Furthermore, the Ripeta team has manually reviewed hundreds of papers for each of our criteria, allowing us to map how and where researchers present certain information.

Once the team has finished its pre-processing work, the NLP models begin scanning articles and extracting text that they recognize as matches for their respective criteria.

Though this system has yielded excellent results, Ripeta continues to push for more accurate systems. The Ripeta team consistently conducts manual quality checks across the NLP models to improve our algorithms. Now, Ripeta is working to automate this process by giving the NLP models direct feedback about their mistakes. The NLP models would then use pattern recognition to automatically revise the phrases and language patterns that they scan for, improving nuance and accuracy.

Reproducibility Criteria

Ripeta currently analyzes articles for six reproducibility criteria, each of which supports responsible science reporting. These criteria were chosen based on an in-depth literature review that the Ripeta team conducted. The review analyzed articles published between 2005 and 2015 that addressed reproducibility

in the biomedical fields. This holistic review yielded a total of 119 unique criteria that impact an article's reproducibility. These criteria could be sorted into five major categories: 1) Research design and aim, 2) Database and data collection methods, 3) Data mining and data cleaning, 4) Data analysis, and 5) Data sharing and data documentation. From these five categories, we chose the six criteria that were most consistently cited as crucial to full reproducibility. Though we would ideally analyze for all relevant reproducibility criteria, we recognize that change is an iterative process. Journals, authors, and publishers must focus on the most important reproducibility criteria before they can delve into the finer points of scientific reporting.

Below, we give the definitions of Ripeta's six criteria and provide brief explanations of how they promote reproducibility.

Study Purpose

Definition: A concise statement in the introduction of the article, often in the last paragraph, that establishes the reason the research was conducted. Also called the study objective.

The inclusion of a study purpose does not, at first glance, impact reproducibility. However, the study purpose grounds the research article, giving context and clarity to the reader and providing a standard for reproducibility. Each piece of the article, from the methods to the conclusion, can be checked against the study purpose for congruency. In the case of a discrepancy, the reader will be tipped off to an issue with the article and can inspect it further. Without a study purpose, the reader has little basis to determine the appropriateness of the methods, analysis, and conclusions.

Data Availability Statement

Definition: A statement, in an individual section offset from the main body of text, that explains how or if one can access a study's data. The title of the section may vary, but it must explicitly mention data; it is therefore distinct from a supplementary materials section.

Data sharing is fundamental to research reproducibility because data allows other researchers to determine if the stated analysis methods and data could have actually produced the stated conclusions. In other words, data sharing allows other researchers to identify the presence of falsified data or faulty analysis methods. The research community widely accepts the benefits of data sharing and has attempted to hold authors accountable by requiring data availability statements. These statements, when properly offset from the main text, provide readers with a quick way to determine how to access data.

Data Location

Definition: Where the author states that the article's data can be accessed, either raw or processed.

Even when articles have data availability statements, their data are not always easily accessible. Where and how authors make their data available strongly influences how easy they are to access. Furthermore, research shows that data location can also serve as a proxy for the completeness of the data; for instance, full data sets are more likely to be available when they are shared in external repositories or upon request

rather than when they are made available in the article or supplemental files. Therefore, Ripeta believes it is very important to track how authors make their data available, not just whether or not authors have included data availability statements. Particularly, we see if authors have stored their data in external repositories, as data tend to be both easier to access and more complete when they are stored in this way.

Ripeta recognizes that there are legitimate reasons for authors to restrict access to certain data, particularly protected health information. We are currently training our software to identify when authors have stated reasons for their data restrictions. This improvement will allow us to differentiate between legitimate and arbitrary data restrictions.

Importantly, Ripeta only checks the author's report of where data can be found; we do not currently check whether data can actually be accessed or are sufficient to reproduce the findings.

Analysis Methods Stated

Definition: The article includes an explanation of the methods used for analysis, including statistical analysis.

Authors routinely provide thorough descriptions of the methods used to conduct their research, but many authors neglect to fully describe their analysis methods. Yet data analysis is how researchers pull their data together to create conclusions, essentially the moment of knowledge creation. Therefore, analysis methods must be both robust and appropriate for the results to hold. So, when authors fail to adequately describe their analysis methods, they leave room to question the quality of their work. Other researchers will not be able to determine if the analysis methods were appropriate or possibly mischaracterized. Furthermore, without analysis methods, it would be difficult to tell if data were falsified or misconstrued. Finally, because the analysis methods affect the results, no other researcher would be able to accurately reproduce the work.

Note, Ripeta currently only checks for an explanation of analysis methods, not the quality of that analysis. Many descriptions of analysis methods are not robust enough for full research replication.

Analysis Software Stated

Definition: Authors have indicated the specific software they used to conduct their analyses.

The analysis methods determine what conclusions are drawn from the data, and software can deeply influence the analysis process. When authors fail to state what software they used, they render their work unreproducible. We analyze papers for the presence of software information, as well as the specific software that authors report using.

Code Shared

Definition: Authors have shared access to the most updated code that they used in their study, including code used for analysis.

For many studies, code is a critical piece of the data analysis process. Yet most authors do not share their code, making it difficult to properly reproduce their studies. Many publishers recognize the importance of code sharing and require it in their author guidelines, but Ripeta has found that compliance is typically below 10 percent. Thus, Ripeta checks if authors have shared access to their code, preferably in an external repository.