Methods and Tools to Enhance Rigor and Reproducibility of Biomedical Research

Halil Kilicoglu, PhD1, Aurélie Névéol, PhD2, Timothy W. Clark, PhD3, Hua Xu, PhD4, Neil R. Smalheiser, MD, PhD5

1National Library of Medicine, Bethesda, MD; 2CNRS, LIMSI, Université Paris Saclay, Orsay, France; 3University of Virginia, School of Medicine, Charlottesville, VA; 4The University of Texas Health Science Center at Houston, Houston, TX; 5University of Illinois at Chicago, Chicago, IL

Abstract

Rigor and reproducibility of biomedical research has been the topic of much debate in recent years. Cases of replication failures, increasing number of retractions, and pervasiveness of questionable research practices lead to a lack of confidence in published findings, indicating that a portion of the biomedical research investment is wasted. Some ongoing efforts aim to address the issues in research conduct and dissemination by focusing mainly on standardization. Guidelines and principles pertaining to data, code, and publications have been proposed. The goal of this didactic panel is to engage the medical informatics community in a discussion about strategies to complement such efforts using informatics methods, tools, and resources. In the panel, first, we will provide a brief overview of standardization initiatives. Next, the panelists will present their informatics-based approaches toward improving rigor and reproducibility of biomedical research, focusing on such areas as information retrieval, natural language processing/text mining, and semantic modeling. Finally, with audience participation, we will discuss challenges facing informatics research aiming to address these problems and seek to identify some potentially fruitful research directions.

Introduction

Articles raising concerns about the rigor, integrity, and reproducibility of scientific research have been commonplace in recent years, in both the research literature and the mainstream media1. Biomedical research has seen its fair share of criticism with regard to rigor and reproducibility, in the form of reports of replication problems2,3, retractions due to scientific misconduct and fraud4, and wasteful research due to poor design and reporting as well as lack of transparency5. An analysis estimated that 85% of the biomedical research funding was avoidably wasted due to research practices leading to reproducibility problems5. Considering that billions of dollars are invested in biomedical research annually in the US alone, the magnitude of the problem is clear. Issues in rigor and reproducibility affect all stakeholders involved in biomedical research, including funding agencies, policymakers, journals, peer reviewers, and researchers. While it is unreasonable to expect to achieve perfect reproducibility, there is substantial room for improvement in how biomedical research is conducted and communicated.

A number of initiatives bringing together various stakeholders have been formed in recent years to promote rigorous and reproducible science and they have developed a considerable number of standards, guidelines, and principles focusing on data, code, reporting, and citation. Such initiatives include NIH Rigor and Reproducibility guidelines6, ICMJE recommendations for the conduct and publication of scholarly work in medical journals7, reporting guidelines for various types of biomedical studies (e.g., randomized controlled trials, observational studies, systematic reviews)8, as well as the FAIR data sharing and stewardship principles9. While such standards are increasingly adopted and have been shown to improve research transparency10, compliance to them remains inadequate and their potential may not be fully realized for years to come.

This panel aims to highlight the contribution of informatics-based approaches to addressing the problems. Such approaches can underpin tools and resources that help the stakeholders in scrutinizing and replicating conducted research in a more systematic manner as well as in managing published research more efficiently to design rigorous and reproducible studies11. In serving these roles, such tools and resources can complement standardization efforts. Recent years have seen some activity in informatics-based approaches. Some research has focused on creating infrastructure for sharing and replication. These include DataMed, a biomedical data discovery index12, MIMIC Code Repository, a centralized location for code operating on the MIMIC dataset to enable reproducibility13, and browser-based interactive notebooks and containers to document and share bioinformatics workflows14. Natural language
processing/text mining (NLP) has also been proposed to address rigor and integrity issues in biomedical research\textsuperscript{11}. Research on identifying rigorous studies for inclusion in systematic reviews\textsuperscript{15}, extracting characteristics of clinical studies from publications\textsuperscript{16,17,18,19}, matching clinical trial registrations with published articles\textsuperscript{20}, detecting plagiarism\textsuperscript{21}, and extracting/contextualizing biomedical claims\textsuperscript{11} can be subsumed in this category. Other research has focused on semantic models to support replication and verification of biomedical research. Such research include development of Micropublications, a model of biomedical claims, evidence, and arguments to improve the quality of biomedical communications\textsuperscript{22}, as well as ProvCaRe, a model that aims to model and extract semantic provenance information from published literature\textsuperscript{23}.

This panel is timely and relevant to AMIA attendees, as there is a practical need for tools and resources to support stakeholders in ensuring the accuracy, completeness, and verifiability of research conduct and dissemination. As noted above, some methods and tools that address specific tasks towards this goal have been developed and reported; however, to our knowledge, there has not been any previous panel discussion bringing together a diverse group of experts for a more comprehensive look from a medical informatics perspective. This panel aims to fill this gap and provide an opportunity to discuss current and future informatics-based collaborations and initiatives.

**Panelists**

Dr. Halil Kilicoglu (Staff Scientist at the National Library of Medicine) will act as a moderator and provide a brief overview of efforts focusing on standardization and guideline development to enhance rigor and reproducibility of biomedical research.

Dr. Aurélie Névéol (Staff Scientist at LIMSI-CNRS, France) will report on initiatives to support and foster reproducibility of research in the MIROR (Methods In Research On Research) European project and CLEF eHealth shared task. In particular, experiences in reproducing work conducted in highly controlled settings will be commented upon to illustrate barriers to reproducibility and yield lessons on how to overcome these obstacles.

Dr. Timothy W. Clark (Associate Professor at University of Virginia, School of Medicine and Data Science Institute) will report on the current state of data and software citation. He will discuss the adoption roadmaps developed by expert groups of publishers, data repository managers, and identifier management systems as well as how these are being integrated into the NIH Data Commons Pilot, and into publisher’s workflows. In addition, he will touch upon how institutional incentive structures are being addressed. He will tie in this discussion with his work in Micropublications, a semantic model of biomedical claims, evidence, and arguments\textsuperscript{22}.

Dr. Hua Xu (Professor at University of Texas Health Science Center) will introduce the development of DataMed\textsuperscript{12}, under the auspices of the bioCADDIE (biomedical and healthCAre Data Discovery Index Ecosystem) consortium. DataMed collects broad types of biomedical datasets of interest from heterogeneous sources, transforms them to a unified representation model, and then indexes them and makes them searchable through a web-based interface, thus enabling more efficient search across domain-specific repositories and making data more discoverable by users.

Dr. Neil R. Smalheiser (Associate Professor at University of Illinois at Chicago) will discuss the subtleties of identifying biomedical publications that exhibit reproducibility and integrity problems, including retractions, major corrections, editorial expressions of concern, and scientific integrity reviews. He will discuss using IR and text mining tools for assessing the reasons for the retractions, and quantifying the severity of the problems for the conclusions of the article itself, as well as the upstream effects upon later scientific investigations.

**Discussion Questions**

Potential questions include:

- What kind of resources are needed to develop methods and tools that address rigor and reproducibility problems? What are the challenges in obtaining/developing such resources?
- How can standardization efforts support automatic approaches? How can they be combined?
- What are the barriers to adoption of informatics tools and resources by stakeholders in biomedical research?
- What are some of the specific tasks that can have an impact in the short term?
- How can we measure the impact of automatic approaches, beyond traditional intrinsic evaluation metrics?
Statement of Participation
The first author affirms that all panel participants have agreed to attend AMIA 2018 Annual Symposium and participate in this panel.

Acknowledgement
HK was supported by the intramural research program at the U.S. National Library of Medicine, NIH.

References