10 years of reproducibility in biomedical research: how can we achieve generalizability and fairness?

Adina Svenja Wagner, Camille Maumet, Melanie Ganz, Cassandra Gould van Praag

▶ **To cite this version:**

Adina Svenja Wagner, Camille Maumet, Melanie Ganz, Cassandra Gould van Praag. 10 years of reproducibility in biomedical research: how can we achieve generalizability and fairness?. ISBI 2023 - 20th IEEE International Symposium on Biomedical Imaging, Apr 2023, Cartagena de Indias, Colombia. , pp.3. hal-04017085

HAL Id: hal-04017085
https://inria.hal.science/hal-04017085

Submitted on 9 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

10 years of reproducibility in biomedical research: **How can we achieve generalizability and fairness?**

Abstract 10 years ago, a series of publications pointed to the difficulty of reproducing scientific findings. This r*eproducibility crisis* was a wake-up call for scientific communities to rethink how we practice and communicate research, and an important driver towards greater transparency and robust results. Ever since, biomedical imaging undertook various efforts to overcome reproducibility issues: From increasing sample sizes for higher statistical power, to data sharing and increased collaborations to acquire such samples, and promoting detailed reporting practices and code sharing to ease computational reproducibility.

But where are we standing with respect to reproducible biomedical imaging now? We discuss recent advances and open questions, and focus on how the conversation has moved beyond efforts to reduce false positive findings to broader questions of generalizability and fairness. How does a finding observed in a given group apply to the population at large? How does a finding obtained with one analysis vary when computed using another tool? How does a finding observed in a given group apply to subgroups of that population, in particular to less represented subgroups? How can open science help with the complex questions of building fair algorithms and fairness in who participates in the process of science?

Introduction

About 10 years ago, a series of publications pointed to the difficulty of reproducing published scientific findings. Evidence of this phenomenon – later referred to as the *reproducibility crisis* – was identified in many biomedical fields including (to name only a few): cancer research, psychology and brain imaging. Overall, realisation of this phenomenon has been a wake-up call for scientific communities to rethink the way we practise and communicate our research and was an important driver towards greater transparency and the production of robust results (often described as "open science").

In the field of biomedical imaging, the early efforts to overcome issues of reproducibility focused on increasing sample sizes in order to achieve higher statistical power. This necessitated data sharing and increased collaboration between labs. In addition, more detailed reporting practices and code sharing were also promoted to ease computational reproducibility (i.e. the ability to reproduce a result using the same data and the exact same protocol and environment).

10 years later, where are we standing with respect to reproducible biomedical imaging? This session will discuss recent advances and open questions in this topic. In particular, we will focus on how the conversation has moved beyond efforts to reduce false positive findings, and on to broader questions of generalizability and fairness.

> How does a finding observed in a given group of participants apply to the population at large? Working with larger datasets has created a shift in the way biomedical analyses can be computed. With more data, coming from external sources, we need new ways to ensure computational reproducibility.

How does a finding obtained with a given analysis vary when computed using another tool or algorithm? The approaches available to analyse a dataset have multiplied and growing evidence is showing that the exact choice of pipeline can impact scientific findings.

How does a finding observed in a given group of participants apply to subgroups of that population and in particular to subgroups that are less represented? While the question of fairness in AI is a topic of acute debate, we are just starting to understand how fairness impacts biomedical findings.

Open science is often described as a must-do to increase reproducibility. But beyond increased statistical power and transparency brought by opening up research artefacts (e.g. open data, open code) how can open science help with the more complex question of building fair algorithms and fairness in who participates in the process of science and who sets the research agenda?

The topics of open science and reproducibility that are at the centre of this session are relevant to the ISBI community at large.

Program

Talk 1. **Towards computational reproducibility when working with very large datasets** by **Adina Wagner**

Psychoinformatics Lab, INM-7 at Juelich Research Centre, Germany.
https://www.adina-wagner.com/

Abstract: The amount of data available to researchers has steadily grown, but over the past decade, a focus on diverse, representative samples has resulted in datasets of unprecedented size. This development is accompanied by a growing awareness of the importance of making the data more findable, accessible, interoperable, and reusable (FAIR), and increasing availability of research standards and tools that facilitate data sharing and management. But large-scale datasets nevertheless constitute challenges. Storage and computational demands easily exceed the capabilities of consumer-grade hardware, and increasingly exceed even those of research institutions infrastructure. The growing complexity of handling large scale datasets makes the resulting large-scale computations more difficult to reproduce, comprehend, and verify, ultimately putting the trustworthiness of derivative data at stake. Together with increasingly complex data analyses, it requires methodological knowledge, but also skills in data and software management, to put open, reproducible, and scalable neuroimaging research into effect. In this talk, I will share the technical challenges we faced when processing the largest neuroimaging datasets reproducibly, and the solutions we put into practice.

Talk 2. **Different analysis pipelines, different results… How to deal with analytical variability?** by **Camille Maumet**

Empenn team at Inria, Univ Rennes, CNRS, Inserm, Rennes, France.
http://camillemaumet.com/

Abstract: Data processing is central to biomedical imaging and over the years the different tools and algorithms available to study a dataset have multiplied. Recently, multiple studies in the literature have demonstrated how the exact choice of analysis pipeline can have an impact on the conclusions of scientific experiments. Those studies focused on varying different components of the analysis pipeline including differences observed across scientific software, across preprocessing pipelines or even across different operating systems. In addition, the development of many-analyst studies -- in which independent teams of experts are tasked to answer the same research questions using the same input data -- have shed lights into how much differences can be expected across pipelines selected by experts from the scientific community. In this talk, we review different examples showcasing this analytical variability. We will reflect on the underlying

reasons that can explain the existence of this variability. We will end by exploring current solutions and in particular how the recording of provenance can help.

Talk 3. **Fairness in biomedical imaging: overview and a case study in neuroimaging** by **Melanie Ganz**

Neurobiology Research Unit at Rigshospitalet, Copenhagen and Department for Computer Science at the University of Copenhagen, Denmark.
https://sites.google.com/view/melanieganz

Abstract: In my talk I will introduce the audience to the broader issue of fairness in biomedical imaging applications. I will do this by going through a recent case study (https://arxiv.org/abs/2204.01737) and then summarising other results and efforts to address the issue of fairness. The case study focuses on the application of modern machine learning algorithms in the application of Alzheimer's disease (AD) diagnosis on the basis of the openly available data from the Alzheimer's disease neuroimaging initiative (ADNI).

Convolutional neural networks (CNN) have enabled significant improvements in medical image-based diagnosis. It is, however, increasingly clear that these models are susceptible to performance degradation when facing spurious correlations and dataset shift, leading, e.g., to underperformance on underrepresented patient groups. In this paper, we compare two classification schemes on the ADNI MRI dataset: a simple logistic regression model using manually selected volumetric features, and a convolutional neural network trained on 3D MRI data. We assess the robustness of the trained models in the face of varying dataset splits, training set sex composition, and stage of disease. In contrast to earlier work in other imaging modalities, we do not observe a clear pattern of improved model performance for the majority group in the training dataset. Instead, while logistic regression is fully robust to dataset composition, we find that CNN performance is generally improved for both male and female subjects when including more female subjects in the training dataset. We hypothesise that this might be due to inherent differences in the pathology of the two sexes. Moreover, in our analysis, the logistic regression model outperforms the 3D CNN, emphasising the utility of manual feature specification based on prior knowledge, and the need for more robust automatic feature selection.

Talk 4. **How does inclusivity support fairness in biomedical imaging?** by **Cassandra Gould van Praag**

Department of Psychiatry at University of Oxford, UK.
https://www.psych.ox.ac.uk/team/cassandra-gould-van-praag

Abstract: Advocates for open science are often (but not always) advocates for inclusivity. Is this coincidence, or does orientation to these two issues arise from a common values base? I will argue that there is both a rational and moral motivation for improving the inclusivity of your research practice, and both lead to more robust and generalisable findings. Participants will be given a tour of the key terminology in the practice of inclusivity, asked to consider where they have experienced the negative impact of power and dominance hierarchies, and consider where they have the opportunity to redress the balance. I will also discuss the practice of "decolonisation" of research, where we consider how our work is influenced by extractive, capitalistic drives which have become entrenched as a global Eurocentric bias. Participants will be presented with the impact of this colonial practice and invited to consider where they can decolonise their work, if they agree that they have a moral obligation to do so.