The need of standardised metadata to encode causal relationships: Towards safer data-driven machine learning biological solutions

Beatriz Garcia Santa Cruz[1,2,*], Carlos Vega[2], Frank Hertel[1,2]

[1] National Department of Neurosurgery, Centre Hospitalier de Luxembourg
[2] Luxembourg Centre for Systems Biomedicine, University of Luxembourg

*corresponding author: garciasantacruz.beatriz@gmail.com 0000-0002-0939-4443

Keywords: confounders, causality, metadata, machine learning, systems biology.

Abstract

In this paper, we discuss the importance of considering causal relations in the development of machine learning solutions to prevent factors hampering the robustness and generalisation capacity of the models, such as induced biases. This issue often arises when the algorithm decision is affected by confounding factors. In this work, we argue that the integration of causal relationships can identify potential confounders. We call for standardised meta-information practices as a crucial step for proper machine learning solutions development, validation, and data sharing. Such practices include detailing the dataset generation process, aiming for automatic integration of causal relationships.

1 Introduction

The number of scientific publications in the biological field employing machine learning (ML) is rapidly growing. Both as a result of better access to larger amounts of data generated using the latest technology (e.g. high throughput screening) and the computational capacity together with the faster development in the ML area, especially in deep learning (DL). Such trend and its direct consequences in biological healthcare applications call for standardised guidelines to ensure the quality of each stage of the research and application pipelines. Among other objectives, these guidelines aim to establish better data sharing and appropriate foundations for good appraisal and reproducibility. The data sharing goal aims to ensure good data management not only to advance in knowledge discovery and innovation but also to allow for proper data reuse. Better appraisal and reproducibility can be achieved through standardised reporting guidelines that guarantee the report of key dataset elements (for example dataset generation details) as an essential step for dataset comparison and validation. In this way, remarkable efforts have been done in the recent years. Here we highlight the FAIR principles [1] and the DOME recommendations [2].

The FAIR principles (Findability, Accessibility, Interoperability, and Reusability) aim to increase data usability, with special emphasis on machine-readable and actionable datasets. This need arises because machines, in contrast to humans, lack a natural ability to identify and interpret the context, becoming more likely to make errors contextualising data. However, machines can overcome humans' main limitations operating at the scope, scale, and speed that the current e-Science scenario requests. Thus, different mechanisms and protocols seeking machine self-guidance for data exploration need to be developed [1]. More specific to the area of applied ML for biological analysis we can find DOME (Data, Optimization, Model and Evaluation), a collection of recommendations focused on proper reporting of supervised ML in biological studies [2].

The rest of the paper is structured as follows: with FAIR and DOME as guidance, we discuss the importance of considering potential confounders in the data, especially after the paradigm change from classical statistical modelling to ML. Below, we explore its impact on the ML-based biological applications emphasising human disease biology. Finally, we discuss potential solutions with a special focus on the standardised metadata that aims to encode causal relationships. Although every step is critical to develop better models, this paper focuses on the datasets and their standardised reporting.

2 Considerations for the development and reporting of ML solutions

The study of biological systems involves either inferences or predictions. Inference aims to create a mathematical model about the data-generation process, testing a hypothesis or formalising our understanding of how the studied systems behave. It is used to understand the mechanism of the studied event, for instance, how the accumulation of one specific protein affects the system. In contrast, prediction purpose is to forecast future behaviour, without necessarily understanding the mechanism behind it. For example, to predict which treatment is better based on the specific level of a determined protein. Despite both statistics and ML can be used to predict and make inferences, traditionally, statistical methods have focused on inference and ML in prediction [3]. The choice between prediction or inference depends on the ultimate analysis goal.

In short, the statistical approaches are useful when we want to understand each variable influence; but in general (not always), have less prediction power, frequently because only few variables and linear relationships are considered. Conversely, when large and high dimensional datasets are analysed with prediction as a goal, ML is chosen [3]. One of the most crucial handicaps in current ML is the lack of *model traceability*. However, the high prediction power of these methods promotes their use for biological applications. Although, such methods are not exempt from risks.

Current limitations in biomedical ML solutions. The number of scientific works using ML techniques has increased during the last years exponentially, and it is progressively translating into real-world applications, including High-Stakes domains such as health, conservation, employment, education or justice. High-Stakes AI domains are characterised by their significance and lasting impact on both individuals and society [4]. Unfortunately, despite these models report excellent results during their model training and testing steps, there are notorious cases when the accuracy dropped significantly during their real application, in some cases with harmful repercussions when happened in High-Stakes domains. Systematic errors have been reported in commercial software employing ML models for tracking, face detection, criminal justice and hiring recommendations. Such errors include systematic biases against concrete populations [5] and limitations in model generalizability and transportability which are well-known issues in biomedical applications [6].

Origin and errors types. The reasons and solutions are complex, but one of the most notable factors is the dataset composition, including the consideration and intervention to avoid undesirable biases and potential confounders. Simply put, the main error sources fall into two subtypes of biases. First, ***conveyance of systematic biases in the datasets***. These patterns represent actual real-world bias that we do not want to convey in the data. For instance, a dataset may reflect an unfair systematic historical discrimination against a particular group of people that may undesirably be perpetuated or even amplified if is not controlled for [5]. Second, ***biases induced during the collection, annotation, preprocessing, and learning strategies***. In this case, biases arise from one or multiple steps of the data pipeline. For example, the data collection process might be biased, or the training process may wrongly use features, producing a biased model [7]. Although the origin and impact of both type of errors is different, the solution to both involves improving model *traceability* in different ways for a better understanding of the model's decisions. Efforts on this direction are applied in the whole MLOps pipeline, from data collection, modelling and post model evaluation.

This paper focuses on the induced biases as they are an important concern in biolog-

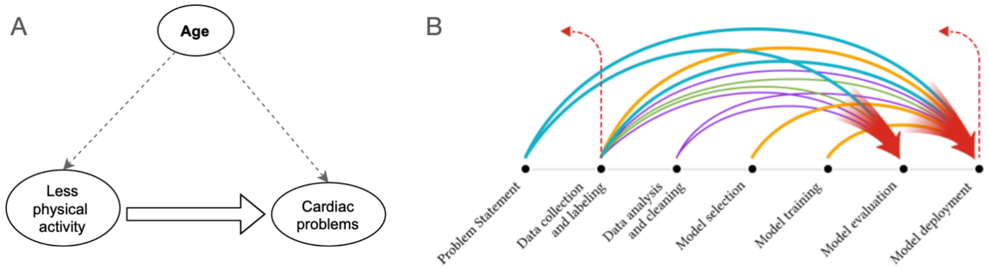

Figure 1: Panel A: Causal diagram depicting a confounding factor (age) acting (dashed lines) in both the predictor and outcome. Panel B: Diagram depicting data cascades extracted from [4]. Thick red arrows show the point where compounding effects become visible; dotted red arrows depict abandoning or restarting of the ML data process. Coloured lines show different cascades: Interacting with physical world brittleness (yellow); Inadequate application domain expertise (blue); Conflicting reward systems (green); Poor cross-organisational documentation (purple).

ical studies. A special focus is put on dataset documentation as a way to reduce the risk of such undesirable biases.

Limitations of the current evaluation system. At this point, one may wonder how models suffering from such issues could satisfy all the requirements needed for their deployment. This issue is explained to a large extent because the metrics currently employed by practitioners (for instance, F1, accuracy, AUC, Matthews correlation coefficient) assess the goodness of the model fitting the data but do not express the phenomenological fidelity and validity of the data. The phenomenological fidelity refers to the representation of the modelled phenomena, while the validity of the data indicates how well the data captures the phenomena in order to explain it [4]. Current applications measure how good the models perform in the test data, which generally is a subdivision of the same dataset or a dataset collected under similar conditions. Such a score does not express how well the model captures the behaviour of the real phenomena, for which data is just an approximate representation of reality. For instance, they do not measure whether all the event variations are considered or if the capturing methods have enough sensitivity. Moreover, the 'black box' nature of current ML models hinders transparency regarding the features or combination of features employed during predictions. While common ML safe practices like cross-validation or class imbalance control aim to minimise model issues such as model over-fitting, their use draws from the premise that data is a solid representation of the modelled phenomena. Such practices cannot overcome data collection issues, leading to poor consideration of the dataset harvest and documentation work. These practices have been proven to impact ML projects. For instance, decreasing the accuracy of IBM's cancer treatment AI solution and causing that Google Flu Trends underestimates a flu peak by 140% [4]. In absence of context, models may use undesirable bias or confounders during the training.

Towards potential solutions. The first step to avoid, or detect, potential bias in datasets and models is to improve the documentation of the dataset generation process. One of the most common sources of induced bias includes unknown confounders as well as selection, acquisition, and annotation biases. As shown in Figure 1 panel A, confounders (age) are variables that affect both the potential predictor variable (physical activity) and the outcome (cardiac problems). When the presence of confounders is unknown and in lack of experiments specifically designed to minimise them (for example randomised controlled trials), we cannot control for them. Uncontrolled confounders lead to conclude that a given feature may be a strong predictor of the outcome when in reality the association is spurious and it does not hold anymore when the sample comes from a different setting where the confounder is differently expressed. When the model learns spurious associations between predictors and outcomes, an undetected overfitted model is produced, resulting in poor generalization capabilities that eventually unveil during its translation into real-world settings [6].

Recently, the concept of data cascades (DC) was presented as one of the main issues in the current life-cycle of an ML systems. DC are compounding events provoking negative downstream impact from data issues causing a technical debt over time. DC describes and identifies how induced biases are generated during the design and data collection process: From the problem statement, dataset collection, data labelling, data analysis and cleaning; as well as model selection, model training, model evaluation until model deployment. This is represented in Figure 1 panel B.

Before ML-based analysis expansion, researchers generally employed statistical modelling of biological processes to make inferences from observational data. Generally, in statistical modelling, there is a tight control of potential confounders. This close analysis allows to include functional assumptions that affect the relationships between variables. Bearing all the previous concerns, intervention in the developing and reporting systems of ML-based solutions must be addressed before its translation into real-world settings.

3 Relevance of induced bias in biological studies for ML analysis

There is a large diversity in the biological ML applications, particularly in the areas focused on the understanding the biological process that underlines the human diseases. Such studies aim to understand the core biological processes like transcription, translations, signalling or metabolism, including tissues and organs. The current size and precision of omics data opens the door for insight modelling using among others ML techniques. From gene expression, splicing, single-cell data to neuroscience [8].

The batch effect (BE) is a known source of confounders in the area of biology. The BE refers to the different factors when comparing sample lots that affect the measurements masking the biological variation impact. BE is the consequence of different laboratory conditions, reagent lots, machine calibration, software and even personnel differences. For example, a strong laboratory-specific effect has been reported when comparing multiple micro-array experiments. Another example concerns gene expression studies, in which large variations are associated with the data processing and the specific settings of micro-array work. Consequently, several papers, including relevant studies published in high-impact journals, were retracted [9] on the basis of such errors.

This is usually addressed during the experimental design thanks to randomisation, stratification, replications and inclusion of both positive and negative controls. However, dataset reuse and dataset mix (often produced with different settings) may impede controlling for such factors. Therefore, this is a lurking problem in biological data analysis and ML solutions employing mixed high-throughput datasets. Finally, although BE often focus on the preparation and measure conditions of the samples, other induced bias may arise in the subsequent pipeline steps in the form of data cascades.

As depicted in Figure 2, there are five general steps from which these potential induced biased can arise. In first place, the biological source. For instance, variations in population, disease penetrance, phenotypic manifestation, environmental conditions or sample techniques may induce biases during human sampling. Next, human or machine sample preparation may be sensitive to the machines employed, reactants, protocol settings and in-house calibration. In the same line, different settings conditions may affect the signal measurement. Then data analysis is conditioned by the approach employed for data cleaning, normalisation and labelling. Finally, data sharing is not exempt from issues if decisions are taken to modify or remove features before the data distribution.

4 Potential solutions: Metadata Standardisation

Open science and open innovation allow fulfilling a basic principle of science, reproducibility. The main principles include open code, open data and open publications. But while open data allows reproducing the reported results, ensuring data reusability entails proper description of the whole generation process. At this point, it is clear that

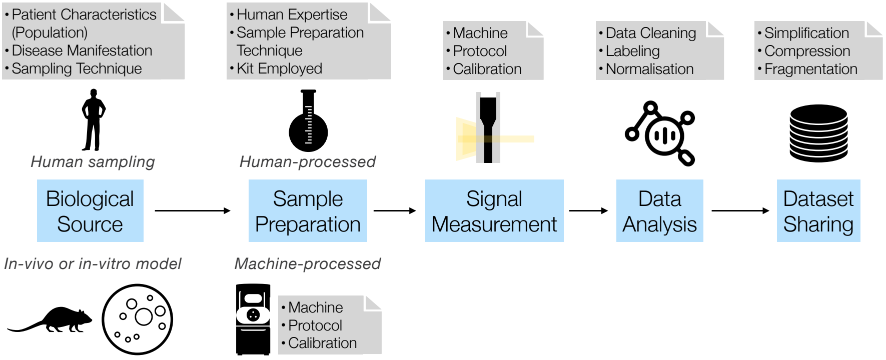

Figure 2: General workflow diagram showcasing the generation process of biological datasets. Gray boxes contain potential sources of induced bias.

datasets must be paired with proper documentation and accompanied with additional information, indeed it is one of the principles of relevant guidelines such as DOME [2]. However, encoding such information together with the dataset is not always possible, often requiring sidecar files better suited to express dataset properties differently. Metadata has emerged as a crucial component for reproducibility in the research life cycle [10]. Additionally, the full potential of metadata is still open to unexplored opportunities associated with the area of biological ML [2]. FAIR principles reflect the need of reusability and interoperability, suggesting extensive documentation to satisfy data management and stewardship needs. Similarly, DOME guidelines aim for proper data provenance and safe model evaluation. However, such principles do not demand further metadata encoding the causal assumptions made during the data collection process and the intentions of the original study for which it was collected.

The final aim of such metadata should convey the generation process enabling its comparison across datasets to identify differences in the generation process and inform of potential induced biases as the first step of its control. To ensure its correct comparison the metadata should be standardised. The metadata standardisation is already present in domain-specific repositories such as Genbank or UniProt which are highly curated and include specific metadata. However, general-purpose solutions are still missing [1]. Other data and metadata standard include DICOM (Digital Imaging and Communications in Medicine), FHIR (Fast Healthcare Interoperability Resources), Functional Annotation of ANimal Genomes (FAANG) and Observational Health Data Sciences and Informatics (OHDSI). This information may also include causal graphs representing the assumptions considered during the data collection process. Such additional information could prevent issues in which the modeller is unaware of known confounders between the variables or samples, with unexpected consequences. For example, an inappropriate split may break the assumption that data is independent and identically distributed, in other words, that all samples stem from the same generative process which has no memory of past generated samples. For instance, a medical dataset containing multiple samples from the same patients without stating the patient id (or another patient dependent variable), precluding group-wise division of the dataset. In this example, the training process could be compromised due to potential data leakage caused by the presence of samples from the same patient in both the train and test sets. Similarly, if the samples stem from a time-dependent process, a time-wise scheme is at hand. In any case, such data generative process must be properly documented beforehand. In short, standardised metadata may include the following principles. **Interoperability**: through automatic data generation. **Usability**: easy-to use integration when human input in required. **Adherence**: provide an interface to which general and domain-specific standards may adhere. **Integrative**: employ already existing guidelines. **Privacy**: provide features to comply with current data protection (such as GDPR).

5 Conclusion

In this paper we present the general issues affecting ML solutions for High-Stakes domains, such as biomedical research, caused by induced biases derived from the collection, annotation and preprocessing stages of the ML model production pipeline. In particular, we address the lack of information context which enables ML models to learn spurious associations between variables that might be affected by endogenous confounding factors derived from the particularities of the pipeline setting (such as instrument noise, laboratory protocols). This issue is amplified in high-throughput scenarios when comparing sample lots generated in different settings. The solution involves bringing data work to the foreground of the ML model production pipeline. Increased domain knowledge, data excellence incentives and improved feedback channels in the AI data life-cycle are good starting points [4]. However, such goals must be translated into material actions. In this paper we propose increased documentation of the dataset generation process as an essential safety practice. This includes the use of dataset-wise standardised metadata and incorporation of causal relationship information regarding dataset variables. The former guideline is suited for the particularities of biological datasets and aimed at easing dataset mixture and comparison. The latter practice would prevent confounding effects by encoding the assumptions concerning the dataset generation process. The production of standardised documentation and metadata may not only ease the data work but also open the door for actionable-metadata and incorporation of causal relationships during the model training. We believe these strategies will help mitigating potential risks of ML solutions in real-world scenarios.

Acknowledgments

The authors would like to thank Andreas Husch and Matias Bossa for their support and review efforts. BGS work is supported by the FNR-PRIDE17/12244779/PARK-QC and Pelican award from the Fondation du Pelican de Mie et Pierre Hippert-Faber.

References

[1] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J. ”The FAIR Guiding Principles for scientific data management and stewardship” *Scientific data*, v 3(1), pp.1-9

[2] Walsh, Ian and Fishman, Dmytro and Garcia-Gasulla, Dario and Titma, Tiina and Pollastri, Gianluca and Harrow, Jennifer and Psomopoulos, Fotis E and Tosatto, Silvio CE. ”DOME: recommendations for supervised machine learning validation in biology” *Nature Methods*, pp. 1-6, 2021.

[3] Danilo Bzdok, Naomi Altman and Martin Krzywinski ”Statistics versus machine learning” *Nature methods*, v. 15, n.4, pp. 223, 2018

[4] Sambasivan, Nithya and Kapania, Shivani and Highfill, Hannah and Akrong, Diana and Paritosh, Praveen and Aroyo, Lora M. ”Everyone wants to do the model work, not the data work”. *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1-15, 2021.

[5] Mitchell, Margaret and Wu, Simone and Zaldivar, Andrew and Barnes, Parker and Vasserman, Lucy and Hutchinson, Ben and Spitzer, Elena and Raji, Inioluwa Deborah and Gebru, Timnit ”Model cards for model reporting” *Proceedings of the conference on fairness, accountability, and transparency* pp. 220-220, 2019.

[6] Garcia Santa Cruz , Beatriz and Bossa, Matias Nicolas and Soelter, Jan and Husch, Andreas Domink ”Public Covid-19 X-ray datasets and their impact on model bias-a systematic review of a significant problem”. *Medical Image Analysis* 2021.

[7] Castro, Daniel C and Walker, Ian and Glocker, Ben ”Causality matters in medical imaging” *Nature Communications* v.11, n. 1, pp. 1-10, 2020.

[8] Ching, Travers and Himmelstein, Daniel S and Beaulieu-Jones, Brett K and Kalinin, Alexandr A and Do, Brian T and Way, Gregory P and Ferrero, Enrico and Agapow, Paul-Michael and Zietz, Michael and Hoffman, Michael M and others. ”Opportunities and obstacles for deep learning in biology and medicine”. *Journal of The Royal Society Interface*, vol.15, pp. 20170387, 2018.

[9] Leek, Jeffrey T and Scharpf, Robert B and Bravo, Héctor Corrada and Simcha, David and Langmead, Benjamin and Johnson, W Evan and Geman, Donald and Baggerly, Keith and Irizarry, Rafael A ”Tackling the widespread and critical impact of batch effects in high-throughput data” *Nature Reviews Genetics* v.11, n.10, pp. 730–739, 2010.

[10] Leipzig, Jeremy and Nüst, Daniel and Hoyt, Charles Tapley and Ram, Karthik and Greenberg, Jane ”The role of metadata in reproducible computational research” *Patterns* v.2, n.9, pp. 100322