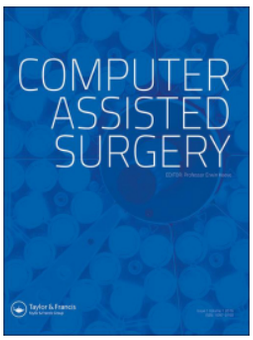
Bias in machine learning for computer-assisted surgery and medical image processing

John S. H. Baxter, Pierre Jannin



Computer Assisted Surgery

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/icsu21

Bias in machine learning for computer-assisted surgery and medical image processing

John S. H. Baxter & Pierre Jannin

To cite this article: John S. H. Baxter & Pierre Jannin (2022) Bias in machine learning for computer-assisted surgery and medical image processing, Computer Assisted Surgery, 27:1, 1-3, DOI: 10.1080/24699322.2021.2013619

To link to this article: https://doi.org/10.1080/24699322.2021.2013619

Published online: 08 Feb 2022.

COMPUTER ASSISTED SURGERY
2022, VOL. 27, NO. 1, 1–3
https://doi.org/10.1080/24699322.2021.2013619

Taylor & Francis
Taylor & Francis Group

EDITORIAL

Bias in machine learning for computer-assisted surgery and medical image processing

Members of the academic community would undoubtedly say that machine learning has changed the face of research in computer-assisted surgery and medical image processing and many laud this change. For some, the rise of new high-performance techniques implies that untold barriers of accuracy and efficiency are about to be demolished. For others, a degree of skepticism is shown toward techniques that claim results that, until recently, would have been considered patently ridiculous, resulting in a new light being shed on traditional aspects of algorithm verification (such as evaluation metrics [1,2] and experimental design [3] that influence the reported performance of these techniques.

Throughout all of this, *bias* in machine learning has become something of a watchword, a response to a myriad of problems currently seen in the literature regarding how machine learning is used in computer-assisted surgery and medical image processing. (Note that here we are discussing *biases in the evaluation of a model* rather than biases in the model's predictions themselves, although they are often related.) Many in the skeptical camp latch upon these four letters as foundational to their critiques. Upon further investigation, however, the concept of bias itself is not as a monolithic whole, but as a collection of inter-related considerations. For the moment, we will split these considerations into two types: *human-centric elements of bias* and *methodology-centric elements of bias*.

The human-centric elements of bias have permeated more widely into the general population's perception of machine learning. Racial, gender, age, and socio-economic biases, in particular, are indeed still issues. These biases arise from an under-, over-, or mis-representation of a particular group, possibly in the data itself, possibly in the applications chosen by researchers, and possibly in the composition of the researcher community. Building awareness and addressing these biases is highly important, and there are many voices more-qualified than mine dedicated to doing exactly that.

However, there are less-discussed methodology-centric elements (given in Table 1) also present in the literature. These biases concern largely whether or not the results of a paper are representative of their actual clinical context and arise from the methodology of the experiment itself in ways that may be stated or unstated. In addition, some of these biases overlap, such as a particular lack of representation (e.g. if images of a particular type of patient are annotated in a particular way, such was the case for several COVID19 detection models [4] leading to feature leakage (e.g. finding this annotation determines the type of patient). Some of these biases appear to be a fundamental component of science, at least in a Lakatosian sense [5]; science either relies on them to advance (i.e. incrementally improving the best-known models) or progressively identifies and models them (i.e. measuring how methods perform on different data and understanding those differences).

To take an example of the former, a degree of *model selection bias* is necessary to advance science in our field or at least is unavoidable due to the nature of science as a human endeavor. Every time one looks at the literature in order to narrow what models or hyper-parameters to use, one is fundamentally introducing a bias. Similarly, with the lack of papers displaying negative results or sub-par performance, the literature as a whole presents a much rosier and more optimistic situation for research prototypes than the clinic shows for well-validated systems. *Metric/ranking selection bias* is a large but necessary problem as metrics are necessary for interpretation but can be highly sensitive and opaque [6].

For the latter, consider *distribution biases*. With the exception of already identified elements, *unrepresentativeness biases* are by definition epistemic: one cannot know if the data that they have collected will be representative of clinics in general with respect to factors that have yet to be described. Although elements of this can be theoretically minimized by collecting more data, this is a relatively passive bias reduction strategy, as opposed to the active strategy of identifying and controlling these factors, but neither is completely certain to eliminate bias altogether. Some distribution biases related to data selection, even when they are identified, are epistemologically impossible to reduce, for example, the impossibility of measuring the results of different, mutually exclusive surgeries on the exact same patient. Others, such as *temporal/causal shifts in the distribution* even imply that any individual result would be different *at the current time* than it was at the time of the study, implying that some studies should come with an expiration date. A possible source for these biases is the way the annotations themselves are generated as the datasets generally only contain data where it is feasible to have annotations. In addition, this level of feasibility and accuracy change over time with different annotation and visualization tools [7] as well as being dependant on the experience of the annotators themselves [8]. Additionally, annotations can introduce

Table 1. Some examples of methodology-centric elements of bias and their definitions.

Distribution biases

Unrepresentativeness: when the distribution of the evaluation data used in the paper differs from that observed in the clinic, due to the influence of human-centric elements of bias or other (possibly unknown) factors

Data selection bias: when the dataset itself captures only a portion of the relevant distribution due to experimental design, technology, or feasibility reasons

Annotation bias: when the annotations themselves include a consistent bias or source of error, such as simplification or differences between the in-plane and out-of-plane quality

Temporal/causal shift: when the distribution itself changes over time, possibly due to the influence of the system under investigation

Methodology selection biases

Model selection bias: when the model with the highest measured performance is selected from a group of measured models and said measurement performance is directly reported (i.e. no independent evaluation)

Metric/ranking selection bias: differences between the metrics used to evaluate a framework and the underlying ill-defined notion of clinical quality (an example of this is how one deals with class imbalance)

Data leakage

Feature leakage: when certain features are provided as input to an algorithm when they would be unavailable in practice

- *temporal leakage:* when information that is normally not available until after the machine learning task is provided as input for said task
- *correlation leakage:* when a feature is irrelevantly correlated with the prediction in the experiment but not in intended use

Train/test leakage: information from the evaluation (i.e. testing) dataset is used at least partially to develop the models which are then evaluated on said dataset

- *normalization/imputation/augmentation leakage:* when knowledge of the evaluation dataset is used to develop algorithms besides the primary one
- *stacking/boosting/parameter leakage:* when the evaluation set is used in the process of combining or selecting from multiple models or configurations

other biases due to systematic biases in the annotations themselves, such as prioritizing smoothness only in the in-plane directions due to disagreement between a single user's segmentation across consecutive slices. Although this bias can be heuristically measured when there are multiple annotators, this is generally only considered to put an upper bound on the performance, rather than be interpreted as a negative bias [3].

In contrast to these more high-level and loosely defined biases, *data leakage* represents a collection of biases that are more grounded, less epistemic, and thus easier for us as a community to address. In fact, it is possible, and often still practical, to eliminate some in their entirety as they are simply showing *the wrong information at the wrong time*. For example, to avoid *temporal leakage* and *correlation leakage*, one must only ensure that the data provided is clearly separated into what can be collected prior to the actual diagnosis or image-processing step and what cannot, and for this separation to be validated by a clinical collaborator or in the clinic itself. (For a more detailed look at the various kinds of feature leakage in a more general context, see ([9], Chapter 24). Eliminating feature leakage in its entirely depends heavily on a functioning researcher-clinician relationship and a strong understanding of the clinical workflow and problem domain, something we should come to expect of our research community and, to our credit, is often the case.

However, *train/test leakage* is more insidious and some studies have shown it to have a large effect on the medical imaging and computer-assisted surgery literature, thus going unnoticed in the review process [10,11] both investigate this in two different applications, for example). Simply put, train/test leakage is when information from the test set is provided in training time, although the simplicity of this formula denies the complexity of the problem as different data points themselves are correlated and provide information about each other. To give an example, several medical image computing datasets include multiple images of the same patient across different time points or even just two-dimensional slices from the same three-dimensional volume. (One reason why these types of errors are so insidious is the complexity of some evaluation methods, such as cross-validation and nested cross-validation, where the difference between a correct and a leaking implementation can be difficult to detect for both authors and reviewers.) There are obvious correlations here that could be leaked if images from the same patient found themselves in both the training and the evaluation datasets at the same time. However, this is also the case with particular hospital centers. Should multi-center experiments require all of the images from each center to fall on the same side of the training/evaluation divide?

In our field of study, there is no simple answer to these questions as bias is not a matter of type but of degree. Similarly, for algorithms adjacent to the primary machine learning task, (such as data normalization or detecting invalid input,) some small amounts of leakage might have negligible detrimental effects, but also non-negligible positive effects. For example, examining the entire dataset to see what constitutes an artifact or a missing value (e.g. is a default value used? a NaN?) is technically leakage as you would also be examining the evaluation dataset. However, the positive gain from this (i.e. the removal of elements that would otherwise break the algorithm or the design of non-intelligent methods to detect these invalid inputs in the clinic) would be immense, although this is a manner

of degree in terms of the methods being implemented and the knowledge is extracted.

One reaction to this would be complete isolation, that researchers identify and remove all possible elements of bias, and this would be a valid response if these biases were a matter of *type* and not of *degree*. For example, in some contexts, it may be appropriate to ensure that all data arising from one hospital center is kept entirely separate from another. In others, the bias introduced by mixing centers in the dataset is not only negligible but irrelevant and damaging. For some research, we benefit from researchers building upon others methods using the same open datasets. In others, the improvement garnered reflects community-level overfitting more than real progress.

In my opinion, we should be unequivocal in asking authors to more carefully perform experiments and for reviewers to more carefully examine papers, cognizant of different biases. However, I don't think we should dogmatically call for their elimination. Instead, we should look for *transparency* and *justification*. Instead of eliminating all bias root-and-stem at the risk of suppressing research and meaningful contributions, we should critically question and make explicit to what degree the different biases exist and their effect. We should justify the level of bias and make them transparent for future research. We will likely find a certain level of bias to be unacceptable. (in my opinion and in the opinion of many others [10,11], mixing data from a single patient across the training/testing divide requires an almost insurmountable level of justification) but this can only come about through a more open and nuanced conversation about bias. Thus, it would be a matter of informed judgment about whether or not a paper should be accepted and its results to be believed. This is a level of judgment that we should, as members of a mature research community, demand of ourselves.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

[1] Mason A, Rioux J, Clarke SE, et al. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images. IEEE Trans Med Imaging. 2020; 39(4):1064–1072.

[2] Reinke A, Eisenmann M, Tizabi MD, et al. 2021. "Common limitations of image processing metrics: A picture story." *arXiv preprint arXiv:2104.05642*.

[3] Maier-Hein L, Reinke A, Kozubek M, et al. BIAS: transparent reporting of biomedical image analysis challenges. Med Image Anal. 2020;66:101796.

[4] DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021; 3:610–619.

[5] Lakatos I. The methodology of scientific research programmes. Philosophical Papers. 1987;1:135.

[6] Maier-Hein L, Eisenmann M, Reinke A, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun. 2018;9(1):1–13.

[7] Duncan D, Garner R, Zrantchev I, et al. Using virtual reality to improve performance and user experience in manual correction of MRI segmentation errors by non-experts. J Digit Imaging. 2019;32(1):97–104.

[8] Kohlberger T, Singh V, Alvino C, et al. 2012. Evaluating segmentation error without ground truth. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer.p. 528–536.

[9] Larsen KR, Becker DS. 2021. Automated machine learning for business. Oxford: Oxford University Press.

[10] Samala RK, Chan H-P, Hadjiiski L, et al. 2020. Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks. In: *Medical Imaging 2020: Computer-Aided Diagnosis*, Vol. 11314, International Society for Optics and Photonics. p. 1131416.

[11] Yagis E, Workalemahu Atnafu S, García Seco de Herrera A, et al. Deep learning in brain MRI: Effect of data leakage due to slice-level split using 2D convolutional neural networks. Sci Rep. 2021;11(1–23).

John S. H. Baxter and Pierre Jannin
Laboratoire Traitement du Signal et de l'Image (LTSI - INSERM UMR 1099), Universite de Rennes 1, Rennes, France
jbaxter@univ-rennes1.fr