Heart Metab. 2020;82:33-35

Avoiding big data pitfalls

Pablo Lamata, PhD

Department of Biomedical Engineering, King's College London, UK

Correspondence: Pablo Lamata, PhD, Dept of Biomedical Engineering – 5th floor Becket House, 1 Lambeth Palace Road, London SE1 7EU, UK
E-mail: Pablo.Lamata@kcl.ac.uk

Abstract: Clinical decisions are based on a combination of inductive inference built on experience (ie, statistical models) and on deductions provided by our understanding of the workings of the cardiovascular system (ie, mechanistic models). In a similar way, computers can be used to discover new hidden patterns in the (big) data and to make predictions based on our knowledge of physiology or physics. Surprisingly, unlike humans through history, computers seldom combine inductive and deductive processes. An explosion of expectations surrounds the computer's inductive method, fueled by the "big data" and popular trends. This article reviews the risks and potential pitfalls of this computer approach, where the lack of generality, selection or confounding biases, overfitting, or spurious correlations are among the commonplace flaws. Recommendations to reduce these risks include an examination of data through the lens of causality, the careful choice and description of statistical techniques, and an open research culture with transparency. Finally, the synergy between mechanistic and statistical models (ie, the digital twin) is discussed as a promising pathway toward precision cardiology that mimics the human experience. ■ *Heart Metab*. 2020;82:33-35

Keywords: artificial intelligence; big data; digital twin

"Big data" is the computer-enhanced version of inductive reasoning

There is an exciting future for medicine where decisions are informed by precise patient-specific data and risk models. Exploiting the "big data" in health care is one of the main engines working toward this future. We have learned from promising case studies, for example, the importance of access to the right source of evidence to select the right therapy for a pediatric lupus patient,[1] but also from epic failures, such as the unmet expectation to predict seasonal flu from Internet searches.[2]

It is useful to conceptualize "big data" as an evolution of traditional statistical methods, now able to harness the value of much larger and heterogeneous sources of information. As such, its ability to learn new biomarkers and predictors will always be subject to the same fundamental limitations of inductive reasoning: can we generalize the findings; are they really true? "Big data" is simply the computer-enhanced version of human inductive reasoning. The scientific method teaches us that the interplay between inductive and deductive reasoning is the way forward. We need to build a hypothesis from the observations and then advance to make predictions and to run more experiments to verify them.

In this context, this article reviews the risks and potential pitfalls of our computer-enhanced ability to reveal the hidden patterns in the data that predict cardiovascular outcomes. It then sets out some recommendations to minimize the risks of spurious patterns: here the main message is the need to examine the revelations through the lens of causality; do they make sense? We need to interpret the experimental findings and see if they fit our framework of a plausible mechanistic explanation.

Heart Metab. 2020;82:33-35

How to learn from data and how to fail in that endeavor

Our goal is to improve our future clinical decisions by learning from our past experiences, old-fashioned clinical observation. Our best tool to implement this collective knowledge is the use of clinical guidelines, the compendium of current best evidence mixed with opinion (class of evidence and level of recommendation). The future potential is to evolve and accelerate beyond this model by generating evidence using the "big data" that is becoming available. And this task is shared between humans and computers; there is a continuum between human and machine interactions to build predictive models.[3]

The main pitfall is to take the generality of our findings for granted and assume past experience predicts events in other cohorts and future patients. This can only be verified with external validation, a new cohort of patients, ideally from different clinical centers and/or geographic regions with a different mix of patients. Unfortunately, most studies will not include this critical step.[4,5]

Methodologically, the risk is that of "garbage in, garbage out" (GIGO): the validity of our findings strongly depends on the quality of the data learned from. Confounding biases will lead to surprising associations between health scores and risk factors that were actually driven by a lurking variable. Selection biases may lead to false conclusions and to ethical risks of models that create or exacerbate existing racial or societal biases in health care systems.[6]

Beyond these traditional statistical risks, the numerous comparisons and searches within the big data exacerbate other potential issues. We have many more chances of finding spurious correlations, such as the high-school basketball searches to predict seasonal flu burden.[2] And we suffer from the "dimensionality curse": the more variables you combine, the higher the chances of a spurious positive finding (eg, a false-positive of a dimension in which healthy and diseased subjects differentiate).

Big data is also characterized by three more features[7]: its heterogeneity (where inferring the mixture model would be a challenge), noise accumulation (so, selecting features would be better than trying to include all), and incidental endogeneity (that would make variable selection quite challenging). The reader is referred to previous works[8] for their detailed description.

Recommendations to avoid the pitfalls

The best attitude when reading "big data" studies is to be cautiously skeptical: a positive finding of the predictive value when working with thousands of variables is at best only a first step toward the right direction. The immediate next step is the preparation of the external validation tests.[4,5]

When conducting research in this area, the obvious recommendation is to apply adequate techniques. Selection bias can be overcome by weighting. The high-dimensionality curse and its dire consequences can be addressed by dimensionality-reduction techniques, where principal component analysis is the most common approach. And there are specific solutions for each of the challenges of big data: penalized quasi-likelihood, sparsest solution in high-confidence set, or independence screening among others (see ref 8 for further details).

The main challenge of induction is the generality of the findings, especially hard given the difficulty to have stable and comparable measurements across cases and time. The subtle differences in the appearance of an echocardiographic or magnetic resonance image across manufacturers is a well-known bottleneck in the imaging community. The homogenization of techniques and protocols is indeed one of the main strategies to alleviate this.[5]

As a community, the strongest recommendation in order to accelerate the generality of findings, and to eventually make an impact in patients, is the promotion of a culture of transparency[2] and open research. The effort to recruit and follow up a cohort of patients is huge, as it is the development of the information infrastructure to allow the access to the electronic health record of a large population. There are indeed ethical and societal barriers to release this data for research purposes, but we must learn to give the adequate value and credit to these contributions so that clinicians and researchers do not feel they are losing a competitive advantage.

The most difficult decision is when to include findings in clinical guidelines. The minimum is to have the positive result subjected to external validation. Even here one can critically challenge the generality of findings, and the recommendation is to take a practical sceptic approach: adopt while monitoring real-world results.

The challenge of generality will be only addressed by the formulation of the mechanistic hypothesis that

offers a plausible explanation of the findings, closing the induction step of the scientific method. In this context, computers can also be used to enhance our deductive reasoning skills: they can make predictions based on mechanistic simulations of our cardiovascular system.[9] The opportunity is thus to exploit the synergy between mechanistic and statistical computational models that is the core of the vision of the digital twin[10] and mimics the way clinicians have worked for millennia. ■

Disclosure/Acknowledgments: *Pablo Lamata was commissioned to write this article by, and has received honoraria from, Servier. Pablo Lamata sits on the advisory board of Ultromics Ltd and Cardesium Inc. Support from the Wellcome Trust Senior Research Fellowship (209450/Z/17/Z) is acknowledged. No conflict of interest is reported.*

REFERENCES

1. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. *N Engl J Med*. 2011;365(19):1758-1759. doi:10.1056/NEJMp1108726.
2. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science*. 2014;343(6176):1203-1205. doi:10.1126/science.1248506.
3. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317. doi:10.1001/jama.2017.18391.
4. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Heal*. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2.
5. Dey D, Slomka PJ, Leeson P, et al. Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *J Am Coll Cardiol*. 2019;73(11):1317-1335. doi:10.1016/J.JACC.2018.12.054.
6. Nordling L. A fairer way forward for AI in health care. *Nature*. 2019;573(7775):S103-S105. doi:10.1038/d41586-019-02872-2.
7. Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. *Int J Inf Manage*. 2015;35(2):137-144. doi:10.1016/J.IJINFOMGT.2014.10.007.
8. Fan J, Han F, Liu H. Challenges of big data analysis. *Natl Sci Rev*. 2014;1(2):293-314. doi:10.1093/nsr/nwt032.
9. Niederer SA, Lumens J, Trayanova NA. Computational models in cardiology. *Nat Rev Cardiol*. 2019;16(2):100-111. doi:10.1038/s41569-018-0104-y.
10. Corral-Acero J, Margara F, Marciniak M, et al. The "Digital Twin" to enable the vision of precision cardiology. *Eur Heart J*. March 4, 2020. Epub ahead of print. doi:10.1093/eurheartj/ehaa159.