Accurate data-driven prediction does not mean high reproducibility

Article *in* Nature Machine Intelligence · January 2020

DOI: 10.1038/s42256-019-0140-2

4 authors:

Jiuyong Li
University of South Australia
343 PUBLICATIONS **4,995** CITATIONS
SEE PROFILE

Lin Liu
University of South Australia
222 PUBLICATIONS **2,158** CITATIONS
SEE PROFILE

Thuc D le
University of South Australia
146 PUBLICATIONS **1,238** CITATIONS
SEE PROFILE

Jixue Liu
University of South Australia
162 PUBLICATIONS **1,528** CITATIONS
SEE PROFILE

Accurate data driven prediction does not mean high reproducibility[*]

A valid machine learning model is predictive, but a predictive model may not be valid. The gap between these two can be larger than many practitioners may expect.

Jiuyong Li, Lin Liu, Thuc Duy Le and Jixue Liu

University of South Australia, Australia

Email: firstname.lastname@unisa.edu.au

Low reproducibility is a major concern in data driven discovery with machine learning. Reproducibility requires that the relationships discovered from data are valid, i.e. they are causal and true in a real system. A model which can provide accurate prediction of an outcome does not mean that the predictors used by the model are likely causes of the outcome. It is generally understood that association does not necessarily indicate causation. However, since causes can be used to make quality predictions, many practitioners take prediction accuracy as an indicator of how likely a predictor is a cause of the outcome. In fact, prediction accuracy and causal validity are measures in two different worlds, and a wrong link between them is very harmful for data driven discovery. This Comment discusses the reason of low reproducibility of data driven discovery and provides practical guidelines for using general machine learning methods for data driven discovery.

Low reproducibility — a challenge for data driven discovery

The recent success of machine learning and AI has given a lot of hope for revolutionising scientific discovery with data driven approaches in many fundamental fields such as genomics, oncology and earth system sciences [1, 7].

A main task of scientific discovery is to identify causal relationships, i.e. whether and how the change of a variable (cause) alters the value of another variable (effect). In practice, a major challenge for data driven discovery with machine learning is the low reproducibility of the findings, which means that the relationships discovered in data can rarely be validated in a real world system. For example, a recent survey on early diagnosis and prognosis prediction of tongue squamous cell carcinoma examined 150 biomarkers discovered by various machine learning methods reported in 96 papers, only 10 were identified as promising candidates with a

[*] Citation: Jiuyong Li, Lin Liu, Thuc Duy Le and Jixue Liu. Accurate data-driven prediction does not mean high reproducibility. Nature Machine Intelligence 2 (1), 13-15, 2020.

clinical relevance [2]. Genome-wide association studies (GWAS) in the past decade have identified thousands of compelling associations between complex traits and diseases using various computational methods, but a major concern about the studies is that most genes signalled by the associations have no direct biological relevance to the diseases [8]. With online advertisement, it was found that the effect of display-advertising campaign on increasing keyword search for the displayed brand was greatly overestimated using machine learning methods, with the estimated increases ranging from 871% to 1198%, while the increase reported by a controlled experiment is only 5.4% [3].

In this Comment, we will explain the reason for the low reproducibility and provide some practical guidelines for using general machine learning methods for data driven discovery.

Reproducibility relies on validity of discovery

Reproducibility relies on the validity of the discovery by a machine learning model, i.e. the identified relationships between the predictors (features) and the outcome are causal and true in a real world system, as causal relationships imply the underlying mechanism of a system, which is supposed to keep unchanged in different studies. However, the majority of machine learning methods are association based. Associations indicate dependency between variables, e.g. two variables are linearly related in data. Associations are used for building a prediction model which maps feature values to class or outcome labels. The accuracy of a model indicates the level of consistency between the predicted class labels and known class labels in existing data.

Linking prediction accuracy and validity is misleading. It is well understood that association is not causation, but because causes can be used to make high quality predictions, many users take prediction accuracy as an indicator of how likely a predictor is a cause of the outcome. Moreover, the prediction accuracy of state-of-the-art machine learning models is so high, it is easily for people to believe that the models may code some causal relationships. We will discuss in the next section that there may not be a link between accuracy and validity.

Validity can be irrelevant to accuracy

To elaborate our points, we use the notation and concepts of the well-established graphical causal model, causal Bayesian networks [5]. For a set of variables in a domain, a causal Bayesian network comprises a causal structure represented by a causal directed acyclic graph (DAG) and the joint probability distribution of the set of variables. In a causal DAG, an edge $X \rightarrow Y$ between the two nodes (variables) X and Y indicates that X is a direct cause (parent) of Y, and Y is a direct effect (child) of X. In a causal DAG, a variable is independent of all of its non-descendants given the set of all its parents (known as the Markov condition).

Let us assume that a true causal mechanism governs the generation of data, as illustrated in Figure 1 (where the DAG representing the true causal mechanism is highlighted in red). From the data, we can obtain the joint probability distribution of the variables, which also indicates the associations and conditional independence between variables. A prediction model is built based on the associations learned from data. However, such a model does not guarantee the validity of the discovery no matter how accurate the predictions are, as the model may not imply the true causal mechanism).

Referring to Figure 1, from data or the joint probability distribution it is possible to infer the causal mechanism, with the assumptions of causal sufficiency (all common causes of two or more variables are included in the data) and faithfulness (every conditional independence encoded in the data is entailed by the Markov conditions in the causal DAG). However, the reality is that there can be multiple causal DAGs which represent the same joint probability distribution and entail the same Markov conditions (i.e. there is an equivalence class of causal DAGs which are Markov equivalent), but only one of the causal DAGs represents the true causal mechanism. This indicates the uncertainty in data driven discovery and that valid discoveries cannot be obtained from data alone. With hidden variables, the uncertainty is even higher. Although all the causal DAGs in the equivalence class are faithful to the probability distribution, and thus an accurate prediction model can be built based on any of them (or even based on a structure that is not in the equivalence class), accuracy does not indicate validity, i.e. whether the prediction model represents the true causal mechanism.

Prediction accuracy and causal validity are measures in two worlds. Prediction accuracy is measured in the world of observations or data. The consistency between observed and modelled joint probability distributions can be unbiasedly estimated when there is no noise in data. In other words, prediction accuracy can be high when a data set is adequately large and has no noise. Causal validity, on the other hand, is measured in the physical world with controlled experiments and it is impossible to quantify validity using data alone due to the uncertainty in data driven discovery.

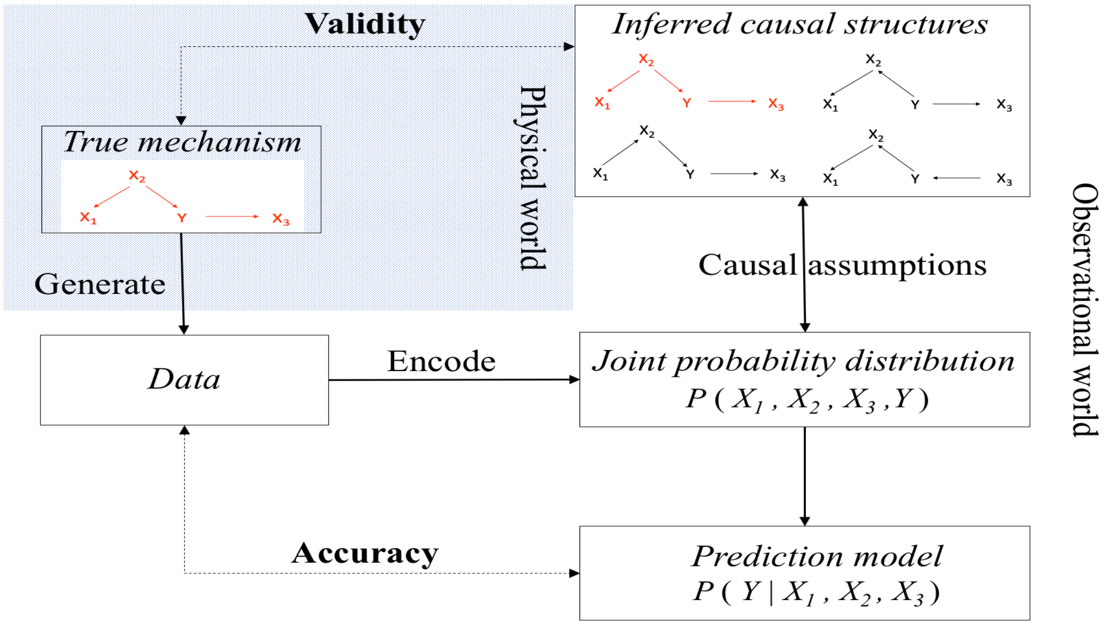

Figure 1: Accuracy versus validity. In the illustrative example, X_1, X_2 and X_3 are three observed variables, and *P()* denotes probability. Accuracy measures the consistency between a model and data, whereas validity indicates the link between an inferred causal structure and the real causal mechanism. The two measures are used in two different worlds, one observational and the other physical. Building a prediction model does not involve a causal structure but data driven discovery does. All the inferred causal structures shown are equivalent from the data viewpoint but they encode different causal relationships. Domain knowledge or known manipulations in an experiment helps identify the true causal structure.

Cross validation does not test reproducibility

Cross validation is frequently used in machine learning for measuring model accuracy. It splits a data set into two sub data sets: a training set and a held-out test data set, and the accuracy of the model built on the training data set is estimated on the test data set. To avoid the randomness of a split, the average accuracy over multiple rounds of splits is considered as the model accuracy. It is an effective means to prevent a model from overfitting the training data set. A model is said to overfit a data set if it provides highly accurate predictions on a training data set but does not perform well on a test data set.

Cross validation does not test reproducibility, although users might think so because cross validation accuracy is obtained from data unseen to the model. However, a held-out test data set in a cross validation is not an independent data set. It is from the same experiment (or observation) as the training data set. Therefore, a test on a held-out data set may reconfirm the same biases or spurious relationships and hence does not indicate validity. Essentially any evaluation methods using observational test data cannot test reproducibility since the test data, same as the training data, is incapable of telling the true causal mechanism from the other Markov equivalent causal structures, and thus cannot be used to demonstrate the validity of the model tested.

Some practical recommendations

In this Comment, we consider reproducibility as a property of valid discovery, and we have explained that accurate model may not code valid relationships. A practical question is then how to discover valid relationships from data, and the solutions belong to causal inference in data [4,7]. Methods in causal inference are based on strong assumptions, and therefore are not very practical yet, especially for large and high dimensional data sets. The majority of machine learning methods are for data driven prediction. The practice of using data driven prediction methods for data driven discovery will continue, but users should be aware of its limitations and do not overclaim what have been discovered. We have the following suggestions for using data driven prediction methods in a discovery process.

Domain knowledge is necessary for building a valid model and testing the validity of the model. For example, domain knowledge should be used for feature selection before model building, by including known and potential causes of the outcome and excluding effect variables of the outcome. Effect variables, if not excluded, may introduce spurious relationships. For example, in a model predicting Food Poisoning if the variable Fever (an effect of food poisoning) is selected as a predictor, irrelevant variables such as Influenza (another cause of Fever) could become associated with food poisoning because when given Fever is true, one cause of Fever could explain away the other cause. When strong predictive variables are identified with a prediction model, they should be checked against domain knowledge for validity. Data driven discovery needs a collaboration between domain experts and machine learning researchers/practitioners, which has been demonstrated in many real world cases, such as in the studies of complex climate and ocean systems [9].

Repeatedly discovered relationships in multiple independent data sets from different experiments/observations present evidence for their validity. However, users should be aware that current machine learning algorithms may not support such a test of reproducibility.

Accuracy (or a fitness measure) is often used to filter out insignificant results or suppress generating those results at all, and hence true causal signals may not be included in a prediction model. To identify or test valid relationships based on their reproducibility in different experiments or observations, it is better to use a machine learning method which finds and keeps the complete relationships instead of only those contributing to prediction accuracy. Such a method helps discover or test valid relationships from multiple data sets without revisiting the data sets.

A take-home message of this Comment is that machine learning researchers and practitioners should not focus only on the accuracy (or its variations) of a method when the goal is for data driven discovery since accuracy (even obtained by cross validation) does not indicate the validity of discovery. Practitioners should use domain knowledge to guide data driven discovery during feature selection and result validation, and use multiple independent data sets for discovery and validation. Fundamentally, association is not causation. Hill's criteria for causation [2] are useful guidelines for machine learning practitioners who try to interpret data driven discovery as causal. For machine learning researchers, focus should be switched from accuracy model building to robust model building and causal inference [7] for data driven discovery.

References

1. Editorial. Deep learning for genomics. *Nature Genetics*, 51(1):1–1, 2019.
2. A. B. Hill. The environment and disease: association or causation?. *Proceedings of the Royal Society of Medicine*. 58 (5): 295–300, 1965.
3. A. A. Hussein, T. Forouzanfar, E. Bloemena, J. de Visscher, R. H. Brakenhoff, C. R. Leemans, and M. N. Helder. A review of the most promising biomarkers for early diagnosis and prognosis prediction of tongue squamous cell carcinoma. *British Journal of Cancer*, 119:724-736, 2018.
4. G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, USA. 2015.
5. R. A. Lewis, J. M. Rao, and D. H. Reiley. Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th International Conference on World Wide Web*, pages 157–166, New York, NY, USA, 2011.
6. E. Ntzani and J. Ioannidis. Predictive ability of DNA microarrays for cancer outcomes and correlates: An empirical assessment. *Lancet,* 362:1439–1444, 2003.
7. J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.
8. J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Communication of the ACM*, 62(3):54–60, 2019.
9. D. F. Ransohoff. Rules of evidence for cancer molecular marker discovery and validation. *Nature Reviews Cancer*, 4:309–314,2004.
10. M. Reichstein, G. Camps-Valls, B. Stevens, et al. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204, 2019.

11. J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Gly- mour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, and et. al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

12. V. Tam, N. Patel, M. Turcotte, Y. Boss, G. Par, and D. Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467– 484, 2019.

13. A. Walther, E. Johnstone, C. Swanton, R. Midgley, I. Tomlinson, and D. Kerr. Genetic prognostic and predictive markers in colorectal cancer. *Nature Reviews Cancer*, 9(7):489–499, 2009.

There are no conflicts of interest.

Acknowledgements

This work has been supported by ARC Discovery Project grant DP170101306 and NHMRC Grant 1123042.