Accurate data-driven prediction does not mean high reproducibility

A valid machine 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

Low reproducibility is a major concern in data-driven discovery with machine learning. Reproducibility requires that the relationships discovered from data are valid — that is, they are causal and true in a real system. A model that can provide accurate prediction of an outcome does not mean that the predictors used by the model are likely to be 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 that 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.

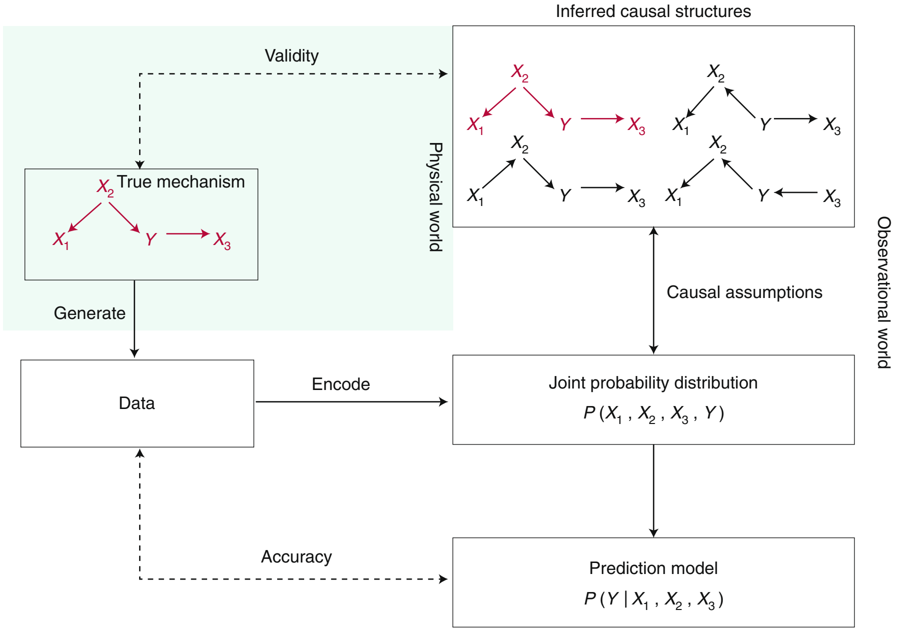

Fig. 1 | Accuracy versus validity. In the illustrative example, X_1, X_2 and X_3 are three observed variables, and $P()$ denotes probability. Accuracy measures the consistency between a model and data, whereas validity indicates the link between an inferred causal structure and the real causal mechanism. 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.

Low reproducibility in data-driven discovery

The recent success of machine learning and AI has given a lot of hope for revolutionizing scientific discovery with data-driven approaches in many fundamental fields such as genomics, oncology and earth system sciences[1,2].

A main task of scientific discovery is to identify causal relationships — that is, 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, in a recent survey on early diagnosis and prognosis prediction of tongue squamous cell carcinoma that examined 150 biomarkers discovered by various machine learning methods reported in 96 papers, only ten were identified as promising candidates with a clinical relevance[3]. 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[4]. With online advertisement, it was found that the effect of display-advertising campaigns on increasing keyword searches for the displayed brand was greatly overestimated using machine learning methods, with the estimated increases ranging from 871% to 1,198%, while the increase reported by a controlled experiment was only 5.4%[5].

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 — that is, 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 — for example, two variables are linearly related in data. Associations are used for building a prediction model that 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 easy 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[6]. 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 the causal DAG, an edge between two nodes (variables) — for example, $X \rightarrow Y$ indicates that X is a direct cause (parent) of Y and Y is a direct effect (child) of X — and 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 Fig. 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 Fig. 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 that represent the same joint probability distribution and entail the same Markov conditions — that is, there is an equivalence class of causal DAGs that 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 — that is, 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, and indicates the consistency between observed and modelled joint probability distributions. Prediction accuracy can be high when a dataset 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.

Cross-validation does not test reproducibility

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

Cross-validation does not test reproducibility, although users might think it does because cross-validation accuracy is obtained from data unseen to the model. However, a held-out test dataset in a cross-validation is not an independent dataset. It is from the same experiment (or observation) as the training dataset. Therefore, a test on a held-out dataset 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 models may not code valid relationships. A practical question then is how to discover valid relationships from data, and the solutions are being sought by researchers in the field of causal inference in data[6,7]. However, current causal inference methods are based on strong assumptions, and therefore are not very practical yet, especially for large and high-dimensional datasets. On the other hand, the majority of machine learning methods are for data-driven prediction and are effective. Hence 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 has been discovered based on accuracy. 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 that '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 the collaboration between domain experts and machine learning researchers and practitioners, which has been demonstrated in many real-world cases, such as in the studies of complex climate and ocean systems[8].

Repeatedly discovered relationships in multiple independent datasets 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 relationships that do not contribute significantly to predictions or suppress generating those relationships at all, and hence true causal relationships 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 that finds and keeps the complete set of relationships instead of only those contributing to prediction accuracy. Such a method helps discover valid relationships repeatedly appear in multiple datasets.

The 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 their goal is data-driven discovery, since accuracy (even obtained by cross-validation) does not indicate the validity of the discovery. Practitioners should use domain knowledge to guide data-driven discovery during feature selection and result validation, and use multiple independent datasets for discovery and validation. Fundamentally, association is not causation. Hill's criteria for causation[9] 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 accurate model building to robust model building and causal inference[10] for data-driven discovery. ❐

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

School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, South Australia, Australia.

**e-mail: jiuyong.li@unisa.edu.au*

Published online: 17 January 2020
https://doi.org/10.1038/s42256-019-0140-2

References

1. *Nat. Genet.* **51**, 1 (2019).
2. Runge, J. et al. *Nat. Commun.* **10**, 2553 (2019).
3. Hussein, A. A. et al. *Br. J. Cancer* **119**, 724–736 (2018).
4. Tam, V. et al. *Nat. Rev. Genet.* **20**, 467–484 (2019).
5. Lewis, R. A., Rao, J. M. & Reiley, D. H. in *Proc. 20th International Conference on World Wide Web* 157–166 (ACM, 2011).
6. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge Univ. Press, 2009).
7. Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge Univ. Press, 2015).
8. Reichstein, M. et al. *Nature* **566**, 195–204 (2019).
9. Hill, A. B. *Proc. R. Soc. Med.* **58**, 295–300 (1965).
10. Pearl, J. *Commun. ACM* **62**, 54–60 (2019).

Acknowledgements

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

Competing interests

The authors declare no competing interests.