Systematic human learning by literature and data mining for feature selection in machine learning

Herdiantri Sufriyana

(1) Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. (2) Department of Medical Physiology, Faculty of Medicine, Universitas Nahdlatul Ulama Surabaya, 57 Raya Jemursari Road, Surabaya 60237, Indonesia. https://orcid.org/0000-0001-9178-0222

Yu Wei Wu

(1) Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. (2) Clinical Big Data Research Center, Taipei Medical University Hospital, 250 Wu-Xing Street, Taipei 11031, Taiwan. https://orcid.org/0000-0002-5603-1194

Emily Chia-Yu Su (emilysu@tmu.edu.tw)

(1) Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. (2) Clinical Big Data Research Center, Taipei Medical University Hospital, 250 Wu-Xing Street, Taipei 11031, Taiwan. (3) Research Center for Artificial Intelligence in Medicine, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. https://orcid.org/0000-0003-4801-5159

Method Article

Keywords: causal diagram, feature selection, machine learning, medical history, electronic health record

Posted Date: October 13th, 2021

DOI: https://doi.org/10.21203/rs.3.pex-1634/v1

Abstract

We proposed a learning algorithm for human to conduct literature and data mining for causal factor discovery. The applicability is to select features for a machine learning prediction model, including but not limited to that using real-world, time-varying data from electronic health records. This protocol is relatively quick to find potentially actionable predictors for a clinical prediction while dealing with high dimensionality in big data. However, this protocol might not find a potentially novel cause, since this only exhaustively examines the existing evidences in a single study. The key stages consisted of systematic human learning, causal diagram construction, data preprocessing, causal inference modeling, and development and validation of a prediction model to describe the explainability.

Introduction

Insurance-based healthcare has been widely implemented worldwide, which urges more preventive intervention to improve patient outcome and reduce healthcare utility. Clinical prediction models are needed to achieve such purpose. Although several machine learning algorithms have been shown satisfying predictive performances for health outcomes,[1-3] these were exposed to optimistic bias due to no independent test set, no data partition, or high dimensionality relatively compared to sample size.[4] Clinical prediction also should be actionable. Identifying causal effects on a predicted health outcome enables targeted interventions to that condition. However, machine learning cannot infer causality yet.

The best practices established in medicine are developed using hypothetico-deductive reasoning. Any prior knowledge is collected by human learning through literature to generate a hypothesis. Subsequently, statistical methods are used to verify the assumption using available data. In traditional approach, randomized controlled trial (RCT) design have been used to collect data in a causal inference study because of its robustness to remove effects from common causes or confounding factors. However, this is limited due to some considerations (e.g. ethical issues). Even if RCT is possible, a preliminary study of a causal effect using observational data is still warranted to reduce resource waste and potential harm in human research. To conduct such study, a causal diagram is constructed based on prior knowledge as a central assumption.[5] Available data (e.g. electronic health records) may be used to verify this assumption. Solely using statistical analysis on available data without contextual knowledge can introduce data-driven bias.[6] Meanwhile, although human learning can use contextual knowledge to prevent such bias, we still need machines to deal with big data.

To solve data-driven bias using contextual knowledge, we proposed a learning algorithm for human to systematically construct a causal diagram by literature mining. Then, one of the generalized (G) methods, i.e., inverse probability weighting (IPW), is used to verify each causal factor using available data.[7,8] This method was designed for time-varying exposures which are typically available data from electronic health records.[9,10] Eventually, all causal factors are included in a prediction model to describe the explainability. The use of causal factors in a model is not necessarily be the highly predictive. Better performances of prediction model are normally achieved by exploiting confounding factors.[5] However, by

differentiating causal and non-causal predictors, this can warn a human user when conducting a critical appraisal a machine learning model. If a causal effect is large, this also provides a pathway for a preventive intervention. In addition, compared to systematic review and meta-analysis, our learning algorithm is relatively quick and low-intensive labor, which follows human intuition on learning through literatures.

We applied this protocol to several studies which were parts of a project that applied our human-machine learning algorithm to a variety of predicted outcomes. Human learning by this protocol was one of the comparators beside those applying standard machine learning prediction by PROBAST guidelines.[11] Ethical clearance was waived by the Taipei Medical University Joint Institutional Review Board (TMU-JIRB number: N202106025). This protocol aimed to propose a protocol for feature selection in statistical machine learning by an algorithm for human to systematically construct a causal diagram by literature mining, and to verify the causal assumptions from prior knowledge by statistical modeling, including but not limited to those using real-world, time-varying data from electronic health records.

Reagents

Equipment

We used R 4.0.2 programming language (R Foundation, Vienna, Austria) to conduct data analysis. The integrated development environment software was RStudio 1.3.959 (RStudio PBC, Boston, MA, USA). To ensure reproducibility, we used Bioconductor 3.11;[12] thus, versions of the included R packages were all in sync according to versions in this Bioconductor version. For statistical machine learning, we used an R package of caret 6.0.86 that wraps R packages for a modeling algorithm, which was glmnet 4.1. We created R packages for many steps in the data analysis, which are medhist 0.1.0 and gmethods 0.1.0. All of these packages are available for download from this repository https://github.com/herdiantrisufriyana. Details on other R package versions and all of the source codes (vignette) for the data analysis are available in https://github.com/herdiantrisufriyana/shl.

To reproduce our work, a set of hardware requirements may be needed. We used a single machine. It was equipped by 8 logical processors for the 3.40 GHz central processing unit (CPU) (Core(TM) i7-4770, Intel®, Santa Clara, CA, USA), and 16 GB RAM. But, one can use a machine with only 4 logical processors and 4 GB RAM, if the sample size is smaller than that of dataset we used in this protocol.

Procedure

1. Choose one or more literature databases

A systematic human learning was conducted by literature mining in a particular period. This drew on our assumption of causality. For simplicity and to avoid redundant records, we only used PubMed because it

is the most frequently updated (daily), has the longest period coverage (1950 to the present), and is a life science-focused literature database.[13] This database also allows use of specific terms in the Medical Subject Headings (MeSH) vocabulary thesaurus from the National Library of Medicine, National Institutes of Health (Bethesda, MD, USA).

2. Look for a document from an authoritative institution

We adopted snowball sampling method by starting with convenience sampling,[14,15] which was a document from an authoritative institution, to obtain a similar sense with human intuition when learning through the literature. We used the keywords '"Fetal Membranes, Premature Rupture"[Mesh]' to find the document for an outcome of prelabor rupture of membranes (PROM) in the literature database, as a convenience sampling step. This led to *Practice Bulletin No. 172* from the American College of Obstetrics and Gynecology (ACOG).[16] We only considered pregnant women as the population of which those studies investigated. The initial document was denoted d_0 (Algorithm 1).

Algorithm 1. Snowball sampling modified by starting from convenience sampling to obtain an initial document (d_0)

Require: d_0

01: $A = \emptyset$

02: $L = \emptyset$

03: $k_0 = \text{r}ead(d_0)$

04: **while** $ks \Delta A \neq \emptyset$ **then**

05: .....$a_s \leftarrow k_s$

06: .....$d_s = search(k_s)$

07: .....**if** $ds \neq \emptyset$ **then**

08: .....$k_{s+1} = read(d_s)$

09: .....**if** $causal(k_{s+1})$ **then**

10: .....**pass**

11: .....**else**

12: .....$l_s \leftarrow k_{s+1}$

13: .....**end while**

14: .....**else**

15: .....**end while**

3. Learn causal factors from the initial document

We denoted causal factors of PROM as *A*, while the confounders were denoted *L*. Confounders are causal factors of a causal factor of PROM. This means *L* represents the same factors that cause both *A* and PROM. Initially, there was no *A* or *L*. By reading an article/document d_0, we identified $a \in A$ to determine k_0 keywords that refer to *a* at the *s=0* stage. The next steps were iterative until no k_s keywords referred to any $a \in A$.

4. Search for causal factors for each causal factor of the outcome from either the initial document or the subsequent documents

We assigned k_s to a_s and searched for the document d_s using k_s for causal factors of a_s. If a document was found, then we continued; otherwise, the iteration ended. We continued by reading d_s to determine k_{s+1} keywords. This refers to a causal factor of $a \in A$ that is referred to by the previous k_s keywords.

5. Identify whether the causal factors from previous step are also causal factors of the outcome

Documents were searched and read to check if the k_{s+1} keyword also refers to causal factors of PROM. If yes, then the k_{s+1} keyword was passed to the *s+1* stage; otherwise, we assigned k_{s+1} to l_s and the iteration ended.

6. Construct a causal diagram for each proposed causal factor of the outcome

Factors of *A* and *L* are called first- or second-level factors of PROM, while only first-level ones are causal factors. This determined the position of factors within a circular network depicting a causal diagram which we used for causal inferences. Since first-level factors may come from second-level factors in the process, we could also find inter-causal factor relationships. We included these relationships as edges in the network, because these are needed to construct causal inference formulas. For each causal factor

with the common causes that have available data, a node and an edge to this node were drawn from the node of each variable consisting either a causal factor or the common causes. This node represented measured variables. Another node and an edge from this node were drawn to the represented node. This node represented unmeasured variables that can affect measurement error of the measured variables. Please kindly find out more explanation about constructing a causal diagram in this reference.[5] The source codes (vignette) for this step are available in https://github.com/herdiantrisufriyana/shl.

7. Split a dataset randomly for a discovery and validation set and define variables in the dataset that can represent each causal factor and the common causes

Only this set was used for causal inferences. Later, we also used it for training set of a prediction model. We represented demographics and medical histories as candidate causal factors if applicable. These were respectively binarized into 0 and 1 for negative and positive factors. Details of the ICD-10 codes and demographic variables we assigned to each causal factor are available in the source codes. We provided an R package medhist 0.1.0 consisting functions to extract, preprocess, and transform data into each causal factor from a nationwide health insurance claim data.

8. Define causal inference formula for each proposed causal factor of the outcome based on the causal diagram and available data

Only first-level factors were included in the formulas. For example, both asthma and influenza are first-level factors of PROM, while varicella is a second-level factor of PROM via asthma. To determine the formula for the causal inference of asthma, we included only asthma and influenza. We used only asthma's significance to determine if asthma was a causal factor of PROM. Only the causal factor of interest and confounding factors or common causes were included in the causal formula. We avoided including common effects to prevent collider-stratification bias, or unnecessary inclusion of second-level factors.[5]

9. Conduct causal inference modeling by a generalized (G) method

To verify our assumptions of PROM causality, we applied one of the generalized (G) methods, i.e., IPW, for each causal factor.[7,8] This method was designed for time-varying exposures.[9,10] However, we also conducted outcome regressions for causal inferences, since this is one of the more commonly methods although it does not work in general.[5,17] Another common method is propensity-score matching with various versions, but we did not apply this method for simplicity. While adjusting all confounding factors is difficult, if not impossible, we disclosed open backdoors (confounding factors that were not blocked) because of limitations of providing data for each causal factor. This will help in interpreting the results of

the study with caution.[18,19] An R package is provided for the causal inference modeling using G-methods, which is gmethods 0.1.0.

10. Develop and validate a prediction model to describe the explainability

After verifying causal factors, we only included those in a prediction model that applied a logistic regression with a shrinkage method, as recommended by PROBAST, instead of using a stepwise selection method.[11] We chose an RR, which applies L_2-norm or beta regularization, because this method retains all causal factors within the model after weights are updated by training.[20] The model was evaluated by the area under receiver operating characteristics to find the predictive performance using all confirmed causal factors.

Troubleshooting

A. Step 1

Problem

No exact MeSH term for the outcome

Possible reason

The outcome term may be one of entry terms.

Solution

Choose either the main or entry term. Browse all MeSH categories within the page of the most similar term to find an alternative term.

B. Step 2

Problem

No document from an authoritative institution

Possible reason

The outcome may be either a novel condition or a multidisciplinary problem.

Solution

Search for a narrative, scoping, or systematic review. If not available, search for an original article and explore the introduction and discussion section.

C. Step 3 to 5

Problem C.1

No causal factor of either each causal factor or outcome

Possible reason C.1

Either the causal factor or the outcome may be either a novel or an idiopathic condition.

Solution C.1

Loosen your criteria of causal factors of either each causal factor or outcome to any associated factor.

Problem C.2

Causal factor as a child term of another factor

Possible reason C.2

A causal factor has several subtypes.

Solution C.2

Assign a child term as a mediator of the parent term on another factor or an outcome without removing the direct relationship.

Problem C.3

Several distinguished terms of outcome

Possible reason C.3

An outcome has several subtypes.

Solution C.3

Assign a subtype as a mediator of the causal factor on an outcome without removing direct relationship between the causal factor on the unspecified outcome.

Problem C.4

Unclear causal or non-causal association

Possible reason C.4

Statistical analysis is insufficiently described.

Solution C.4

There is no need to confirm the causal effect. These steps aim to identify existing assumptions from previous researchers about potential causes of an outcome.

D. Step 6 to 9

Problem D.1

No available data for any causal factor

Possible reason D.1.1

A causal factor have no specific diagnosis or procedure code.

Solution D.1.1

Consider other codes that may represent a causal factor.

Possible reason D.1.2

There is no relevant variable in a dataset.

Solution D.1.2

Consider to get another secondary dataset or collect primary data.

E. Step 10

Problem E.1

No confirmed causal factors for a prediction model

Possible reason E.1.1

Diagnosis or procedure codes may not represent a causal factor.

Solution E.1.1

Consider other codes that may represent a causal factor.

Possible reason E.1.2

There is no relevant variable in a dataset.

Solution E.1.2

Consider to get another secondary dataset or collect primary data.

Time Taken

All steps

Approximate time: 4 to 11 days

Step 1 to 5

Approximate time: 6 to 8 hours a day for 3 to 10 days

Step 6 to 8

Approximate time: 20 minutes to 2 hours per causal factor

Step 9

Approximate time: 1 to 20 minutes per causal factor

Step 10

Approximate time: 1 to 20 minutes per causal factor

Anticipated Results

Several causal diagrams are expected to be constructed for an outcome, beginning from outcome-related guidelines by an authoritative institution. A study may indicate >1 relationships between causal factors and either another causal factor or an outcome. Some covariates in a causal diagram may not have available data, but, the diagram is shown to disclose potential backdoors or confounding factors that are not controlled. The outcome regression may show larger effects compared to those by IPW. Since the latter method, which is a G-method, is possible to estimate a causal effect although the causal model is mistakenly specified,[17] the confounding effects are well-removed thus demonstrating smaller effect of a causal factor.

References

1. Fleuren, L.M., et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 46, 383-400 (2020).

2. Lee, Y., et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. J Affect Disord 241, 519-532 (2018).

3. Gonem, S., Janssens, W., Das, N. & Topalovic, M. Applications of artificial intelligence and machine learning in respiratory medicine. Thorax 75, 695-701 (2020).

4. Sufriyana, H., et al. Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis. JMIR Med Inform 8, e16503 (2020).

5. Hernán, M.A. & Robins, J.M. Causal Inference: What If. , (Chapman & Hall/CRC, Boca Raton, 2020).

6. Wilkinson, J., et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health 2, e677-e680 (2020).

7. Hernán, M.A. How to estimate the effect of treatment duration on survival outcomes using observational data. Bmj 360, k182 (2018).

8. Chatton, A., et al. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. Sci Rep 10, 9219 (2020).

9. Naimi, A.I., Cole, S.R. & Kennedy, E.H. An introduction to g methods. Int J Epidemiol 46, 756-762 (2017).

10. Doosti-Irani, A., Mansournia, M.A. & Collins, G. Use of G-methods for handling time-varying confounding in observational research. Lancet Glob Health 7, e35 (2019).

11. Moons, K.G.M., et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med 170, W1-w33 (2019).

12. Huber, W., et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12, 115-121 (2015).

13. Falagas, M.E., Pitsouni, E.I., Malietzis, G.A. & Pappas, G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. Faseb j 22, 338-342 (2008).

14. Goodman, L.A. Snowball Sampling. The Annals of Mathematical Statistics 32, 148-170, 123 (1961).

15. Lee, J. & Spratling, R. Recruiting Mothers of Children With Developmental Disabilities: Adaptations of the Snowball Sampling Technique Using Social Media. J Pediatr Health Care 33, 107-110 (2019).

16. ACOG. Practice Bulletin No. 172: Premature Rupture of Membranes. Obstet Gynecol 128, e165-177 (2016).

17. Dukes, O. & Vansteelandt, S. A Note on G-Estimation of Causal Risk Ratios. Am J Epidemiol 187, 1079-1084 (2018).

18. Hernán, M.A. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. Am J Public Health 108, 616-619 (2018).

19. Hernán, M. The C-Word: The More We Discuss It, the Less Dirty It Sounds. Am J Public Health 108, 625-626 (2018).

20. Van Calster, B., van Smeden, M., De Cock, B. & Steyerberg, E.W. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Stat Methods Med Res 29, 3166-3178 (2020).

Acknowledgements

The social security administrator for health or *badan penyelenggara jaminan sosial (BPJS) kesehatan* in Indonesia gave permission to access the sample dataset in this protocol (dataset request approval number: 5064/I.2/0421). This protocol was funded by the Ministry of Science and Technology (MOST) in Taiwan (grant number MOST109-2221-E-038-018 and MOST110-2628-E-038-001) and the Higher Education Sprout Project from the Ministry of Education (MOE) in Taiwan (grant number DP2-110-21121-01-A-13) to Emily Chia-Yu Su.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppprotocolshl.pdf