Applying PC Algorithm and GES to Three Clinical Data Sets: Heart Disease, Diabetes, and Hepatitis

Nurdi Afrianto, Yopi Azzani, Yuan Sa’adati, Nurhaeka Tou, Putri Mentari Endraswari, Yohani Setiya Rafika Nur, Nur Annisa, Rifai Nur Widyanara, Ridho Rahmadi

Master Program in Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia

E-mail: [nurdi.afrianto, yopi.azzani, yuan.saadati, nurhaeka.tou, putri.endraswari, yohani.nur, nur.annisa, rifai.widyanara]@students.uii.ac.id
E-mail: ridho.rahmadi@uii.ac.id

Abstract. The goal of many sciences, including those related to the clinical domain, is to discover the generative model, that is, to understand how variables in the data take on their values. This goal cannot be addressed directly using approaches such as machine learning and deep learning, as such methods focus more on the association between input and output variables. In this paper, we aim to show to the readers an alternative approach, which can be a more appropriate method to target such aforesaid research goal. This approach is called causal modeling. We will first begin with some application examples of machine learning and deep learning on clinical data, and then show our applications of causal modeling to three clinical real-world data sets. This paper is projected to be a concise guideline for researchers to causal modeling, as well as to choose suitable approaches for problems of interest.

1. Introduction

Machine learning (ML) and deep learning (DL) are advances in Artificial Intelligence, and have long been applied in the clinical domain [1][2][3]. Those two concepts aim to develop (mostly automatic) systems which can learn models from data. The typical learning methods are generally of two paradigms: supervised learning and unsupervised learning. The supervised learning focuses on tasks, e.g., prediction and classification. Examples of this paradigm are Naïve Bayes classifier, support vector machine (SVM), and convolution neural network. The unsupervised learning, on the other hand, focuses on different tasks, e.g., clustering and components analysis. Examples of this paradigm are K-Means, independent component analysis, and Autoencoders.

Both ML and DL, however, mainly target the associations between input variables $X$ and output variables $Y$, that is, $P(y|x)$. Here both $X$ and $Y$ are observations on domains of interest. The learning process here is designed to project output $Y$ based on the given input $X$. A practical example is to learn whether or not an individual has diabetes (output), based on his/her recent clinical examination (input) gathered in a hospital. Such a task is called classification. Despite the accuracy of such classification models, the conducted learning process results in trained model parameters which are numerous and difficult to explain [4]. In some DL models, for
example, the number of parameters reaches thousands or even millions \[5\][6][7]. The lack of explainability of ML and DL models has been discussed in [4].

The goal of many sciences, including those related to the clinical domain, is to discover generative models. Such models explain how variables in the data take on their values [8]. This goal, however, cannot be addressed directly using ML and DL models, which mostly focus on the association between input and output variables \(P(Y|X)\). As an alternative, a method called causal modeling or causal discovery can be used to address the aforesaid generative models. Basically, a causal model learns how variable \(Y\) (effect) changes its value solely based on another variable \(X\) (cause). Often \(X\) is obtained from an intervention, e.g., a medical therapy; here we denote \(do(x)\) to distinguish a result of an intervention from a merely observation \(x\). Unlike \(P(y|x)\) in ML and DL problems, here we are interested in to see the mechanisms entailed by \(P(y|do(x), z)\), that is, “what will be \(y\), if I do \(X = x\), and observe some \(z\)”. Understanding such mechanisms helps us to estimate the model which generates the data we have. In clinical practice, obtaining a causal model can lead us to better develop medication or therapy, because we could focus more on the variables which have more impact on the disease or the disorder. Technically, the causal methods can be divided into two approaches, namely the constraint-based approach and the score-based approach [9]. The constraint based approach uses conditional independence tests and orientation rules to obtain a causal model. The score based approach uses scores, often for model refinements, which are the building blocks of the final causal model.

In this study, therefore, we aim to show causal modeling applications in estimating the generative models of some clinical data, as an alternative method to those of ML and DL, which focus more on the association between input and output variables. In particular, we will apply causal modeling on three clinical data sets, namely, hepatitis [10], diabetes mellitus [11], and heart attack/disease [12]. These three diseases are of the five main causes of death worldwide [13], and thus understanding the causal mechanisms underlying each of the diseases is very important. Especially in recent years, we are faced with the challenges of developing more efficient and effective medications and treatments. We hope that the outcome of this study could become both scientific and practical references in the related fields. Technically, we will employ a score-based method called GES (Greedy Equivalent Search)[14] and a constraint-based method called PC algorithm [15]. We are aware that these two causal methods are not the current state-of-the-art. These methods, however, generally reflect the foundation and simplicity of constraint-based procedure (PC algorithm) and score-based procedure. The recent causal methods, on the other hand, are based on these methods with additional steps and assumptions, or even with more complicated algorithms. As the main purpose of the present study is to introduce causal modeling as an alternative approach, we select simpler methods such as PC algorithm and GES.

This paper is organized as follows. Section 1 describes the background and motivation of the study. Section 2 describes the previous studies which are relevant. Section 3 explains the data and detailed methods. Section 4 discusses the results, and Section 5 concludes the study.

2. Previous Studies
There have been numerous clinical studies on heart disease, which applying machine learning or deep learning. For example, to predict the outcome of patients given some clinical measurements. For example, a study by Rufai and Umar [16] attempted to predict heart disease status from 303 patients using a neural network model. Rajesh et al. [17] reviewed the performance of classification methods such as Naive Bayes classifier and decision trees for predicting heart disease from 300 patients. A research conducted by Sisodia and Sisodia [18] attempted to predict diabetes using decision tree, Naive Bayes classifier, and SVM. The study was conducted on the Pima Indians Diabetes Database. Kumar and Vignaswari [19] administered a comparison study of machine learning applications (logistic regression, random forest, C4.5, and multilayer
perceptron) to hepatitis prediction, based on 155 patients. There have been some causal studies on clinical or psychological cases, e.g., Alzheimer’s disease, chronic fatigue syndrome, and chronic kidney disease [20][21][22]. In this present study we aim to apply causal modeling to clinical data similar to the studies described in the beginning, i.e., heart disease, hepatitis, and diabetes.

3. Data and methods

We use R [23] to compute the analyses. In particular we use package pcalg [24] to apply the PC algorithm and GES to the following data sets. For each data set, we deliberately select some variables which are continuous and of our interest. As our main purpose of the present study is to introduce causal modeling as an alternative method, we will conduct simple applications rather than comparisons. Here, we will apply PC algorithms to the Heart disease and the Diabetes data sets, and GES to the Hepatitis data set.

3.0.1. Heart disease data
The data set is obtained from kaggle.com, containing observations on 303 patients who have heart attack possibility. Each patient is characterized by 14 attributes related to their health conditions. In this study, we select in particular variables of interest, that is, resting blood pressure (trestbps), serum cholesterol (chol), maximum heart rate achieved (thalach), and ST depression induced by exercise relative to rest looks at stress of heart during exercise unhealthy heart will stress more (oldpeak).

3.0.2. Diabetes data
The Pima Indians Diabetes is originally from the National Institute of Diabetes and Digestive and Kidney Diseases, and obtained from kaggle.com. The data set comprises observations on 768 females. Originally this data set contains nine continuous attributes and one outcome. In this study particularly we select variables of interest, namely Glucose, Blood Pressure, Insulin, and BMI.

3.0.3. Hepatitis
The data set is obtained from kaggle.com, containing observations on 142 patients. Originally the data comprises 20 variables. In this study we select some variables of interest, that is, Age, Albumin, Alkaline Phospate, and Protime.

3.1. Methods

3.1.1. PC-Algorithm
PC algorithm [15] seeks causal relations by testing conditional independence among variables. It has two steps. In the first step, PC algorithm starts with a complete graph and repeatedly eliminates edges connecting two variables if those variables are independent based on conditional independence test, resulting in a skeleton. In the second step, the skeleton is directed, resulting in a causal model represented by a partially completed directed acyclic graph (CPDAG). For more detail of the first step, see the Algorithm1. Technically, this step tests for conditional independence among pairs of variable across conditioning set that increases in size \((d + 1)\). When \(X\) and \(Y\) are independent conditional on \(Z\), that is, \(I(X, Y|Z) = \text{TRUE}\), we eliminate the edge between \(X\) and \(Y\) and save \(Z\) in the corresponding separating set. Once the skeleton is estimated, it is oriented with the following rules. For each triple of variables \((X, Z, Y)\) such that \(X\) and \(Z\) are adjacent, \(Z\) and \(Y\) are adjacent, and \(X\) and \(Y\) are not adjacent, orient the edges \(X \leftarrow Z \rightarrow Y\) as \(X \rightarrow Z \leftarrow Y\), if \(Z\) was not in the set conditioning on which \(X\) and \(Y\) became independent and the edge between them was accordingly eliminated. We call such a triple of variables a \(v\)-structure. For each triple of variables such that \(X \rightarrow Z \rightarrow Y\), and \(X\) and \(Y\) are not
adjacent, orient the edge $Z - Y$ as $Z \rightarrow Y$. The pseudocode of PC algorithm is adopted from [25].

**Algorithm 1** PC Algorithm

Input: Dataset $D$ with a set of variables $V$ and significant level $\alpha$
Output: The undirected graph $G$ with a set of edges $E$

Assume all nodes are connected initially
Let depth $d = 0$

1: repeat
2: for each ordered pair of adjacent vertices $X$ and $Y$ in $G$ do
3: if $|\text{adj}(X, G) \setminus \{Y\}| \geq d$ then
4: for each subset $Z \subset \text{adj}(X, G) \setminus \{Y\}$ and $|Z| = d$ do
5: Test $I(X, Y | Z)$
6: if $I(X, Y | Z)$ then
7: Remove edge between $X$ and $Y$
8: Save $Z$ as the separating set of $(X, Y)$
9: Update $G$ and $E$
10: break
11: end if
12: end for
13: end if
14: end for
15: Let $d = d + 1$
16: until $|\text{adj}(X, G) \setminus \{Y\}| < \text{for every pair of adjacent vertices in } G$

3.1.2. GES

The Greedy Equivalence Search [14] (GES; see Algorithm 2) is a score-based causal method, consisting of forward equivalence search (FES; see Algorithm 3) and backward equivalence search (BES; see algorithm 4). GES repeatedly adds causal relations via FES until no more improvement on the score (typically Bayesian Information Criterion) and repeatedly deletes causal relations via BES until no more improvement on the score. The pseudocode of GES is adopted from [26].

**Algorithm 2** Add Operation (Forward phase)

Input: A dataset $D$ and a CPDAG $G$
Output: A CPDAG $H$

Let $M$ be a set of CPDAGs obtainable from $G$ by adding one edge
Let $K$ be a CPDAG from $M$ with the highest score $S(D, K)$

1: if $S(D, G) < S(D, K)$ then
2: \hspace{1cm} $H \leftarrow K$
3: else
4: \hspace{1cm} $H \leftarrow G$
5: end if
6: return $H$
Algorithm 3 Delete Operation (Backward phase)

**Input** A dataset $D$ and a CPDAG $G$

**Output** A CPDAG $H$

Let $M$ be a set of CPDAGs obtainable from $G$ by deleting one edge

Let $K$ be a CPDAG from $M$ with the highest score $S(D, K)$

1. if $S(D, G) < S(D, K)$ then
2. $H \leftarrow K$
3. else
4. $H \leftarrow G$
5. end if
6. return $H$

Algorithm 4 GES

**Input** A dataset $D$

**Output** A CPDAG $G$

1. while $S(D, \text{Add}(D, G)) > S(D, G)$ do
2. $G \leftarrow \text{Add}(D, G)$
3. end while
4. while $S(D, \text{Del}(D, G)) > S(D, G)$ do
5. $G \leftarrow \text{Del}(D, G)$
6. end while
7. return $G$

3.2. Source code

We provide the source code used in all computation in https://github.com/nurdi1995. As an example, the following source code is used for computing heart disease data set using PC algorithm.

```r
1 continuous_data <- data.frame(data$trestbps, data$chol,
2 data$thalach, data$oldpeak)
3 suffStat <- list(C=cor(y), n = nrow(continuous_data))
4 pc.fit <- pc(suffStat, indepTest=gaussCItest,
5 p = ncol(continuous_data), alpha=0.05, verbose = TRUE)
```

4. Results and Discussion

4.1. Heart disease

We applied PC algorithm to the heart disease data set. We assume the variables are normally distributed and use the Gaussian conditional independence test with $\alpha = 0.05\%$.

Figure 1 shows the CPDAG estimated by PC algorithm. The causal relations between chol and trestbps indicates cholesterol factor will have an effect on blood pressure where the body is at rest. The research that has been done by [27] mentioned that increased serum cholesterol levels are able to influence blood pressure, at least during sympathetic stimulation. We also find that oldpeak, which indicated myocardial ischemia (condition where blood flow is obstructed by plaques), are influenced by thalach and trestbps. In studies [28] [29], it is explained that a consistent ST depression elevation hypertension affects left ventricular hypertrophy on the ECG is still within normal limits and an increase in the maximum heart rate will affect ST depression during the treadmill process. In general, the causal relations estimated here are corroborated by previous studies.
4.2. Diabetes

We used PC algorithms to estimate causal relations from the Diabetes data. Figure 2 shows the result. We found several causal relations here: from insulin and blood pressure to BMI, from insulin to BMI, and from insulin to glucose. Studies by Dewi, et al. [30] found that a positive and significant relationship between BMI and glucose. Moreover, the studies in [31] [32] have found an association between insulin and BMI. Overall, mean systolic and diastolic blood pressures increased significantly and linearly across BMI levels. We found gradient increases in blood pressure with higher BMI levels. The fact that this gradient exists in a fully adjusted analysis suggests that BMI can cause an immediate effect on blood pressure, independent of other clinical risk factors. We can see that our findings are in general, accordance with those of some previous studies [33].

4.3. Hepatitis

Figure 3 shows the result of applying GES to the Hepatitis data. We found that age and alkaline phosphatase variables causes the albumin, and albumin causes prothrombin time. Several previous studies are in favor of these findings. The studies in [34] found that a decrease in albumin levels in the blood is significantly associated with aging, i.e., the older a person is, the lower his/her level of albumin in the blood. Moreover, the studies in [35] shows that the Prothrombin time and albumin are positively related to the incidence of hepatitis. Again, in general our causal findings in the Hepatitis data conform those of the previous studies.

5. Conclusion & Future works

ML and DL typically resolves problems by modeling the association between input and output variables. In some domains, however, the main goal is to understand the model generating data. Such case cannot be addressed directly by ML and DL. The aim of the present study is to
show causal modeling as an approach to such problem. Causal modeling attempts to identify causal relations among variables, which constitute the generative model. Understanding such a model is important, e.g., when developing a new clinical treatment for a disease, we need to understand which symptoms are the main causes of the disease, in order to design the best intervention. We showed that our results of applying causal methods on three clinical data sets. The results are, in general, in accordance with those of relevant previous studies. Our present study, however, does not cover practical problems such as missing values or non-normal data distribution. Future work should attempt to address those problems.

6. References

[1] Kumar P S and Pranavi S 2017 Performance analysis of machine learning algorithms on diabetes dataset using big data analytics Int. Conf. on Infocom Technologies and Unmanned Systems pp 509–13
[2] Fatima M and Pasha M 2017 Survey of machine learning algorithms for disease diagnostic J. of Intelligent Learning Systems and Applications vol 09 no 01 pp 1–16
[3] Bakator M and Radosav D 2018 Deep learning and medical diagnosis: a review of literature Multimodal Technologies and Interaction vol 2 no 3 pp 1–12
[4] Pearl J 2019 The seven tools of causal inference, with reflections on machine learning Commun. ACM vol 62 no 3 pp 54–60
[5] Bzdok D, Nichols T E and Smith S M 2019 Towards algorithmic analytics for large-scale datasets Nat Mach Intell vol 1 no 7 pp 296–306
[6] Gupta S, Zhang W and Wang F 2016 Model accuracy and runtime tradeoff in distributed deep learning: a systematic study IEEE 16th Int. Conf. on Data Mining (ICDM) pp 171–80
[7] Sejnowski T J 2020 The unreasonable effectiveness of deep learning in artificial intelligence preprint 2002.04806 [q-bio.NC]
[8] Spirtes P 2010 Introduction to causal inference J. Mach. Learn. Res vol 11 no 5 pp 1644–62
[9] Rahmadi R, Groot P, Heins M, Knoop H and Heskes T 2017 Causality on cross-sectional data: stable specification search in constrained structural equation modeling Appl. Soft Comput. vol 52 no 2 pp 687–98
[10] HariniR 2020 Hepatitis Retrieved on 13 Sep. 2020 from https://www.kaggle.com/harinir/hepatitis
[11] Devisangeetha 2017 Which Factor Causes Diabetes Retrieved on 13 Sep. 2020 from https://www.kaggle.com/devisangeetha/which-factor-causes-diabetes
[12] Bhat N 2020 Health Care: Data Set on Heart Attack Possibility Retrieved on 13 Sep. 2020 from https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility.
[13] Onggo I T and Wijayanti D 2011 5 Pengikut Utama Pencahayaan Nyawa: Jantung - Stroke - Diabetes Mellitus - Kanker - Hepatitis B and C (Yogyakarta: Mitra Buku)
[14] Chickering D M and Meek C 2015 Selective Greedy Equivalence Search: Finding Optimal Bayesian Networks Using a Polynomial Number of Score Evaluations preprint 1506.02113 [cs.LG]
[15] Spirtes P, Glymour C and Scheines R 2000 Causation, Prediction, and Search, Second Edition (Cambridge: MIT Press)
[16] Rufai A U S and Umar M 2018 Using artificial neural networks to diagnose heart disease IJCAI vol 182 no 19 pp 1–6
[17] Rajesh N, Maneesha T, Hafeez S and Krishna H 2018 Prediction of heart disease using machine learning algorithms Int. J. of Eng. and Technol. vol 7 no 2.32 pp 363–6
[18] Sisodia D and Sisodia D S 2018 Prediction of diabetes using classification algorithms *Procedia Comput. Sci.* vol 132 pp 1578–85

[19] Kumar N K and Vigneswari D 2019 Hepatitis-infectious disease prediction using classification algorithms *J. Adv. Pharm. Technol. Res.* vol 12 no 8 pp 3720–5

[20] Rahmadi R, Groot P and Heskes T 2018 Stable specification search in structural equation model with latent variables *preprint* 1805.09527 [stat.ML]

[21] Rahmadi R, Groot P and Heskes T 2017 Causality on cross-sectional data: stable specification search in constrained structural equation modeling *Applied Soft Computing* vol 52 pp 687–98

[22] Rahmadi R et al 2018 Causality on longitudinal data: stable specification search in constrained structural equation modeling *Stat. Methods Med. Res.* vol 27 no 12 pp 3814–34

[23] Douglas A C & D L A, Roos D and Mancini F 2018 *Quantitative Geography: The Basics* (California: Sage Publication)

[24] R Development Core Team 2020 *The Comprehensive R Archive Network* Retrieved from https://cran.r-project.org

[25] Le T D, Hoang T, Li J, Liu L, Liu H and Hu S 2019 A Fast PC Algorithm for High Dimensional Causal Discovery with Multi-Core PCs *IEEE/ACM Trans. Comput. Biol. Bioinform.* vol 16 no 5 pp 1483–95

[26] Coumans V 2017 *Causal Discovery Algorithms and Real World Systems* (Nijmegen: Radboud University)

[27] Alonso-Barba J I, delaOssa L, Gámez J A and Puerta J M 2013 Scaling up the greedy equivalence search algorithm by constraining the search space of equivalence classes *Int. J. Approx. Reason.* vol 54 no 4 pp 429–51

[28] Rampengan S H 2015 *Kegawatdarutan Jantung* (Jakarta: Fakultas Kedokteran Universitas Indonesia)

[29] Radi B, Arson I, Sarvasti D, Tadjodin Y and Tjahjono C 2016 *Pedoman Uji Latih Jantung: Prosedur dan Interprestasi* (Jakarta: PERKI)

[30] Dewi D Y, Subawa A A N and Mahartini N N 2019 Hubungan indeks massa tubuh terhadap kadar gula darah puasa pengunjung lapangan Niti Mandala Renon bulan Juli tahun 2018 *Intisari Sains Medis.* vol 10 no 3 pp 711–4

[31] Jawad S M A, Al-Zubaide B A and Al-Shalan N W K 2019 The relationship between insulin hormone and obesity in type 2 diabetic patients *Drug Invention Today.* vol 11 pp 2589–92

[32] Al-Goblan A S, Al-Alfi M A and Khan M Z 2014 Mechanism linking diabetes mellitus and obesity *Diabetes Metab. Syndr. Obes.* vol 7 pp 587–91

[33] Landi F et al 2018 Body mass index is strongly associated with hypertension: Results from the longevity check-up 7+ study *Nutrients* vol 10 no 12 pp 1–12

[34] Gom I et al 2007 Relationship between serum albumin level and aging in community-dwelling self-supported elderly population *J. Nutr. Sci. Vitaminol.* vol 53 no 1 pp 37–42

[35] Tungadi N 2017 *Hubungan Nilai Prothrombin Time dan Albumin dengan Staging Pasien Sirosis Hepatitis Di RSU DR. Wahidin Sudirohusodo* (Makasar: Universitas Hasanudin)