Causality on Longitudinal Data: Stable Specification Search in Constrained Structural Equation Modeling

Ridho Rahmadi, Perry Groot, Marieke MHC van Rijn, Jan AJG van den Brand, Marianne Heins, Hans Knoop, Tom Heskes, the Alzheimer’s Disease Neuroimaging Initiative*, the MASTERPLAN Study Group, the OPTIMISTIC consortium**

Abstract—Developing causal models from observational longitudinal studies is an important, ubiquitous problem in many disciplines. In the medical domain, especially in the case of rare diseases, revealing causal relationships from a given data set may lead to improvement of clinical practice, e.g., development of treatment and medication. Many causal discovery methods have been introduced in the past decades. A disadvantage of these causal discovery algorithms, however, is the inherent instability in structure estimation. With finite data samples small changes in the data can lead to completely different optimal structures. The present work presents a new causal discovery algorithm for longitudinal data that is robust for finite data samples. The method works as follows. We model causal models using structural equation models. Models are scored along two objectives: the model fit and the model complexity. Since both objectives are often conflicting we use a multi-objective evolutionary algorithm to search for Pareto optimal models. To handle the instability of small finite data samples, we repeatedly subsample the data and select those substructures (from optimal models) that are both stable and parsimonious which are then used to infer a causal model. In order to validate, we compare our method with the state-of-the-art PC algorithm on a simulated data set with the known ground truth model. Furthermore, we present the results of our discovery algorithm on three real-world longitudinal data sets about chronic fatigue syndrome, Alzheimer disease and chronic kidney disease that have been corroborated by medical experts and literature.

Index Terms—Longitudinal data, Causal modeling, Structural equation model, Stability selection, Multi-objective evolutionary algorithm.

1 INTRODUCTION

Extracting causal models from observational longitudinal studies often turns out to be an essential, important, and ubiquitous problem in many disciplines [2], [3], [4]. In the medical domain, especially in case of rare diseases, revealing causal relationships from longitudinal observational data set may lead to improvement of clinical practice, for example, the development of treatment and medication.

Many causal discovery algorithm have been introduced in the past decades, for example, as implemented in [5], [6], [7], [8], [9]. Of aforementioned methods, we can distinguish two approaches: constraint-based and score-based.

R. Rahmadi is with the Department of Informatics, Universitas Islam Indonesia and Institute for Computing and Information Sciences, Radboud University Nijmegen, the Netherlands. E-mail: rrahmadi@cs.ru.nl

P. Groot and T. Heskes are with Institute for Computing and Information Sciences, Radboud University Nijmegen, the Netherlands.

M. van Rijn and J. van den Brand are with Department of Nephrology, Radboud University Medical Center, Nijmegen, The Netherlands.

* Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

M. Heins and H. Knoop are with Expert Centre for Chronic Fatigue, Radboud University Medical Centre, the Netherlands.

**The members of OPTIMISTIC consortium are described in [1].
The former approach starts with a complete graph and constructs a skeleton graph by excluding edges between variables that are conditionally independent. Edges are then oriented to form a causal graph. As this approach does not rely on the causal sufficiency assumption, it is able to detect a common cause of the observed variables [10]. However, the use of independence tests on a large number of conditioning variables makes it less reliable [11]. The latter approach scores graph structures based on their fitness (how well a model fits the data) and the complexity of the graph. The main goal of the score-based approach is to obtain the graph structure with the best score. As this approach measures the reliability of the inferred causal relationships, the result is easier to interpret [12]. However, a score-based method usually relies on the causal sufficiency assumption, making it unable to detect common confounders of the observed variables. And since such an optimization problem is usually NP-hard, different search heuristics are often used.

Both approaches, however, suffer from the inherent instability in structure learning. That is, small changes in finite data samples can lead to entirely different optimal structures. The errors that occur in the algorithm may be propagated and lead to further errors [11].

In our previous paper [13], we developed a robust causal discovery algorithm, designed for cross-sectional data. The method, which is a score-based approach, searches over structures represented by Structural Equation Models (SEMs), a primary language for causal discovery [14]. We subsample the data and search Pareto optimal models for each subset based on the scores. The optimal models are characterized by two objectives: the model fit and the model complexity. These two objectives, however, are often conflicting as a model that fits the data well typically has complex structure. Thus, we apply a multi-objective evolutionary algorithm to search for the optimal models. We then use some thresholds to filter out the so-called relevant model structures from these optimal models. That is, the structures that are considerably stable and parsimonious. At the end, we infer a causal model from these relevant model structures, including reliability scores of each causal relationship.

In this present paper we extend the method to longitudinal data and provide estimation of causal effects. In particular, we explain how to model longitudinal causal relationships from longitudinal data consisting of an arbitrary number of time slices. We also describe how to reshape the longitudinal data correspondingly, so as to match the longitudinal causal model which then can easily be scored using standard SEM software. Moreover, we employ IDA [15] to estimate the causal effects from the inferred causal structure.

To validate, we compare our method with the state-of-the-art PC algorithm [10] on a simulated data set with the known ground truth model. Furthermore, we present the results of our causal discovery algorithm on three real-world longitudinal data sets about chronic fatigue syndrome, Alzheimer disease and chronic kidney disease that have been corroborated by medical experts and literatures.

The rest of this paper is organized as follows. All methods used in our approach are presented in Section 2. The results and the corresponding discussions are presented in Section 3. Finally, conclusions and future work are presented in Section 4.

2 METHODS

2.1 Causal Modeling in SEM

According to [14], there are two common ways to represent a SEM model: drawing them as a causal diagram (graph), see Figure 1, also referred to as a directed acyclic graph (DAG), or using a set of equations stating all relations, called a causal model, which has the following general form:

$$x_i = f_i(p_a, \varepsilon_i), \quad i = 1, \ldots, p.$$  (1)
where $pa_i$ denotes the parents, representing the set of variables considered to be direct causes of $X_i$, and $\epsilon_i$ represents errors on account of omitted factors that are assumed to be mutually independent.

Typically, a SEM is used as follows: 1) set a hypothesis as an initial model, 2) fit the model to the data, 3) evaluate the model, and 4) modify the model to improve the parsimony and score [16]. The last step is called specification search [17], [18].

Moreover, it is often that one has prior knowledge about the domain, for example, that $A$ does not cause $B$ directly, denoted by $A \not\rightarrow B$ which can be translated to a DAG with no directed edge from $A$ to $B$. Our method [13] is able to include such constraints in the models, thus it can work with constrained SEMs.

### 2.2 Multi-objective Optimization

In model structure learning, in line with the principle of Occam’s razor, one should prefer models that fit the data well and are parsimonious. These two objectives are often conflicting. Thus in our method [13], we make as a well-fit model is most likely to be a parsimonious. These two objectives are often conflicting models that fit the data well and are parsimonious.

In multi-objective optimization, optimal solutions are defined in terms of domination. A model $x_1$ is said to dominate model $x_2$, if the following conditions are satisfied [20]:

$$x_1 \preceq x_2 \iff \forall i \in \{1, \ldots, M\} \ f_i(x_1) \leq f_i(x_2) \land \exists j \in \{1, \ldots, M\} \ f_j(x_1) < f_j(x_2)$$

where $M$ is the number of the objectives. The first condition states that the model $x_1$ is no worse than $x_2$ in all objectives $f_i$. The second condition states that the model $x_1$ is strictly better than $x_2$ in at least one objective. By using this concept, given the population of models $P$, we can partition $P$ into $n$ sets called fronts $F_1, \ldots, F_n$, such that $F_k$ dominates $F_l$ where $1 \leq k < l \leq n$ and the models within the same front do not dominate each other. The so-called Pareto Front or non-dominated set $F_1$ includes models that are not dominated by any member of $P$. Essentially, using multi-objective optimization we efficiently find the best fitting models over a whole range of model complexities using a single coherent optimization approach.

### 2.3 Stable Specification Search

Our method, stable specification search [13], can be divided into two phases (see Figure 2). The first phase is search, performing exploratory search over SEM models. Based on the idea of stability selection [21], the method subsamples the data $D$ with size $|D|/2$ without replacement and generates Pareto optimal models for each subset. After that, all Pareto optimal models are transformed into their corresponding model equivalent classes, called Completed Partially Directed Acyclic Graph (CPDAG) [22]. Since all DAGs that are a member of the same CPDAG represent the same probability distribution, they are indistinguishable based on the observational data alone. From these CPDAGs we compute the edge and causal path stability graph, such as Figure 6 by grouping them according to model complexity and computing their selection probability, i.e., the number of occurrences divided by the total number of models for a certain level of model complexity. Stability selection is then performed by specifying two thresholds, $\pi_{sel}$ (boundary of selection probability) and $\pi_{bic}$ (boundary of complexity). For example, setting $\pi_{sel} = 0.6$ means that all causal relationships with edge stability or causal path stability (e.g., Figure 6) above this threshold are considered stable. The second threshold $\pi_{bic}$ is used to control overfitting. We set $\pi_{bic} = 7$ means that all causal relationships with an edge stability or a causal path stability lower than this threshold (e.g., Figure 6) are considered parsimonious. Causal relationships that intersect with the top-left region are considered
both stable and parsimonious and called relevant, from which we can derive a causal model.

The second phase is visualization, combining the stability graphs into a graph with nodes and edges. This is done by adding the relevant edges and orienting them using background knowledge and the relevant causal paths. For more detail of the method, Figure 3 provides pseudocode for our approach.

2.4 Estimating Causal Effects
We employ IDA [15] to estimate the causal effects of a covariate $X_i$ on response $Y$ from the inferred structures. This method works as follows. Given a CPDAG $G$ which contains $m$ different DAGs, IDA applies intervention calculus [14], [23] to each DAG $G_j$ to obtain multisets $\Theta_i = \{\theta_{ij}\}_{j=1,...,m}$, $i = 1, \ldots, p$ where $p$ is the number of covariates, which contain possible causal effects of $X_i$ on $Y$. Different graphs $G_j$ may imply different $\text{pa}_i$ of $X_i$ which then lead to distinct values of $\theta_{ij}$.

Under the assumption that the distribution of the data is normal and the model is linear, the causal effect of $X_i$ on $Y$ is given by $\beta_{i|\text{pa}_i}$, where, for any set $S \subseteq \{X_1, \ldots, X_p, Y\} \setminus \{X_i\}$,

$$\beta_{i|S} = \begin{cases} 0, & \text{if } Y \in S \\ \text{coefficient of } X_i \text{ in } Y \sim X_i + S, & \text{if } Y \notin S, \end{cases}$$

and $Y \sim X_i + S$ is the linear regression of $Y$ on $X_i$ and $S$.

In addition to our method, we apply IDA as follows. We gather $G_{\pi_{bic}}$, the CPDAGs of all optimal models with complexity equal to $\pi_{bic}$. For each CPDAG $G \in G_{\pi_{bic}}$, we compute possible causal effects $\Theta$ of each relevant causal path using IDA, say $X \rightarrow Y$, yielding $\Theta_{k_{X \rightarrow Y}}$, $k = 1, \ldots, N$, where $N$ is the number of subsets (see Figure 2). All causal effect estimations in $\Theta_{k_{X \rightarrow Y}}$ are then concatenated into a single multiset $\Theta_{X \rightarrow Y}$. To represent the estimated causal effects from $X$ to $Y$, we compute the median $\tilde{\Theta}_{X \rightarrow Y}$ and if both $X$ and $Y$ are continuous variables we standardize the estimation using

$$\frac{\tilde{\Theta}_{X \rightarrow Y} \cdot \sigma_X}{\sigma_Y},$$

where $\sigma_X$ and $\sigma_Y$ are the standard deviations of the covariate and the response, respectively. A standardized causal effect allows us to meaningfully compare them.

2.5 “Unrolled” Network
Based on the idea of “unrolling” the network in Dynamic Bayesian Networks [8], [9], we have extended our method to handle longitudinal data. We model longitudinal causal relationships with a SEM model consisting of two time slices (Figure 4a) that can be “unrolled” into a network with an arbitrary number of time slices (Figure 4b). Time slice $t_i$ represents the relationships within a time slice (intra-slice causal relationships, solid arcs in Figure 4a). Causal relationships between time slices (interslice causal relationships, dashed arcs in Figure 4a) always go forward in time, i.e., from time slice $t_{i-1}$ to time slice $t_i$.

2.6 Data Reshaping
To score our models on longitudinal data, we use data reshaping. In the reshaped data, the first $n$ data points contain the relations that occur in the first two time slices $t_0$ and $t_1$. The next $n$ data points contain the relations that occur in time slices $t_1$ and $t_2$. The $i$-th subset of $n$ data points contain the relations in time slices $t_{i-1}$ and $t_i$. The reshaped data then allows us to use standard SEM software to compute the scores.

3 Results and Discussion

3.1 Implementation
We implemented our method in [13] as well as the extension described in this present paper as an R package named stablespec. The package is publicly available at the Comprehensive R Archive Network (CRAN) [1] so it can be installed directly, e.g., from R console by typing install.package("stablespec") or from RStudio by using feature to install package. We also included a package documentation as a brief tutorial of using the functions.

1. https://cran.r-project.org/web/packages/stablespec/index.html
Fig. 2: The proposed method consists of two phases: search and visualization. The search phase is an iterative process using an outer loop and inner loop that combines SEM, NSGA-II, and stability selection, which outputs all relevant edges and causal paths between two variables. The visualization phase displays the relevant relationships as a causal model.

```plaintext
1: procedure stableSpecificationSearch(data set D, constraint C)
2:   \( H \leftarrow () \) \( \triangleright \) initialize
3:   for \( j \leftarrow 0, \ldots, J - 1 \) do \( \triangleright \) \( J \) is number of outer loop iterations
4:     \( T \leftarrow \) subset of \( D \) with size \( \lceil |D|/2 \rceil \) without replacement
5:     \( F_1 \leftarrow () \) \( \triangleright \) initialize Pareto fronts to empty list
6:     for \( i \leftarrow 0, \ldots, I - 1 \) do \( \triangleright \) \( I \) is number of inner loop iterations
7:       if \( i = 0 \) then
8:         \( P \leftarrow N \) random DAGs consistent with \( C \)
9:         \( P \leftarrow \) fastNonDominatedSort(\( P \))
10:      else
11:         \( P \leftarrow \) crowdingDistanceSort(\( F \)) \( \triangleright \) draw the first \( N \) models
12:      end if
13:     \( Q \leftarrow \) make population from \( P \)
14:     \( F \leftarrow \) fastNonDominatedSort(\( P \cap Q \))
15:     \( F_1 \leftarrow \) pareto front of \( F \) and \( F_1 \)
16:   end for
17:   \( H \leftarrow H \cap F_1 \) \( \triangleright \) concatenation
18: end for
19: \( G \leftarrow \) consDag2Cpdag(\( H, C \))
20: edges \leftarrow \) edge stability of \( G \)
21: paths \leftarrow \) path stability of \( G \)
22: end procedure
```

Fig. 3: Stable specification search consists of an outer and an inner loop. The outer loop samples a subset of the data, and for every subset, the inner loop searches for the Pareto front by applying NSGA-II. The Pareto fronts are converted into constrained CPDAGs which are then used to compute the edge and causal path stability graph.
Fig. 4: (a) The longitudinal causal model. (b) The “unrolled” causal graph used to generate longitudinal data. It contains four continuous variables \((X_1, \ldots, X_4)\) in three different time slices \(t_0, \ldots, t_2\).

### 3.2 Parameter Settings

For application to simulated data and real-world data, we subsampled 50 and 100 subsets from the data with size \(\lceil |D|/2 \rceil\), respectively. We did not do comprehensive parameter tuning for NSGA-II, instead, we followed guidelines provided in [24]. The parameters for the application to simulated longitudinal data were set as follows: the number of generations was 30, the size of the population \(P\) was 100, the crossover rate was 0.85, the mutation rate was 0.1, and the selection strategy was binary tournament selection. The parameters for the application to real-world longitudinal data were similar to above except that the number of generations was 35, the size of the population \(P\) was 150, and the mutation rate was 0.07.

We score models using the chi-square \(\chi^2\) and the model complexity. The \(\chi^2\) is considered the original fit index in SEM and measures how close the model-implied covariance matrix is to the sample covariance matrix [25]. The model complexity represents how many predicted parameters the model contains. Assuming that variances of parameters are always predicted, the maximum model complexity with \(p\) variables is given by \(p(p-1)/2\).

When using multi-objective optimization we minimize both the \(\chi^2\) and model complexity objectives. These two objectives are, however, conflicting with each other. For example, minimizing the model complexity typically means compromising the data fit.

### 3.3 Application to Simulated Data

For this experiment, we generated a longitudinal data set with 200 instances from a causal graph as depicted in Figure 4b. The data set consists of three time slices with four continuous variables for each time slice \(\bar{X}_0\) and \(\bar{X}_2\). We searched over SEM models we added prior knowledge that variables \(X_1\) at \(t_i\) and \(X_2\) at \(t_i\) do not cause variable \(X_3\) at \(t_i\) directly.

Our method subdivides the data into a number of subsets, here 50 of size 100 instances, and then runs the NSGA-II to obtain 50 Pareto fronts. In order to conduct a fair comparison, we consider a subsampling version of the PC algorithm, as in [26], with also 50 subsets. We repeated this procedure 10 times and computed the average of both edge and causal path stability.

As the true model is known, we measure the performance of our method by means of the Receiver Operating Characteristic (ROC) [27] for both edges and causal paths. We compute the True Positive Rate (TPR) and the False Positive Rate (FPR) based on the CPDAG of the true model. As for an example, in the case of edge stability, a true positive means that an edge obtained by our method or the PC algorithm is present in the CPDAG of the ground truth.

To compare the ROC curves of our method and those of the PC algorithm, we employed three significance tests. The first two tests, as introduced in [28] and in [29], compare the

2. Available at http://tinyurl.com/hd3wf32
Area Under the Curve (AUC) of the ROC curves by using the theory of U-statistics and bootstrap replicates, respectively. The third test \(^{[30]}\) compares the actual ROC curves by evaluating the absolute difference and generating rank-based permutations to compute the statistical significance. The null hypothesis is that (the AUC of) the ROC curves of our method and the PC algorithm are identical.

Figure 5 portrays the ROC curves for both edge and causal path stability. The black curves and the red curves represent our method’s and the PC algorithm’s results, respectively. For edge and causal path stability, the AUCs of our method are 0.71 and 0.82, respectively, and those of the PC algorithm 0.53 and 0.56. Table 1 contains the results of the significance tests. Following the \(p\)-values, the observed differences in the ROC curves for the edge stability are significant on the first and the second test, and are marginally significant on the third test. Moreover, the differences on the causal path stability are highly significant on all tests. Therefore, we can reject the null hypothesis and conclude that our method outperforms the PC algorithm on these simulated data.

### 3.4 Application to Chronic Fatigue Syndrome Data

For the first application to real-world data, we consider a longitudinal data set about chronic fatigue syndrome (CFS). The longitudinal data consists of 183 subjects and five time slices with six discrete variables \(^{[31]}\). The variables are, fatigue severity, the sense of control over fatigue, focusing on the symptoms, the objective activity of the patient (oActivity), the subject’s perceived activity (pActivity), and the physical functioning. We use Expectation Maximization (EM) to impute the missing values. As all of the variables have large scales, e.g., in the range between 0 to 155, we treat them as continuous variables. We added prior knowledge that the variable fatigue at \(t_i\) does not cause any of the other variables within \(t_i\) directly. We performed the search over 100 subsamples of the original data set.

We set \(\pi_{sel} = 0.6\) and found \(\pi_{bic} = 27\). Figure 6a shows that nineteen relevant edges were found, consisting of eleven intra-slice (blue lines) and eight inter-slice relationships of which six are between the same variables (orange lines) and two are between different variables (black lines). Figure 6b shows that thirty-five relevant causal paths were found, consisting of twelve intra-slice (blue lines) and twenty-three inter-slice relationships of which six are between the same variables (orange lines) and seventeen are between different variables (black lines). For a more intuitive representation, we combine the stability graphs into a causal model using the following procedure. First, the nodes are linked according to the nineteen relevant edges. Second, edges are oriented according to our background knowledge; eight of the inter-slice relationships are oriented from time slice \((t_{i-1})\) to \(t_i\) and the fact that the variable fatigue at \(t_i\) does not directly cause any other variable at \(t_i\) results in five oriented intra-slice edges, which in this case corresponds exactly to the relevant causal paths obtained. Third, the edges are oriented according to the relevant causal paths, which results in another six directed edges. The inferred model is shown in Figure 7. Each edge is annotated with a reliability score (its maximum selection probability in the top-left region of the edge stability graph) and a standardized causal effect estimation. For example, annotation “1/0.5” represents a reliability score equal to 1 and a standardized causal effect equal to 0.5.

The intra-slice and inter-slice causal relationships that we obtained were both relatively high, i.e., the selection probabilities are all pretty close to 1. In the intra-slice causal relationships, we found that all variables are direct causes for fatigue severity. We also found that all variables, except fatigue, are direct causes for the pActivity. Furthermore, the variable control is a direct cause for both focusing on the symptoms and physical functioning. Generally the inter-slice relationships show direct causes between the same variables. In addition, the variables pActivity and control also indicate an inter-slice causal effect. The standardized causal effects estimate to which extent a covariate would cause a response. For example, as shown in Figure 7, the more sense of control over fatigue a subject has, the less fatigue the subject would experience. Overall, the inferred
Fig. 5: ROC curves for (a) the edge stability and (b) the causal path stability. In (a), the AUCs of our method and the PC algorithm are 0.71 and 0.53, respectively. In (b), the AUCs of our method and the PC algorithm are 0.82 and 0.56, respectively.

TABLE 1: Comparison of ROC curves of our method and the PC algorithm.

| Statistics Test | ROC for the edge stability | ROC for the causal path stability |
|-----------------|-----------------------------|----------------------------------|
| DeLong [28]     | Z = 1.95 and p-value = 0.05 | Z = 3.50 and p-value = 0.0004   |
| Bootstrap [29]  | D = 1.98 and p-value = 0.04 | D = 3.65 and p-value = 0.0003   |
| Venkatraman [30]| E = 56 and p-value = 0.09   | E = 256 and p-value = 0.0035    |

structure and the estimated causal effects are consistent with results reported in the medical literature [31], [32], [33].

3.5 Application to Alzheimer’s Disease Data

For the second application to real-world data, we consider a longitudinal data set about Alzheimer’s Disease (AD), which is provided by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) [34], and can be accessed at adni.loni.usc.edu. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information see www.adni-info.org.

Here, we focus on patients with MCI, an intermediate clinical stage in AD [35]. Following a hypothesis discussed in [36], we include only the variables of interest, resulting in 179 subjects with four continuous variables and six time slices. The variables are, the subject’s cognitive dysfunction (ADAS-Cog), the subject’s hippocampal volume (hippocampal_vol), the subject’s whole brain volume (brain_vol), and the subject’s brain glucose metabolism (FDG-PET). Like in the application to CFS, we use EM to impute the missing values. We added prior knowledge that the variable ADAS-Cog does not cause any of the other variables directly. We performed the search over 100 subsamples of the original data set.

We set $\pi_{sel} = 0.6$ and found $\pi_{bic} = 12$. As shown in Figure 8a, we obtained twelve relevant edges, consisting of four intra-slice (blue lines) and eight inter-slice relationships of which four are between the same variables (orange lines) and four are between different variables (black lines). As shown in Figure 8b, we obtained seventeen relevant causal paths, consisting of six intra-slice (blue lines) and eleven inter-slice relationships of which four are between the same variables (orange lines) and seven are between different variables (black lines). Likewise the application to
Fig. 6: The stability graphs for CFS together with $\pi_{\text{sel}}$ and $\pi_{\text{bic}}$, yielding four regions. The top-left region contains the relevant causal relations. (a) The edge stability graph. (b) The causal path stability graph. Orange lines represent inter-slice relationships between the same variables, black lines represent inter-slice relationships between different variables, blue lines represent intra-slice relationships.

CFS, we visualize the inferred causal model by first linking the nodes according to the twelve relevant edges. Second, edges are oriented following our background knowledge; eight of the inter-slice relationships are oriented from time slice $t_{i-1}$ to $t_i$ and the fact that the variable ADAS-Cog at $t_i$ does not directly cause any other variable at $t_i$ results in one directed edge from brain_vol to ADAS-Cog. Third, the edges are then oriented based on the relevant causal paths, giving another three directed edges. The direction of the edge from FGD-PET to brain_vol follows because we do not allow cycles in our model. The final result of our procedure is shown in Figure 9. Each (directed) edge in the inferred causal model is annotated with its reliability score and its standardized causal effect.

Fig. 7: The inferred model of CFS by combining the edge stability and causal path stability graphs. Each edge is annotated by its reliability score (its maximum selection probability in the top-left region of the edge stability graph) and its standardized causal effect. For example, the annotation “1/0.71” represents a reliability score of 1 and a standardized causal effect of 0.71.

Similar to the result on CFS, we also obtained the intra-slice and inter-slice causal relationships with selection probabilities that are all pretty close to 1 (see Figure 9). In the
Fig. 8: The stability graphs for AD together with $\pi_{\text{sel}}$ and $\pi_{\text{bic}}$, yielding four regions. The top-left region contains the relevant causal relations. (a) The edge stability graph. (b) The causal path stability graph. Orange lines represent inter-slice relationships between the same variables, black lines represent inter-slice relationships between different variables, blue lines represent intra-slice relationships.

Fig. 9: The inferred model of AD by combining the edge stability and causal path stability graphs. The dash line represents a strong relation between two variables but the causal direction is undecidable. Each edge is annotated by its reliability score (its maximum selection probability in the top-left region of the edge stability graph) and its standardized causal effect. For example, the annotation "1/0.81" represents a reliability score of 1 and a standardized causal effect of 0.81.

Intra-slice causal relationships, we found that change in \textit{brain vol} is a direct cause of change in \textit{ADAS-Cog}. Both \textit{hippocampal vol} and \textit{FDG-PET} cause change in \textit{ADAS-Cog} through some intermediate variables. In addition, there is a strong direct relation between \textit{hippocampal vol} and \textit{FDG-PET}. In the inter-slice relationships, we found direct causes between the same variables, which is expected. Moreover, the change in \textit{hippocampal vol} and \textit{brain vol} are direct causes of \textit{ADAS-Cog}, while \textit{hippocampal vol} and \textit{ADAS-Cog} are direct causes of \textit{brain vol}. As shown in Figure 9, an increase in a subject’s hippocampal volume would result in a reduction in the subject’s cognitive impairment. In addition to the directed edge from \textit{hippocampal vol} to \textit{brain vol} in the intra-slice causal relationship, the standardized causal effect of 0 means the sign is undecidable whether it is positive or negative. The inferred structure and the estimated causal effects are consistent with results reported in the medical literature [36], [37], [38], [39].
3.6 Application to Chronic Kidney Disease

For the third application to real-world data, we consider a longitudinal data set about chronic kidney disease (CKD), provided by the MASTERPLAN study group [40]. The MASTERPLAN study was initiated in 2004 as a randomized, controlled trial studying the effect of intensified treatment with the aid of i.e., nurse practitioners on cardiovascular and kidney outcome in CKD. This intensified treatment regimen addressed eleven possible risk factors for the progression of CKD simultaneously. The study previously showed that this intensified treatment resulted in fewer patients reaching end stage kidney disease compared to standard treatment [40].

Here we focussed on the potential causal mediators for the protective effect incurred by the intensified treatment with the aid of nurse practitioners. In other words, we aimed to identify which of the treatment targets contributed to the observed overall treatment effect. In the present analysis, we include only variables of interest being treatment status, either nurse practitioner aided care or standard care, as allocated by the randomization procedure (treatment), estimated glomerular filtration rate (gfr)—a marker for overall kidney function, and a variable indicating informative censoring (inf_cens). Informative censoring occurred when patients reached end stage kidney disease requiring renal replacement therapy, such as dialysis or a kidney transplantation, or when they died. Furthermore, we considered treatment targets that were previously hypothesized to contribute most to the overall treatment effect: systolic blood pressure (sbp), LDL-cholesterol (ldl) and para-thyroid hormone (pth) concentrations in blood, and protein excretion via urine (pcr). In total, there are 497 subjects with seven variables (both continuous and discrete) over five time slices. Particularly we set the variable treatment only at \( t_{i-1} \) as it remains the same over all time slices, and the variable inf_cens only at \( t_i \) as it is a consequence of previous treatment or event, e.g, renal replacement or death. We further added the prior knowledge that gfr at \( t_i \) does not directly cause any other variables at \( t_i \), and that there are no relations between any variable and inf_cens within \( t_i \). Both gfr and inf_cens are read-out for CKD progression and are within a time slice always the consequence and never the cause of another variable. Like in the applications to CFS and AD, we imputed the missing values using EM. We performed the search over 100 subsamples of the original data set.

We set \( \pi_{sel} = 0.6 \) and found \( \pi_{bic} = 23 \). Based on Figure 10a, we obtained seventeen relevant edges, consisting of four intra-slice (blue lines) and thirteen inter-slice relationships of which five are between the same variables (orange lines) and eight are between different variables (black lines). Based on Figure 10b, we obtained twenty-six relevant causal paths, consisting of five intra-slice (blue lines) and twenty-one inter-slice relationships of which five are between the same variables (orange lines) and sixteen are between different variables (black lines). As in the applications to CFS and AD, we visualize the inferred causal model by first linking the nodes according to the seventeen relevant edges. Second, the edges are oriented following our background knowledge; thirteen of the inter-slice relationships are oriented from time slice \( t_{i-1} \) to \( t_i \) and the fact that the variable gfr at \( t_i \) causes nothing else at \( t_i \) results in two directed edges: from sbp to gfr and from pth to gfr. Third, the edges are oriented according to the relevant causal paths, yielding another directed edge from pcr to pth. There is one relevant edge between pcr and sbp which direction is undecidable. This particular edge represents a strong association between these two variables. When we lowered \( \pi_{sel} \) from 0.6 to 0.5, we found a causal path from sbp to pcr, suggesting that this direction may be the more likely.

The inferred model is shown in Figure 11. Each edge is annotated with a reliability score and its standardized causal effect estimation. Most of the intra-slice and inter-slice causal relationships are very stable with selection probabilities close to 1. We found inter-slice causal relationships between gfr, pth, pcr, and inf_cens. Furthermore, gfr, sbp and pcr are well known determinants for CKD progression. However, the causal relationship between pth at \( t_{i-1} \) and inf_cens was somewhat surprising, but as pth is
Fig. 10: The stability graphs for CKD together with $\tilde{\pi}_{bic}$ and $\pi_{sel}$ yielding four regions. The top-left region contains the relevant causal relations. (a) The edge stability graph. (b) The causal path stability graph. Orange lines represent inter-slice relationships between the same variables, black lines represent inter-slice relationships between different variables, blue lines represent intra-slice relationships.

Fig. 11: The inferred model of CKD by combining the edge stability and causal path stability graphs. The dash line represents a strong relation between two variables but the causal direction is undecidable. Each edge is annotated by its reliability score (its maximum selection probability in the top-left region of the edge stability graph) and its standardized causal effect. For example, the annotation “0.80/0.07” represents a reliability score of 0.80 and a standardized causal effect of 0.07. In addition, the blue-dashed (undirected) edge represents a strong association between two variables, but the direction is undecidable.

a marker for regulation of phosphate stores in the body and related to overall vascular damage and thereby to mortality. Indeed, literature indicates that lowering $pth$ in dialysis patients resulted in a slight reduction in mortality [41]. The same may hold true for patient who have CKD and who do yet need dialysis treatment. Perhaps most surprising are the relations between $sbp$ and $pcr$ and $gfr$, respectively. From renal physiology we know that higher filtration pressures due to higher blood pressure causes the short term glomerular filtration rate to increase. Likewise, at higher filtration pressure, more and larger proteins are pushed out of the blood stream and into the pro-urine and are ultimately excreted via the urine. In the long term, chronically elevated filtration pressures and elevated levels of protein in the pro-urine
cause kidney damage and ultimately even end stage kidney disease. Overall, the results are consistent with literature and physiology.

4 Conclusion and Future Work

Causal discovery from longitudinal data is an important, ubiquitous problem in many disciplines. In the medical domain, especially in the case of rare diseases, revealing causal relationships from a given data set may lead to improvement of clinical practice, e.g., development of treatment and medication. In the past decades, many causal discovery algorithms have been introduced. These causal discovery algorithms, however, have difficulty dealing with the inherent instability in structure estimation.

The present work introduces a new discovery algorithm for longitudinal data that is robust for finite samples, extending previous method in [13]. Experimental results on both simulated and real-world data sets show that the method results in reliable structure estimates. On simulated data our method outperforms a bootstrap version of the state-of-the-art PC algorithm. Furthermore, the causal structures and causal effect estimates obtained from real-world data on CFS, AD, and CKD patients are corroborated by literature studies [31], [32], [33], [36], [37], [38], [39], [41]. However, the current method considers only longitudinal data with observed variables and cannot handle missing values (other than through imputation as a preprocessing step). Moreover, we still assume that the time intervals between time slices is fairly uniform between subjects. Some existing approaches called random-coefficient models, also termed multi-level or hierarchical regression models [42], [43], are flexible to handle unequal interval between time slices within a subject and/or across subjects. Future research will aim to account for these aforementioned issues.

Acknowledgments

The research leading to these results has received funding from the DGHE of Indonesia and the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 305697.

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] B. van Engelen et al., “Cognitive behaviour therapy plus aerobic exercise training to increase activity in patients with myotonic dystrophy type 1 (dm1) compared to usual care (optimistic): study protocol for randomised controlled trial,” Trials, vol. 16, no. 1, p. 224, 2015.

[2] B. Clarke, “Causality in medicine with particular reference to the viral causation of cancers,” Ph.D. dissertation, UCL (University College London), 2011.

[3] K. D. Hoover, “Causality in economics and econometrics,” The new Palgrave dictionary of economics, vol. 2, 2008.
[4] S. Abu-Bader and A. S. Abu-Qarn, “Government expenditures, military spending and economic growth: causality evidence from Egypt, Israel, and Syria,” Journal of Policy Modeling, vol. 25, no. 6, pp. 567–583, 2003.

[5] A. Riva and R. Bellazzi, “Learning temporal probabilistic causal models from longitudinal data,” Artificial Intelligence in Medicine, vol. 8, no. 3, pp. 217–234, 1996.

[6] J. Parner and E. Arjas, Causal reasoning from longitudinal data. Rolf Nevanlinna Inst., University of Helsinki, 1999.

[7] H. W. Marsh and A. S. Yeung, “Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data.” Journal of educational psychology, vol. 89, no. 1, pp. 41–54, 1997.

[8] N. Friedman, K. Murphy, and S. Russell, “Learning the structure of dynamic probabilistic networks,” in Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1998, pp. 139–147.

[9] K. Murphy, S. Mian et al., “Modelling gene expression data using dynamic Bayesian networks,” Computer Science Division, University of California, Berkeley, CA, Tech. Rep., 1999.

[10] P. Spirtes, C. N. Glymour, and R. Scheines, Causation, prediction, and search. MIT press, 2000, vol. 81.

[11] P. Spirtes, “Introduction to causal inference,” The Journal of Machine Learning Research, vol. 11, pp. 1643–1662, 2010.

[12] D. Heckerman, C. Meek, and G. Cooper, “A bayesian approach to causal discovery,” Computation, causation, and discovery, vol. 19, pp. 141–166, 1999.

[13] R. Rahmadi, P. Groot, M. Heins, H. Knoop, T. Heskes, and The OPTIMISTIC consortium, “Causality on cross-sectional data: Stable specification search in constrained structural equation modeling.” vol. arXiv:1506.05600 [stat.ML], 2015.

[14] J. Pearl, Causality: models, reasoning and inference. Cambridge Univ Press, 2000.

[15] M. H. Maathuis, M. Kalisch, P. Bühlmann et al., “Estimating high-dimensional intervention effects from observational data,” The Annals of Statistics, vol. 37, no. 6A, pp. 3133–3164, 2009.

[16] R. MacCallum, “Specification searches in covariance structure modeling.” Psychological Bulletin, vol. 100, no. 1, p. 107, 1986.

[17] E. E. Leamer, Specification searches: Ad hoc inference with nonexperimental data. Wiley New York, 1978.

[18] J. S. Long, Covariance structure models: An introduction to LISREL. Sage, 1983, no. 34.

[19] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182–197, 2002.

[20] K. Deb, Multi-objective optimization using evolutionary algorithms. John Wiley & Sons Chichester, 2001, vol. 2012.

[21] N. Meinshausen and P. Bühlmann, “Stability selection,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 72, no. 4, pp. 417–473, 2010.

[22] D. M. Chickering, “Learning equivalence classes of Bayesian-network structures,” The Journal of Machine Learning Research, vol. 2, pp. 445–498, 2002.

[23] J. Pearl, “Statistics and causal inference: A review.” Test, vol. 12, no. 2, pp. 281–345, 2003.

[24] J. J. Grefenstette, “Optimization of control parameters for genetic algorithms,” Systems, Man and Cybernetics, IEEE Transactions on, vol. 16, no. 1, pp. 122–128, 1986.

[25] R. Kline, Principles and Practice of Structural Equation Modeling, ser. Methodology in the social sciences. Guilford Press, 2011.

[26] J. Ramsey, “Bootstrapping the PC and CPC algorithms to improve search accuracy,” 2010.

[27] T. Fawcett, “ROC graphs: Notes and practical considerations for researchers,” Machine learning, vol. 31, pp. 1–38, 2004.

[28] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach,” Biometrics, pp. 837–845, 1988.

[29] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, “proc: an open-source package for R and S+ to analyze and compare ROC curves,” BMC bioinformatics, vol. 12, no. 1, p. 77, 2011.

[30] E. Venkatraman and C. B. Begg, “A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment,” Biometrika, vol. 83, no. 4, pp. 835–848, 1996.

[31] M. J. Heins, H. Knoop, W. J. Burk, and G. Bleijenberg, “The process of cognitive behaviour therapy for chronic fatigue syndrome: Which changes in perpetuating cognitions and behaviour are related to a reduction in fatigue?” Journal of psychosomatic research, vol. 75, no. 3, pp. 235–241, 2013.

[32] J. Vercoulen, C. Swanink, J. Galama, J. Fennis, P. Jongen, O. Hommes, J. Van der Meer, and G. Bleijenberg, “The persistence of fatigue in chronic fatigue syndrome and multiple sclerosis: development of a model,” Journal of psychosomatic research, vol. 45, no. 6, pp. 507–517, 1998.

[33] J. F. Wiborg, H. Knoop, L. E. Frank, and G. Bleijenberg, “Towards an evidence-based treatment model for cognitive behavioral interventions focusing on chronic fatigue syndrome,” Journal of psychosomatic research, vol. 75, no. 5, pp. 399–404, 2012.

[34] M. W. Weiner, P. S. Aisen, C. R. Jack, W. J. Jagust, J. Q. Trojanowski, L. Shaw, A. J. Saykin, J. C. Morris, N. Cairns, L. A. Beckett et al., “The alzheimer’s disease neuroimaging initiative: progress report and future plans,” Alzheimer’s & Dementia, vol. 6, no. 3, pp. 202–211, 2010.

[35] R. C. Peterson, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, “Mild cognitive impairment: clinical characterization and outcome,” Archives of neurology, vol. 56, no. 3, pp. 303–308, 1999.

[36] T. J. Haight, W. J. Jagust, A. D. N. Initiative et al., “Relative contributions of biomarkers in alzheimer’s disease,” Annals of epidemiology, vol. 22, no. 12, pp. 868–875, 2012.

[37] W. Henneman, J. Sluimer, J. Barnes, W. Van Der Flier, I. Sluimer, N. Fox, P. Scheltens, H. Vrenken, and F. Barkhof, “Hippocampal atrophy rates in alzheimer disease added value over whole brain volume measures,” Neurology, vol. 72, no. 11, pp. 999–1007, 2009.

[38] D. Mungas, B. Reed, W. Jagust, C. DeCarli, W. Mack, J. Kramer, M. Weiner, N. Schuff, and H. Chui, “Volumetric mri predicts rate of cognitive decline related to ad and cerebrovascular disease,” Neurology, vol. 59, no. 6, pp. 867–873, 2002.

[39] H. Rusinek, S. De Santi, D. Frid, W.-H. Tsui, C. Y. Tarshish, A. Convit, and M. J. de Leon, “Regional brain atrophy rate predicts future cognitive decline: 6-year longitudinal mr imaging study of normal aging 1,” Radiology, vol. 229, no. 3, pp. 691–696, 2003.

[40] M. J. Peeters, A. D. van Zuilen, J. A. van den Brand, M. L. Bots, M. van Buren, M. A. ten Dam, K. A. Kaasjager,
G. Ligtenberg, Y. W. Sijpkens, H. E. Sluiter et al., “Nurse practitioner care improves renal outcome in patients with ckd,” *Journal of the American Society of Nephrology*, vol. 25, no. 2, pp. 390–398, 2014.

[41] G. M. Chertow, G. A. Block, R. Correa-Rotter, T. B. Drueke, J. Floege, W. G. Goodman, C. A. Herzog, Y. Kubo, G. M. London, K. W. Mahaffey et al., “Effect of cinacalcet on cardiovascular disease in patients undergoing dialysis.” *The New England journal of medicine*, vol. 367, no. 26, pp. 2482–2494, 2012.

[42] S. W. Raudenbush and A. S. Bryk, *Hierarchical linear models: Applications and data analysis methods*. Sage, 2002, vol. 1.

[43] I. G. Kreft, I. Kreft, and J. de Leeuw, *Introducing multilevel modeling*. Sage, 1998.