Constructing causal life course models
Comparative study of data-driven and theory-driven approaches
Petersen, Anne Helby; Ekstrøm, Claus Thorn; Spirtes, Peter; Osler, Merete

Published in:
American Journal of Epidemiology

DOI:
10.1093/aje/kwad144

Publication date:
2023

Document version
Peer reviewed version

Document license:
Unspecified

Citation for published version (APA):
Title: Constructing causal life course models: Comparative study of data-driven and theory-driven approaches

Authors: Anne Helby Petersen, Claus Thorn Ekstrøm, Peter Spirtes, and Merete Osler

Correspondence Address: Correspondence to Anne Helby Petersen, Section of Biostatistics, Department of Public Health, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen K, Denmark (email: ahpe@sund.ku.dk)

Affiliations: Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark (Anne Helby Petersen, and Claus Thorn Ekstrøm); Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA (Peter Spirtes); Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospitals, Copenhagen, Denmark (Merete Osler); Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark (Merete Osler).

Funding: This work was funded by Independent Research Fund Denmark grant 8020-00031B and National Institutes of Health contract R01HL159805.

Data Availability Statement: The data are not available online for replication because they cannot be anonymized. Researchers interested in gaining access to the data may contact the Public Health Database at the Department of Public Health, University of Copenhagen.

Thanks: We owe great thanks to Ida Kim Wium-Andersen and Matilde Winther-Jensen for the time and effort they devoted to constructing the expert models.

Conference presentation: N/A

Preprint Information: N/A
Disclaimer: N/A

Conflict of Interest: None declared.

Running Head: Constructing causal life course models

Key words: causal discovery; life course study; causal model; structure learning; longitudinal studies; PC algorithm; directed acyclic graph

Abbreviations: TPC algorithm (Temporal Peter-Clark algorithm), DAG (Directed Acyclic Graph), CPDAG (Completed Partially Directed Acyclic Graph), TPDAG (Temporal Partially Directed Acyclic Graph), BMI (Body Mass Index), GLM (Generalized Linear Model), TPC-S (Temporal Peter Clark Search), TPC-P (Temporal Peter Clark pre-specified), PC Algorithm (Peter-Clark Algorithm), CHD (coronary heart disease), ADMG (Acyclic Directed Mixed Graph), FCI algorithm (Fast Causal Inference algorithm), PAG (Partial Ancestral Graph).
Constructing causal life course models: Comparative study of data-driven and theory-driven approaches

Anne Helby Petersen
Claus Thorn Ekstrøm
Peter Spirtes
Merete Osler

Abstract

Life course epidemiology relies on specifying complex (causal) models that describe how variables interplay over time. Traditionally, such models have been constructed by perusing existing theory and previous studies. By comparing data-driven and theory-driven models, we investigate whether data-driven causal discovery algorithms can help this process. We focus on a longitudinal dataset following a cohort of Danish men. The theory-driven models are constructed by two subject-field experts. The data-driven models are constructed by use of temporal Peter-Clark (TPC) algorithm. TPC utilizes the temporal information embedded in life course data.

We find that the data-driven models recover some, but not all, causal relationships included in the theory-driven expert models. The data-driven method is especially good at identifying direct causal relationships that the experts have high confidence in. Moreover, in a post-hoc assessment we found that most of the direct causal relationships proposed by the data-driven model, but not included in the theory-driven model, were plausible. Thus, the data-driven model may propose additional meaningful causal hypothesis that are new or have been overlooked by the experts. In conclusion, data-driven methods can aid causal model construction in life course epidemiology, and combining both data-driven and theory-driven methods can lead to even stronger models.

1 Introduction

Life course studies require specification of complex (causal) models in order to estimate causal relationships between exposures and outcomes of interest. These models are systems of assumptions about potential cause-effect relationships and they are typically produced by reviewing relevant theory and previous empirical studies. While this tried and tested strategy has generated numerous important results, it is fundamentally confirmatory by nature. Thus, it may suffer from confirmation bias, because the model specification by design must fit into existing theories of cause and effect across the life course.

Causal discovery (also referred to as (causal) structure learning) provides a data-driven alternative to theory-driven model specification. Here, statistical algorithms are used to infer causal models directly from observational data. There exists several correct and complete algorithms, which can infer all identifiable causal information from a given dataset, but the algorithms require certain assumptions to be fulfilled, some of which are not testable [1,2]. Moreover, it is not yet clear how the algorithmic guarantees of correctness translate to finite-sample data [3,4]. Hence, although causal discovery may in principle be tremendously useful for constructing causal life course models, its practical utility is still not well understood.

In this study, we aim to compare classical, theory-driven approach to constructing life course models with a data-driven causal discovery approach. Life course studies are an ideal use case for such a methodological comparison because the temporal information embedded in life course data eases the causal discovery task and helps guide theoretical model construction as well.
We focus the comparison on a life course epidemiological dataset based on the Metropolit Cohort (5). This is a longitudinal observational study following a cohort of Danish men born in 1953 with several data collections throughout their lives, as well as register-based follow-up. Two subject-matter experts construct theory-driven causal models for the variables in this dataset (without access to the actual data). We compare these expert models to models obtained by applying a causal discovery method called the temporal Peter-Clark (TPC) algorithm (6) to the dataset.

The article is structured as follows: First, we introduce causal model terminology and assumptions in Section 2. Next, we present the data and specify how the theory-driven and data-driven model constructions are conducted in Section 3. We present the resulting models in Section 4 and finally discuss strengths, limitations and directions for future research in Section 5.

2 Causal models and assumptions

We represent systems of causal assumptions using directed acyclic graphs (DAGs), and we also refer to these as causal models. We here provide a brief and informal introduction to relevant concepts and assumptions, and we refer to (7) for a more thorough description.

DAGs are useful for summarizing causal assumptions, and thereby guiding an appropriate statistical analysis targeting the causal estimands of interest (8,9), and using DAGs is becoming increasingly widespread in health and clinical research (10). A causal DAG uses nodes to represent random variables, while directed edges are used to indicate causal relationships. A directed edge from one variable X to another variable Y, denoted X → Y, means that X is a direct cause of Y (relative to the variables in the system). This means that intervening on X will have an effect on Y, but not vice versa, and that this effect is not entirely mediated via other variables in the system. Two variables are said to be adjacent if there exists an edge between them. A DAG constructed like this will fulfill the Markov property (2), which ensures that we can use the causal information in the DAG to deduce what conditional independence relationships must be fulfilled by data that come from the DAG.

Relying on DAGs to represent our causal models implies some assumptions about the causal mechanisms studied. First, we assume causal sufficiency, that is, that we are considering all relevant variables. More specifically, we need to include all confounding variables and we are not allowed to condition on any collider variables that are not in the DAG (see (9) for definitions). If this is not fulfilled, a DAG is not a meaningful representation of the causal mechanism (2). Secondly, we assume acyclicity, which means that there cannot be feedback loops between variables: X cannot be a cause of Y, if Y is also a cause of X. Note that these assumptions are always required when using DAGs over observed variables to represent causal models, and hence they are (sometimes implicitly) applied in all epidemiological studies utilizing DAGs.

DAGs are attractive causal model representations because they are relatively intuitive to work with, and because many useful statistical results exist for DAGs (11). However, it is not generally possible to infer a DAG from observational data. The problem is that different DAGs can imply the same conditional independencies, and hence potentially the same observed data; the structure embedded in a DAG is not sufficient to uniquely specify the statistical properties of observed data. Instead of inferring DAGs from the data, we therefore infer Markov equivalence classes of DAGs. A Markov equivalence class is the collection of all DAGs that produce data with the same conditional independencies. Fortunately, these equivalence classes are very well understood so we can still provide a detailed description of which causal models agree with a given observational dataset, and which causal models are not compatible with it.

A Markov equivalence class can be represented by a graph itself, namely a completed partially directed acyclic graph (CPDAG). A CPDAG has directed edges wherever all DAGs in the Markov equivalence class agree on the orientation of that edge, while they have undirected edges whenever at least two DAGs in the equivalence class disagree on the orientation of an edge. Note that this means that all DAGs in an equivalence class share the same skeleton, that is, there are edges between the same variables. Only certain edge orientations may differ among members from the same Markov equivalence class. Figure 1 provides an example of two DAGs that make up a Markov equivalence class along with the CPDAG that represents the equivalence class.

If additional external background knowledge is available, it may be possible to rule out certain members of the equivalence class. In this article, we focus on a specific type of external temporal information, which assigns each variable to one specific period in life (for example childhood or adulthood). This additional temporal information rules out certain edge orientations: it is not possible to have causal links against the direction of time. We refer to a CPDAG that respects such temporal information as a temporal partially directed acyclic graph (TPDAG). As an example, consider the CPDAG...
in Figure [1](C). If we assume that \(X_1 \) and \(X_2 \) belong to one period (e.g., childhood), while \(X_3 \) and \(X_4 \) belong to another period (adulthood), then we can rule about the (b) DAG from the equivalence class, as this model then has a causal link against the direction of time. Therefore, the (a) DAG is the only remaining member of the equivalence class, and hence the (a) DAG is the TPDAG for this causal mechanism.

In order to infer a CPDAG or TPDAG from observational data, an extra assumption is generally needed: Faithfulness. This assumption is perhaps better understood by its complement: A key example of non-faithfulness is to have causal effects with opposite signs cancel each other out. When we assume faithfulness, we assume that such phenomena do not occur. More generally, faithfulness rules out that two variables appear to be unrelated in the data (in the large sample limit), even though they are causally related in the data generating mechanism. More formally, the faithfulness assumption ensures that all conditional independencies observed in the data originate from causal unrelatedness in the data generating mechanisms [7], and it is hence the reverse implication of the Markov property.

We use a causal discovery algorithm to construct data-driven models. We want to stress that causal discovery algorithms do not violate the general rule of causal inference, namely correlation does not imply causation, but instead makes use of its complement; causation does in fact imply association (under the assumption of faithfulness). In other words, causal relationships in the data generating mechanism will leave behind certain associational patterns in the data. The totality of the associations found in a dataset can be used to rule out certain causal relationships and help us reconstruct a causal model for the potential relationships that remain. Under the additional assumption of causal sufficiency, a CPDAG (or TPDAG) summarizes what parts of the causal mechanism that can be reconstructed (directed edges) and which parts remain unknown (undirected edges). For a more formal introduction to how and why this works, we refer to [12].

3 Methods

3.1 Data

We use data from the longitudinal Metropolit study [5]. This study includes a cohort of Danish men born in the Copenhagen metropolitan area in 1953. Several data sources are combined: Information from birth registers, a study conducted in childhood (age 12), conscript board examination data (age 18-30), and survey information (age 51). We furthermore use information from the extensive Danish national registers. This provides information for an additional follow-up period after the final survey, until the men are 64 years old in 2017. The records from these data sources are matched using unique personal identification numbers. We only consider cohort members who participated in the 2004 survey, and who were furthermore alive and residing in Denmark at end of follow-up December 31, 2017. The study is thus retrospective.

We consider a total of 22 variables and each variable is assigned to one of 5 periods: Birth, childhood, youth, adulthood, or early old age (Table [1]). All variables are either numeric or binary, and we consider only complete cases, which gives us a sample size of \(n = 3145\). Web Appendix 1 provides further details about the data, including the exclusion process (13 % excluded due to missing or erroneous information, Web Figure 1) and marginal distributions (Web Table 1).

3.2 Theory-driven model construction

Two health researchers with experience in epidemiology were recruited to construct theory-driven models. We refer to them as the experts in the following. Expert 1 has a background in biology and a PhD in heart disease epidemiology, while expert 2 is a medical doctor specialized in psychiatry.

The two experts were provided with a list of the 22 variables, but no access to data, and were asked to construct a causal DAG from these variables. They were instructed to annotate each edge with an indication of whether they had “high” or “moderate” confidence in the suggested direct causal link. We provide the full instructions given to the experts in Web Appendix 2. The clarity of these instructions was tested in a small pilot study using a third medical doctor with only little research experience.

The two experts first constructed their own models independently without consulting each other. Afterwards, we organized and moderated a meeting where the two expert brought their individual models and together constructed a new consensus model. This consensus model serves as our primary expert model, but we also compare the individual models in Section [4.1].
3.3 Data-driven model construction

The data-driven models were constructed using the temporal Peter-Clark (TPC) algorithm\(^6\). TPC is an extension of the well-known causal discovery algorithm PC, modified to take into account the temporal structure of life course data.

TPC uses information about conditional independencies among variables to infer a TPDAG. Therefore, it requires a statistical test for conditional independence. However, it is mathematically impossible to construct a test of conditional independence between combinations of numeric and binary variables that controls both type I and type II errors simultaneously (no matter how large the sample), if we do not make any assumptions about the distributions of these variables\(^13\). Instead, we use a test of non-association suggested in\(^6\), which tests a necessary (but not sufficient) condition for conditional independence using generalized linear models (GLMs) with spline-expansions to handle non-linearities.

TPC requires choosing a significance level, \(\alpha\), for the tests. No general guidelines exist for the choice of \(\alpha\) for TPC (or PC). Which value of \(\alpha\) performs best depends on graph size and sample size\(^4\), and most likely also on data distribution, choice of test and whether background information is included. We consider two approaches for choosing this significance level:

**TPC search (TPC-S):** Choose the significance level such that the number of edges in the estimated graph is as close as possible to the number of edges in the expert consensus graph.

**TPC pre-specified (TPC-P):** Use a pre-specified significance level of \(\alpha = 0.01\).

TPC-S makes use of the fact that adjacency retention is generally high as \(\alpha\) decreases\(^6\). This means that the difference between the discovered skeletons for e.g. \(\alpha = 0.1\) and \(\alpha = 0.01\) will be that the former graph can be constructed by adding edges to the latter graph. Hence, we can make sure the final graph has a certain number of edges by searching for the relevant value of \(\alpha\).

This approach requires an expert graph, which may not be available in many use cases where TPC could be applied. Therefore, we also consider the TPC-P approach. This resembles the arbitrary choice of significance level often used in statistical testing. Our specific choice of \(\alpha = 0.01\) has also been considered by others\(^14\), but we wish to emphasize that the value is nonetheless somewhat arbitrary.

We use a complete case analysis, that is, we exclude all observations that have missing information. In a sensitivity analysis, we furthermore apply test-wise deletion, as suggested by\(^15\). When using test-wise deletion, observations are excluded test by test rather than globally, and therefore, the statistical power is reduced less.

We investigate the stability of the results by use of bootstrapping. We repeat the TPC-S and TPC-P procedures 100 times on random bootstrap samples, and report the stability of an edge as the number of times it occurs across the bootstrap repetitions.

All computations are conducted in R, primarily using the causalDisco package\(^12\). The code is available in Web Appendix 3.

4 Results

We first describe the expert models in Section 4.1 and the data-driven models in Section 4.2. Afterwards, we compare the two in Section 4.3.

Note that we can only determine whether two models agree on the orientation of a specific edge if that edge is in fact included in both models. Therefore, when comparing two models, we first describe agreement with respect to adjacencies, and then discuss orientation agreement among the shared edges.

Note also that the maximal number of possible edges in a DAG (or CPDAG) with 22 variables is

\[
n_{\text{max}} = \sum_{i=1}^{22-1} i = \frac{22 \cdot 21}{2} = 231.
\]

4.1 Theory-driven models

**Consensus model** Figure 2 presents the expert consensus DAG. This DAG has a total of 30 edges, 7 of which the experts have high confidence in (marked with blue). We find that most proposed direct causal links are either within a time period or between variables from two adjacent time periods. We find only two direct causal links with a larger time lag (i.e., connections that surpass several periods), namely edges between positive attitude towards school in childhood and undergraduate
education in adulthood, and maternal smoking in childhood and total years of smoking in adulthood. The occurrence of time-lagged direct causal effects is often of special interest in life course epidemiology because it may provide evidence in favor of e.g., critical period theories, which propose that some early exposures have effects in specific periods much later in life that cannot be intervened upon and diminished after the critical period has ended (16).

### Inter-expert agreement

The two experts provided rather different individual model suggestions (Web Figures 2 and 3). The expert 1 model has 19 edges, while the expert 2 model has 37 edges. Table 2 compares the adjacencies found in two expert models. We see that the two experts agree on 15 adjacencies, while expert 2 claims 22 adjacencies that expert 1 does not, and expert 1 claims 4 adjacencies that expert 2 does not. Among the 15 adjacencies that the experts agree on, they also agree on the orientation of 13 edges.

Expert 1 marked 6 edges as high confidence, while expert 2 marked 13 edges as high confidence. Five of these edges were shared by both experts, and both experts agreed on the orientation of these edges.

### 4.2 Data-driven models

**TPC-S**

TPC-S was used to construct a model with the same number of edges as the expert consensus model. This was achieved at $\alpha_{\text{search}} = 0.00015625$.

Figure 3 shows the estimated TPDAG. We note that this TPDAG is quite informative, as only 2 out of 30 edges are undirected. Hence it makes claims about the causal direction of the remaining 28 edges.

We find a high degree of stability for the edges included in this model. Using bootstrapping, we find that 26 of the 30 edges in the TPC-S model occur more than 60 % of the time, and 15 of these edges occur more than 80 % of the time (Web Appendix 4, Web Table 2). If we only consider adjacencies, and disregard edge orientation, even higher stability measures are found.

**TPC-P**

We also consider using TPC-P with a pre-specified significance level of $\alpha = 0.01$ (Web Figure 4). All adjacencies found in the TPC-S model are also present in the TPC-P model, but the TPC-P model furthermore has 13 additional edges.

The TPC-P model is also very informative, since only two edges are unoriented. The unoriented edges are the same as for TPC-S. The two procedures agree on the orientation of 26 of their shared 30 edges, while 4 edges are oriented differently between the two graphs. These are edges between low paternal social class and intelligence test score (both in childhood), binge drinking and total years of smoking (both in adulthood), weekly contact with friends and cohabitation status (both in adulthood), and depression and retirement (both in early old age).

We again find a high degree of stability: 32 of the 43 edges in the TPC-S model occur more than 70 % of the time, but an additional 5 edges occur in more than 75 % of the bootstrap repetitions if we disregard the edge orientation (Web Appendix 4, Web Table 3).

### Handling of missing information

Both TPC-S and TPC-P produced identical models to those presented above when using test-wise deletion rather than complete case analysis to handle missing information. However, TPC-S found its target significance level to be somewhat different when using test-wise deletion; we found $\alpha_{\text{twd search}} = 0.0000843$.

### 4.3 Comparison: Theory-driven and data-driven models

We first compare the TPC-S model and the consensus expert model, see Table 3. We find large agreement about absent adjacencies, but less agreement for claimed adjacencies. Only 10 adjacencies are present in both the TPC-S model and the expert consensus model, while both models have 20 additional adjacencies that the other does not. The symmetry is by design, since both models have 30 edges. Among the 10 shared adjacencies, the two models large agree on orientation: 9 of these have the same orientation, while only one does not. This is the edge between weight and length (both at birth), which is undirected in the TPC-S model.

If we only consider edges marked as high confidence in the expert consensus model, we see that TPC-S performs quite well: Out of the 7 edges marked with high confidence in the expert consensus model, 6 are also present in the TPC-S model, and they all have the same orientation in both models. Only the edge from bullied in school to positive attitude towards school (both in childhood) is not...
present in the TPC-S model. The high confidence edges are found to be very stable: 5 of them are found in all bootstrap repetitions, while the last edge is found in 88% of the bootstrap repetitions for TPC-S.

We also note that the TPC-S model has more lagged causal relationships spanning several periods than the expert models, including a direct causal link that skips two periods (maternal smoking in childhood to hospital contact due to heart disease in early old age).

Turning to the TPC-P model, we see that it has found 3 additional edges from the expert consensus model, compared with TPC-S, but that this comes at the cost of 30 suggested edges that are not present in the expert consensus model, which is 10 more than the TPC-S model (Table 4). Moreover, 10 among the 13 adjacencies shared with the expert consensus model also had the same orientation. All 6 high confidence edges were found in all bootstrap repetitions and hence their estimation was highly stable.

4.3.1 Causes of depression
To exemplify what the differences in the two models can imply, we now focus on depression in early old age. Zooming in on this variable, we find quite large differences between the theory-driven and data-driven models. The data-driven TPC-S model identifies only three direct causes of depression (employment status in adulthood, retirement in early old age, and heart disease in early old age), while the expert consensus model proposes five additional direct causes (cohabitation, BMI, contact with friends, binge drinking, and number of children, all in adulthood). Moreover, all of these variables additionally serve as confounders of the effect of heart disease on depression according to the expert consensus model, while this effect is unconfounded in the TPC-S model. Hence, the differences in how the two models describe causal links to depression not only include potential causes, but also have implications for how these causal effects may be estimated, as it is necessary to adjust for confounders in order to obtain unbiased causal effect estimates.

If we move beyond only considering direct causal effects, we find further similarities between the two models. In several instances, the models agree on indirect causes of depression, although they do not agree on the intermediate causal pathways. For example, although the two models disagree on whether binge drinking in adulthood is a direct cause of depression in early old age, they do agree that there is an indirect causal pathway from binge drinking to depression. The same can be said for the effect of BMI.

On the other hand, the two models do not agree on whether cohabitation, contact with friends and number of children in adulthood are indirect causes of depression in early old age. In the TPC-S model, these variables are not indirect causes of depression while they are direct (and sometimes indirect) causes in the expert consensus model. Hence, the two models are in larger agreement about the roles of binge drinking and BMI than the other three variables for which they did not agree about direct effects.

4.3.2 Plausibility of additional edges proposed by the data-driven model
We conducted a post-hoc investigation of whether the extra edges proposed by TPC, but not included in the expert model, may aid causal model building. Table 5 provides an overview of the extra edges, and we have assigned a plausibility classification to each of them to describe whether the proposed direct causal relationship is supported by existing knowledge. Web Appendix 5 presents arguments underlying the classifications. The classifications were conducted post hoc, and hence should be interpreted cautiously. We find that only 3 of the extra edges are classified as low plausibility, while 6 edges are moderately plausible, and 11 are highly plausible.

5 Discussion
For the theory-driven approach, we found that the two experts provided somewhat different individual model suggestions, although we find higher agreement about proposed adjacencies between the two experts than between the TPC-S and expert consensus models. We consider this as evidence in favor of expert models having higher credibility than data-driven estimates. However, the proposed consensus strategy will likely produce an even stronger model, and of course the more experts involved, the higher the credibility. But using a larger number of experts is perhaps not feasible, since causal model construction is very time consuming and requires highly specialized knowledge. Hence, including a few experts with different specializations, as we did in this study, can be a reasonable compromise.
For the data-driven approach, we found that TPC identifies some, but not all of the direct causal links proposed in the expert consensus model. Notably, the TPC-S model correctly identified almost all direct causal links marked with high confidence by the experts. This indicates that TPC-S does indeed capture causal relationships. TPC-S also produced a number of direct causal links not proposed by the experts, and correspondingly missed some direct causal links from the expert model. However, we found that the majority of the edges proposed by TPC-S, but not by the expert model were in fact epidemiologically plausible, while only very few suggested edges are outright implausible. Hence, conducting data-driven model construction does indeed seem to contribute with new insights and meaningful hypotheses that may have been overlooked in the expert model construction process.

We therefore propose that TPC can be a useful supplement to classical theory-driven model construction, and that combining theory-driven and data-driven approaches may lead to stronger causal models that include both well-known and newly discovered causal relationships. We found it useful to use the TPC-S approach specifically, so that TPC obtains the same number of edges as the expert model, because this will allow the data-driven approach to contribute with as much information as the theory-driven approach. TPC-P provided insights into what may happen when we allow the data-driven method to contribute with more information: While TPC-P added 13 additional edges, compared to TPC-S, only 3 of them were also found in the expert consensus model. We consider this a worse trade-off between informativeness and correctness than what TPC-S produced. We thus suggest a pipeline where the initial expert model is first constructed, then a data-driven search is performed targeting the same number of edges, and finally the results of the data-driven search should be critically assessed and be incorporated into a new combined model.

This study is the first to compare theory-driven and data-driven causal model construction for life course studies and, to our knowledge, epidemiology more generally. A few related studies exist. (17) search for causal effects in a database of pneumonia patients, and compare the discovered effects with expert judgments. However, they do not construct full causal models, but instead only consider whether potential instrumental variables can be used to identify causal relationships between pairs of variables. They do not incorporate temporal background information into the search. (18) propose to use the PC algorithm to add additional edges to expert graphs (used in the PC algorithm as background knowledge), thereby allowing the data-driven method to make further suggestions for direct causal links. However, their procedure fails and results in faulty causal conclusions if there are any wrongly oriented edges in the expert model. In contrast, our proposed pipeline would still have the chance to find a conflicting orientation in the data-driven model and thereby draw the experts’ attention towards a potentially problematic edge. (19) construct causal models for coronary heart disease (CHD) using both causal discovery and so-called knowledge engineering. For the latter, they consider regression models from published studies on CHD, and construct causal models where all covariates from these regression models are considered potential direct causes of the outcome, but otherwise unrelated. This may not be an advisable strategy, since covariates are often included in regression models because they are potential confounders, and hence this approach may not describe the intended causal model. (20) also compare human and computer methods for constructing causal models for health sciences. In this study, however, the baseline true models are constructed with computer simulations, and medical students were given data simulated from these models and tasked to construct causal models using the data. This did not prove an easy task, and we suggest that it is hence more useful to have humans construct models from theory and computers infer models from data, as in the current study.

An open question is how much the results presented here depend on sample size. TPC conducts a large number of tests of non-association, and hence sample size can potentially have a notable impact on the statistical decisions made along the way. The PC algorithm is quite sensitive towards sample size and may require large samples to adequately estimate a causal model (4). But due to the use of external information, TPC does not conduct as many tests as PC, and hence it may perform relatively better for small or moderate samples. We consider the high degree of bootstrap stability found in this study to support this conjecture. Further work should be devoted to investigating this systematically.

The choice of test may also influence the performance of TPC. We use a test that tests a necessary, but not sufficient, condition for conditional independence (namely non-association in a GLM setting). If the test is not testing a good proxy for conditional independence, the algorithm is not guaranteed to work well. The fact that we did identify all but one direct causal relationships marked with high confidence in the expert consensus model indicates that we are testing something that can be used to infer conditional independence (and thereby, causal models). We thus believe that studies such as the present may help inform whether the test is appropriate.

Another more direct consequence of the choice of test is that the study was restricted to binary and numeric variables, as the test used here only accommodates these data types. As a result, we have used dichotomized versions of some categorical variables that were originally measured using more
categories (bullying in childhood, positive attitude towards school in childhood and paternal social class at birth and in childhood). Such a dichotomization may introduce bias if the true causal effects operate at the level of the more nuanced categories that were collapsed into two categories. However, the original categories may be less well-defined (i.e., how much bullying a person was subjected to) and hence suffer from poor external validity. All in all, it is not clear whether the restriction towards binary variables is problematic in this study, and we suggest that each application of TPC should critically assess this. If dichotomization is not meaningful in a specific application, TPC can still be used, but the test we have used here needs to be replaced by an alternative that accommodates categorical variables with more than two categories.

We did not find any differences in the estimated models from TPC-S nor TPC-P, depending on whether we used complete cases analysis or test-wise deletion to handle missing information. This may be due to the rather low amount of missing information (13%). There is thus no noticeable gain in statistical power.

Another, more severe, issue related to missing information is that both complete case analysis and test-wise deletion are only generally guaranteed to be unbiased if whether a given observation is missing or not is independent of all both observed and unobserved variables (i.e., missing completely at random using Rubin’s terminology (21)). This is a quite strong assumption which is likely to be violated to some extent in the current study. However, we believe that the rather low amount of missing information makes it unlikely that such bias drives the conclusions. Some work has investigated alternative approaches to handling missing information in the PC algorithm under less restrictive assumptions (22, 15) An interesting line of future research is to extend these analyses to TPC as well.

Moreover, there are still many open questions about the statistical properties of causal discovery algorithms. While their mathematical foundations are well-understood, we still need more knowledge about how finite data and imperfect statistical testing impacts their performance. In particular, we lack means to control type I and II errors, and this has important implications: While we may find the mathematical assumption of faithfulness uncontroversial, we may still be subject to “empirical unfaithfulness” simply due to an under-powered study leading to type II errors in statistical tests of conditional independence. Moreover, we do not yet have a systematic way to describe the statistical uncertainty of the procedures. Using bootstrapping, we have investigated how stably each edge is estimated across resampling, and this provides some insight into the marginal uncertainty of each estimated edge. However, the goal of causal discovery as applied here is not to estimate singular edges, but to provide models for the full data generating mechanism. Hence, an appropriate uncertainty measure should quantify uncertainty over the full model, not edge by edge. We hope that new methodological developments will allow us to supplement analyses such as the current with more adequate uncertainty quantifications in the future.

A more fundamental limitation of the current study, and TPC more generally, is the assumption of causal sufficiency. In many epidemiological datasets, it may not be reasonable to assume no unobserved confounding. In the presence of unobserved confounding, PC (and by extension also TPC) is known to produce models with spurious adjacencies [1]. Hence, absent adjacencies (i.e. non-connected variables) will generally still be correct (of course subject to statistical uncertainty) even in the presence of unobserved confounding. We do believe that this issue is likely to affect the models presented here. Although we have aimed to include as many relevant variables as possible, there will most likely be some unobserved confounders that we have left out, and hence, the TPC-S and TPC-P models may include spurious adjacencies. This could explain the implausible edges discussed above, but these could also result from statistical uncertainty, even if there is no unobserved confounding. In any case, these imperfect models may be the best solution at this time, and corresponds well with current practice in epidemiology, where DAGs are often used, even though their use implicitly assumes no unobserved confounding. In the presence of unobserved confounding, PC (and by extension also TPC) is known to produce models with spurious adjacencies [1]. Hence, absent adjacencies (i.e. non-connected variables) will generally still be correct (of course subject to statistical uncertainty) even in the presence of unobserved confounding. We do believe that this issue is likely to affect the models presented here. Although we have aimed to include as many relevant variables as possible, there will most likely be some unobserved confounders that we have left out, and hence, the TPC-S and TPC-P models may include spurious adjacencies. This could explain the implausible edges discussed above, but these could also result from statistical uncertainty, even if there is no unobserved confounding. In any case, these imperfect models may be the best solution at this time, and corresponds well with current practice in epidemiology, where DAGs are often used, even though their use implicitly assumes no unobserved confounding.

Extending the current work to handle unobserved confounding may be possible, but there are several challenges. First, a DAG is no longer adequate for describing the causal data generating mechanism, and instead experts would need to construct for example acyclic directed mixed graphs (ADMGs). Secondly, TPC cannot be used. Instead, the fast causal inference (FCI) algorithm [1] can be applied, as this algorithm is correct even under non-sufficiency. However, a temporal extension of FCI would have to be developed and implemented in R. (23) have proposed a related extension of FCI, which does include temporal background knowledge, but this algorithm assumes that there is no unobserved confounding between variables from different periods. We do not consider this a realistic assumption in the present context. Third, FCI also learns an equivalence class rather than the full causal model, and when there is unobserved confounding, this equivalence class is described by a partial ancestral graph (PAG). However, PAGs are quite complex and difficult to interpret. Moreover, comparing a PAG to an ADMG is not straightforward, since their edges do not have the same interpretation. Even if the ADMG belongs to the equivalence class the PAG represents, they do not even share
the same adjacencies. Hence, all in all, handling unobserved confounding when constructing data-driven causal life course models requires new methodological developments, and we plan to address this in future research.

References


Figure 1: An example of two directed acyclic graphs (DAGs), (A) and (B), that belong to the same Markov equivalence class, and the completed partially directed acyclic graph (CPDAG) that represents this equivalence class (C). If variables $X_1$ and $X_2$ are assigned to one period, while $X_3$ and $X_4$ are assigned to a different period, the graph in (A) is furthermore the temporal partially directed acyclic graph (TPDAG).

Figure 2: Expert consensus model. Edges colored blue are those that the experts marked high confidence in, while the remaining edges were marked as moderate confidence by the experts. BMI denotes body mass index.

Figure 3: Model obtained by using temporal Peter-Clark algorithm search (TPC-S), that is, TPC with $\alpha$ search ($\alpha_{\text{search}} = 0.00015625$). BMI denotes body mass index.
Table 1: Overview of variables, their periods and the age and year associated with each period.

<table>
<thead>
<tr>
<th>Period</th>
<th>Age</th>
<th>Year</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth</td>
<td>0</td>
<td>1953</td>
<td>Mother married, Low paternal social class, Length, Weight</td>
</tr>
<tr>
<td>Childhood</td>
<td>12</td>
<td>1965</td>
<td>Maternal smoking, Intelligence test score, Positive attitude toward school, Bullied in school, Low paternal social class</td>
</tr>
<tr>
<td>Youth</td>
<td>18-30</td>
<td>1971-1983</td>
<td>BMI, Intelligence test score</td>
</tr>
<tr>
<td>Adulthood</td>
<td>51</td>
<td>2004</td>
<td>Employment status, Binge drinking, BMI, Total years of smoking, Weekly contact with friends, Cohabitation status, Number of children, Undergraduate education</td>
</tr>
<tr>
<td>Early old age</td>
<td>52-64</td>
<td>2005-2017</td>
<td>Depression, Hospital contact due to heart disease, Retired</td>
</tr>
</tbody>
</table>

Abbreviation: BMI, body mass index.

Table 2: Comparison of adjacencies\(^a\) included in the two individual expert models\(^b\).

<table>
<thead>
<tr>
<th>Expert 1</th>
<th>Adjacency</th>
<th>Non-adjacency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjacency</td>
<td>15</td>
<td>22</td>
</tr>
<tr>
<td>Non-adjacency</td>
<td>4</td>
<td>190</td>
</tr>
</tbody>
</table>

Abbreviation: DAG, directed acyclic graph.

\(^a\) For each of the 231 potential adjacencies in the DAG, we classify whether both experts included the adjacency \((n = 15, 13\) of which also agree on orientation), neither expert included the adjacency \((n = 190)\), only expert 1 included the adjacency \((n = 4)\) or only expert 2 included the adjacency \((n = 22)\).

\(^b\) Note that the two individual expert models do not contain the same number of edges.
Table 3: Comparison of adjacencies\(^a\) found in the TPC-S model and the expert consensus model\(^b\).

<table>
<thead>
<tr>
<th></th>
<th>Expert consensus</th>
<th>TPC-S</th>
<th>Adjacency</th>
<th>Non-adjacency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjacency</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>181</td>
</tr>
<tr>
<td>Non-adjacency</td>
<td>20</td>
<td>20</td>
<td>181</td>
<td>181</td>
</tr>
</tbody>
</table>

Abbreviations: DAG, directed acyclic graph. TPC-S, temporal Peter-Clark algorithm search.

\(^a\) For each of the 231 potential adjacencies in the DAG, we classify whether both approaches included the adjacency \((n = 10, 9 \text{ of which also agree on orientation})\), neither approach included the adjacency \((n = 181)\), only the expert consensus model included the adjacency \((n = 20)\) or only the TPC-S model included the adjacency \((n = 20)\).

\(^b\) Note that the two approaches result in the same total number of adjacencies by design, and hence the symmetry in the two disagreement categories is not incidental.

Table 4: Comparison of adjacencies\(^a\) found in the TPC-P model\(^b\) and the expert consensus model.

<table>
<thead>
<tr>
<th></th>
<th>Expert consensus</th>
<th>TPC-P</th>
<th>Adjacency</th>
<th>Non-adjacency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjacency</td>
<td>13</td>
<td>13</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Non-adjacency</td>
<td>17</td>
<td>17</td>
<td>171</td>
<td>171</td>
</tr>
</tbody>
</table>

Abbreviations: DAG, directed acyclic graph. TPC-P, temporal Peter-Clark algorithm pre-specified.

\(^a\) For each of the 231 potential adjacencies in the DAG, we classify whether both approaches included the adjacency \((n = 13, 10 \text{ of which also agree on orientation})\), neither approach included the adjacency \((n = 171)\), only the expert consensus model included the adjacency \((n = 17)\) or only the TPC-P model included the adjacency \((n = 30)\).

\(^b\) Using significance level \(\alpha = 0.01\).
Table 5: Overview of edges proposed by the TPC search model that are not present in the expert consensus model.

<table>
<thead>
<tr>
<th>From-variable</th>
<th>Edge type</th>
<th>To-variable</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother married (B)</td>
<td>$\rightarrow$</td>
<td>Low paternal social class (B)</td>
<td>Low</td>
</tr>
<tr>
<td>Mother married (B)</td>
<td>$\rightarrow$</td>
<td>Intelligence test score (C)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Low paternal social class (B)</td>
<td>$\rightarrow$</td>
<td>Intelligence test score (C)</td>
<td>High</td>
</tr>
<tr>
<td>Low paternal social class (B)</td>
<td>$\rightarrow$</td>
<td>Intelligence test score (Y)</td>
<td>High</td>
</tr>
<tr>
<td>Weight (B)</td>
<td>$\rightarrow$</td>
<td>BMI (Y)</td>
<td>High</td>
</tr>
<tr>
<td>Maternal smoking (C)</td>
<td>$\rightarrow$</td>
<td>Hospital contact due to heart disease (E)</td>
<td>High</td>
</tr>
<tr>
<td>Intelligence test score (C)</td>
<td>$\rightarrow$</td>
<td>Low paternal social class (C)</td>
<td>Low</td>
</tr>
<tr>
<td>Positive attitude towards school (C)</td>
<td>$\rightarrow$</td>
<td>Intelligence test score (Y)</td>
<td>Low</td>
</tr>
<tr>
<td>Low paternal social class (C)</td>
<td>$\rightarrow$</td>
<td>Undergraduate education (A)</td>
<td>High</td>
</tr>
<tr>
<td>Intelligence test score (Y)</td>
<td>$\rightarrow$</td>
<td>Total years of smoking (A)</td>
<td>High</td>
</tr>
<tr>
<td>Intelligence test score (Y)</td>
<td>$\rightarrow$</td>
<td>Retirement (E)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Employment status (A)</td>
<td>$\rightarrow$</td>
<td>Total years of smoking (A)</td>
<td>High</td>
</tr>
<tr>
<td>Employment status (A)</td>
<td>$\rightarrow$</td>
<td>Cohabitation status (A)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Binge drinking (A)</td>
<td>$\rightarrow$</td>
<td>Total years of smoking (A)</td>
<td>High</td>
</tr>
<tr>
<td>Binge drinking (A)</td>
<td>$\rightarrow$</td>
<td>Cohabitation status (A)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Total years of smoking (A)</td>
<td>$\rightarrow$</td>
<td>Retirement (E)</td>
<td>High</td>
</tr>
<tr>
<td>Weekly contact with friends (A)</td>
<td>$\rightarrow$</td>
<td>Cohabitation status (A)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Number of children (A)</td>
<td>$\rightarrow$</td>
<td>Cohabitation status (A)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Undergraduate education (A)</td>
<td>$\rightarrow$</td>
<td>BMI (A)</td>
<td>High</td>
</tr>
<tr>
<td>Undergraduate education (A)</td>
<td>$\rightarrow$</td>
<td>Retirement (E)</td>
<td>High</td>
</tr>
</tbody>
</table>


*Variable names are annotated with their period.

*We use three categories of plausibility. "Low" indicates that no evidence in favor of the relationship exist, and that we have no suggestions for convincing new hypotheses that explain the relationship. "Moderate" indicates that some evidence in favor of the relationship exist, or that we have meaningful suggestions for new hypotheses that explain the relationship. "High" indicates that strong evidence in favor of the relationship exists. Motivations for each plausibility classification are provided in Web Appendix 5."
Figure 1

A) $X_1 \rightarrow X_3 \rightarrow X_4$
B) $X_1 \rightarrow X_3 \rightarrow X_4$
C) $X_1 \rightarrow X_3 \rightarrow X_4 \rightarrow X_2$
Figure 2