Author's response to reviews

Title: Combining Directed Acyclic Graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology

Authors:

David W Evans (davidwevans1@gmail.com)
Basile Chaix (chaix@u707.jussieu.fr)
Thierry Lobbedez (lobbedez-t@wanadoo.fr)
Christian Verger (c.verger@wanadoo.fr)
Antoine Flahault (Antoine.Flahault@ehesp.fr)

Version: 3 Date: 28 August 2012

Author's response to reviews: see over
Dear Dr Taljaard,

Thank you very much for the review of our manuscript, Combining Directed Acyclic Graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology, and for the opportunity to submit a revised version.

We very much appreciated the detailed reading by the reviewers and their insightful and helpful comments. The revised manuscript is, in our opinion, considerably improved thanks to this exchange. We have uploaded a version with changes to the text shown in bold.

You will see that the revised manuscript is longer than the original version owing to the addition of new paragraphs to address the reviewers’ points. Whilst we understand that this length still meets journal policy, we may be able to move some text into an additional file if preferred.

We appreciate your considering our revised manuscript for publication in BMC Medical Research Methodology and look forward to hearing from you in due course.

Yours sincerely,
Dr David Evans, for the authors.
Authors’ response to reviewers

Reviewer: Ian Shrier

GENERAL COMMENTS
I want to thank the authors for writing this manuscript. The general idea has been around for a few years and it is nice to see someone attempt to tackle the associated challenges. The text is generally well written but I do have some concerns that I think the authors should address.

MAJOR COMPULSORY REVISIONS
1. Defining meaningful change: This is the crux of the problem. The authors suggest the 10% rule, and maybe to use an absolute difference. It seems to me that more thought is required for this step. In this context, it is not just what is “meaningful” that is important, but also what is expected. I would expect that as measurement becomes less precise (i.e. random error increases as opposed to systematic measurement error), the expected variation in effect estimates with the addition of different variables (i.e. the plus one side of the figures) might change. Further, when comparing the results of different sufficient sets (as suggested by the authors for bias amplification, and a method that could be more widely used, see comment below), I would expect large random error would greatly affect the plus-one graph. Perhaps the authors feel that the addition of “measurement error” nodes would solve this, and if so, they should discuss it. For me, that would mean every variable would always have to be associated with a node for measurement error, and the DAG would become impractical. So, if a node for measurement error would solve the issue, the authors should provide some guidance as to how much random error would need to occur before it is necessary to include a measurement error node in the DAG. In addition, the authors should also comment on factors that might affect what a meaningful change might be.

We thank the reviewer for his thoughtful feedback on this core issue. We have considerably expanded the manuscript (pp.11-12) to address the issue of expected variation, following these comments as well as those from Dr Weng and Dr Fleischer. We agree that the addition of multiple measurement error nodes would be impractical. We preferred to develop the idea, originally mentioned as a potential extension in the Discussion, of bootstrapping the add-one and minus-one patterns to estimate variability. We propose calculating the estimated proportion of bootstrap samples which fall outside the meaningful threshold (i.e. which would have led to a revision of the DAG), reporting these proportions for transparency, and using them in a sensitivity analysis when reviewing the prior DAGs. We have updated the empirical example accordingly. As we acknowledge that the proposed cut-off of 50% of bootstrap samples falling outside of the threshold in the sensitivity analyses is arbitrary, we flag this in the Discussion as an area requiring further work. We consider exploring the performance of different thresholds by simulation studies out of scope for this first article.

Thanks to the reviewer’s comment, we have also revised the section on “Defining a
meaningful change” (p.11) to clarify the thinking behind proposing an absolute change (compared with the 10% relative change in common practice) and to address in more detail how a researcher should choose a meaningful threshold based on considerations of relevance to clinical or public-health decisions. This choice, of course, remains arbitrary but we feel that the proposed approach has the merit of clearly communicating the threshold and the reasons for its choice to other researchers.

MINOR ESSENTIAL REVISIONS

2. One step that I think the authors have under-emphasized is that there is more than one minimally sufficient set within any DAG. Therefore, in addition to the change in estimate by adding or subtracting variables, one can also use the same principles to compare different minimal sufficient sets – there should not be any change in estimate between these sets if the DAG is correct. The authors do mention this strategy when discussing bias amplification, the same strategy should be used to test mis-specification of the DAG for any reason – there is nothing unique about bias amplification in this regard.

We appreciated this suggestion and have revised the section presenting the proposed steps to include a comparison of the effect estimate adjusted on all minimally sufficient adjustment sets (p.10). As this will be particularly helpful in distinguishing between different plausible DAGs, we recommend it for comparing different candidate DAGs after initial selection based on the add-one and minus-one patterns. We have also updated the Results section to highlight the two minimally sufficient adjustment sets in the DAG in Figure 1 (p.15).

3. The authors state that the objective of the method is to develop the best working DAG rather than the “true” DAG. However, I think there are times when several different DAGs might be consistent with the data, or perhaps, equally inconsistent with the data. In these cases, one does not know which DAG is correct. I believe that in these cases, authors should present all the different DAGs as plausible, and this will help focus future research on the most important areas of uncertainty. As more knowledge becomes available, the various DAGs would be tested again and inappropriate DAGs would be dropped. A sentence or two discussing advantages of their approach in this regard could improve the manuscript further.

This is a very welcome observation and we have revised the manuscript accordingly. In the steps on pp.9-10, we now explicitly recommend checking all the plausible prior DAGs for consistency with the observed change-in-estimate patterns (steps 8 and 9) and presenting all consistent DAGs identified as final DAGs (step 12). We have also added two sentences to the Discussion (p.23) to highlight how presenting all plausible DAGs consistent with the observed patterns will communicate uncertainties to other research teams and so help them support or discount the DAGs using different datasets.

4. Page 14: define PKD the first time it is used

This has been corrected (p.19).
5. Figure 9: I am not clear on why “type of peritoneal dialysis” exists. First, if it is important to exist, there must be an arrow from Peritoneal/Hemodialysis box. But I would think it is just one form of peritoneal dialysis, and therefore the Peritoneal/Hemodialsis is simply a variable with many levels instead of being dichotomous. I would also think there would be an arrow from Peritoneal/Hemodialysis to death since people die without it, and since peritoneal dialysis has a limited time frame before one needs to switch to hemodialysis.

Type of peritoneal dialysis refers to continuous ambulatory peritoneal dialysis (CAPD) and automated peritoneal dialysis (APD). We have clarified this in the text (p.19) and in the footnotes to the figures, and have also added an arrow from PD/HD to type of PD as kindly signalled by the reviewer. We preferred to keep the type of PD node in the DAG. This variable is frequently adjusted on in observational PD research as a prior confounder, whereas we consider it to be downstream of the exposure for most research questions. The DAG helps show this. We also prefer to maintain the separation of the PD/HD and Type of PD nodes since this clearly shows PD/HD as a selection variable for the Registry. We have already conditioned on PD/HD during data collection; but we can choose whether to condition on CAPD/APD during data analysis. We did not include an arrow joining PD/HD and death as mortality does not appear to differ between dialysis modalities. We do not think that there are (many) patients needing dialysis who do not receive it in the French healthcare system, so that this variable does not have a “No dialysis” class.

DISCRETIONARY REVISIONS

6. Page 13: Bias amplification is not limited to instrumental variables. It will exist even for “confounders” if the variable is more strongly related to the exposure than the outcome (under the additive model – under the multiplicative model it is more complex). If the authors can fix this sentence without making the text more confusing, it would be helpful.

We thank the reviewer, Dr Fleischer, and Dr Ogburn for flagging the need to clarify instrument-like variables. The fact that confounders can be instrument-like is indeed crucial to the bias amplification discussion. We have now updated the text on p.17 to define instrument-like, to explain how confounders can be instrument-like, and to contrast instrument-like variables with standard instrumental variables. We have also clarified why $C_2$ and $C_3$ in Figure 8 have different instrument strengths (pp.17-18).

7. Page 15: The sentence “The minimally sufficient adjustment set...” should be re-worded.
Although this is actually the only minimally sufficient set in this DAG, most DAGs have more than one minimally sufficient set. Therefore, the authors should say “There is only one minimally sufficient adjustment set from the prior DAG and it is simply....”

This has been clarified (p.20).

8. Figures: The figures should stand-alone. The break line between add-on and minus-one needs to be explained. For example, it should say that “minus 1” means you remove
covariates from the minimally sufficient set, etc. I think it would be better for the minimal adjustment set to be listed in the figure rather than inside the figure legend. Also, when DAGs are being redrawn with changes, it would be helpful if there were some indication in the figure of what the change is so one doesn’t have to flip back and forth to figure it out. For example, the authors could use light gray arrows to show arrows that are deleted and dotted arrows (or thicker arrows) to show arrows that are added. I’m not sure what would work best, but the current style is sometimes difficult to follow.

We have included the adjustment set at the top of each Figure and have revised the figure legends to explain the meaning of the add-one and minus-one boxes. We agree with the reviewer that a way of following changes to the DAGs would help the reader but our attempts to do this by modifying the arrows were very unsatisfactory. We therefore decided to include the Powerpoint version of the DAGs as an additional file, as the changes are easily identified by flicking back and forth between figures. To encourage readers to actually do this, we also direct them to the additional file in the beginning of the Results section (p.14).

9. The authors may want to read Pearl and Bareinboim technical report R-372A on Selection Diagrams. Although these methods are originally described for translating information about assumptions related to generalizability, they can also be used to encode uncertainty about arrow presence/absence in a DAG. I don’t think the authors should include this concept in the current manuscript, but a reference to it might be helpful to help move the field forward.

We thank the reviewer for pointing us to this interesting report. We have revised the text to acknowledge selection nodes as a possible extension to the approach in the final paragraph of the Discussion (p.27).

10. Web Appendix: The authors say that Type of Nephropathy is not a causal contrast because the variable cannot be manipulated. The authors should acknowledge that not everyone agrees that this restriction should be included in the definition of cause, or else the idea will continue to be propagated without critical thought. I understand that some authors do require this as part of their definition, but others do not (Pearl, Principal Stratification article in Int J Biostats 2012), and I fall into the latter camp. For example, an earthquake cannot be manipulated but causes damage and death, and I can’t believe that some people would say it is not the cause. Of course, they may say that they can create the effects of an earthquake with a bomb, and it is these effects that cause the death, but that is just a question of proximal and distal causes. Now, in the example of PKD, we don’t even have to worry about this debate because strictly speaking, nephropathy can be manipulated. One can certainly cause several different types of nephropathy (e.g. ischemic, myoglobin-associated damage, etc). In addition, there are different genetic causes of PKD, and these conditions have associated animal models, i.e. we can and do cause PKD in animals (doi: 10.1111/j.1365-2796.2006.01743.x). With a little thought, almost all of the causes that are considered non-manipulated can indeed be manipulated, or one can envision a time when technology will allow us to manipulate them. Does that mean that something is not a cause today, but will become a cause when humans develop a particular technology? That would
seem to be a strange definition of “cause”. A compromise would be to consider the restriction mentioned to highlight “causes of interest”, or “causes that can be directly studied today”, rather than “causes” per se. I would therefore suggest the authors either soften the language to acknowledge different opinions on this topic, use the reference provided to say that PKD can be manipulated in animals, or just delete this sentence altogether.

We are grateful to the reviewer for this comment and, in fact, share his view of the importance (or not) of being physically able to manipulate a variable in causal comparisons. As suggested, we chose to delete the sentences covering this point. In fact, we have been challenged over manipulability of PKD (and other variables) in other settings and so appreciated the reviewer’s insights into the manipulability of some nephropathies, including PKD in animals, and the strangeness of considering a non-manipulable exposure as a non-cause now but a cause in the future, once technology to manipulate it is available – points we will be sure to raise in future discussions!
Reviewer: Nancy Fleischer

Overall, this is a nice paper that provides a valuable extension of the DAGs literature. My suggestions are to help clarify various points, particularly in the Methods and Results.

Major Compulsory Revisions
1. The paper would be improved by a thorough explanation in the Methods concerning consistency and inconsistency with the DAG (as outlined in steps 7 and 8). It would be helpful if the authors discussed, in a methodical manner, how to assess the differences in estimates with the “add-one” and “minus-one” methods, and what that means. For instance, what do different directions and magnitudes of change mean in these approaches? Help the reader assess, in a stepwise manner, what they are seeing in Figure 3 (e.g.)

We thank the reviewer for this suggestion. We have added a paragraph immediately following the steps (p.10) to state the expected add-one and minus-one patterns and to give an example of the expected patterns on different adjustment sets in Figure 1. We have also rephrased the discussion of Figure 3 in the example (pp.14-15) to state the implied patterns before discussing how the observed patterns differ from these. These paragraphs deal with the magnitude of the change (falling outside of the defined threshold); dealing with the direction of the change is flagged as a possible extension, using signed DAGs, at the end of the Discussion (pp.27-28).

2. In the Results, it would be helpful if the authors are more explicit with their first example (Confounding, mediation, collision in Figures 1-3). Although the figure legend for Figure 3 explains the solid horizontal line, etc, it would be helpful to have this in the text. Also, this is a good place to walk the reader through the comments on the consistent/not consistent results (mentioned in my previous comment).

We have substantially revised these paragraphs (pp.14-15). We now first present the starting and alternative prior DAGs, followed by the implied add-one and minus-one patterns, before discussing whether the implied patterns are consistent with the observed patterns. We have also added an explanation of the dotted line in Figure 3 to the text, as helpfully suggested. We hope these changes make this part easier to read.

3. The section on Bias amplification is a bit confusing. It would help to better explain why C2 & C3 have different “instrument strengths,” and how that amplifies the bias (paragraph 2 of section).

Following this comment and similar feedback from Dr Shrier and Dr Ogburn, we have revised the text on p.17 to clarify the meaning of “instrument-like” and to explain how confounders can have instrument-like properties. We also explain why C2 and C3 in Figure 8 have different instrument strengths, per the reviewer’s suggestion. We feel that a detailed explanation of why bias amplification occurs is beyond the scope of this article and so refer the interested reader to recent articles addressing the topic.
4. It would be nice to have more discussion on the implications if the empirical assessments of the DAG relationships do not match the theoretical DAG, which was likely based on previous relationships determined from data. What if this is due to important differences in the underlying populations?

This was a helpful comment. In response, we have added a paragraph to the section on preparing the prior DAG (pp.7-8) to address the possibility of differences in the populations from which prior knowledge is derived and the population generating the dataset to be analyzed. The use of selection nodes (see Dr Shrier’s comment on this) may be an interesting future addition to help address this.

Minor Essential Revisions

1. In the Abstract, reword the (iii) point under results to use clearer language more consistent with text of paper (i.e., “add-one” and “minus-one” with definitions). It may help to define the minimally sufficient set under point (ii) as “S”, and refer to that in point (iii).

We have revised these points in the Abstract as suggested (p.2). The points now use the same language as the steps in the Methods.

2. In the second paragraph of the Methods, the authors state that “Conditioning on a variable reverses its status.” This language is not clear. What is the “status” the authors are referring to?

By “status”, we were referring to whether a variable is open or closed on a path, i.e. conditioning on a collider opens the collider whilst conditioning on a non-collider closes the non-collider. We have, however, deleted the paragraph with this sentence following the feedback from Dr Weng on the relevance of this section for the reader (please see response to Dr Weng below).

3. Please explain Step 5 in the Methods: Examine the change of estimate between the models graphically. Perhaps a brief example would help. Is this referring to what the authors did in Figure 3, for example?

We have changed the wording for step 5 in the Methods (p.9) based on the reviewer’s suggestion. We preferred not to add an example in the steps in the Methods but now refer back to step 5 when presenting Figure 3 in the Results (p.14).

4. What is the role of statistical inference in assessing the magnitudes of change in the “add-one” and “minus-one” scenarios?

In response to this question and to the comments from Dr Shrier and Dr Weng, we have added a section on estimating the variability in the add-one and minus-one patterns and on using this variability in a sensitivity analysis when revising the prior DAGs (pp.11-12). (Please also see...
the response to Dr Shrier above.)

5. Specify the “uncertain confounding pathways” mentioned in paragraph 3 of the Methods. Also, why is Figure 1 referenced here for an uncertain confounding pathway, if it is the unknown best? It seems obvious that, if Figure 1 is the unknown best, that it would have consistent add-one, minus-one patterns.

We have revised the article to clarify that the “uncertain pathways” in the original version refer to alternative variable relationships considered plausible during the preparation of the prior DAGs (p.7). The revised version now refers to alternative prior DAGs. We have also modified the text to state more clearly that Figure 1 is both the unknown best working DAG and one of the alternative prior DAGs identified during preparation (p.14). Indeed, Figure 1 does have consistent patterns by design.

Discretionary Revisions
1. It may be beyond the scope of the paper, but it would be nice to also have a section on effect modification and the implications of the add-one, minus-one assessments in that context.

We agree that effect modification is an important question in covariable selection using the proposed method (and, indeed, any method) but we also feel that it is beyond the scope of the present paper. We hope to address it in future work and have listed it as an extension at the end of the Discussion (p.28).

2. It would be helpful, in the “Drawing up a prior DAG” section of the Methods, to number the steps explained in the second sentence (currently separated by a semi-colon).

We have broken these points into a numbered list (p.7).
Reviewer: Hsin-Yi Weng

In this manuscript, the authors proposed a novel approach of using the observations in empirical data to assist in examining and revising the prior DAG. This brings up an immediate logical concern about whether it is appropriate to use samples to test the assumptions about the underlying population where the samples are from. After reviewing the manuscript I am still not fully convinced by the authors that their approach is appropriate. This is mainly because the authors did not critically discuss how the implications of sampling variation may affect their approach. For example, a non-confounding variable (in the population) may cause a change in estimates (or vice versa) in a sample by chance alone. The empirical example presented in this manuscript, which included only 17 cases in the PKD group, is a good example to illustrate this concern. The concern might be exacerbated by increased number of covariates to be considered or by weak confounders. I do not agree that empirical data from a single sample can be used to make decision on complex causal assumptions that may involve a set of confounding variables and different biases.

We thank the reviewer for this important and helpful comment, also raised by Dr Shrier and Dr Fleischer. Based on this feedback, we have substantially added to the manuscript to address variability in the add-one and minus-one patterns and on considering variability in reviewing the DAGs. The revised articles addresses the sources of variation in the change in estimate (sampling error, model instability owing to small sample size or too many covariates, etc.); suggests a method of estimating variability by bootstrap, and proposes a sensitivity analysis to consider this variability when revising the DAG (pp.11-12). Please see the response to Dr Shrier’s first comment for more details.

We share the reviewer’s unease about using empirical data from a single sample to make decisions about causal relations. However, epidemiologist make decisions about adjustment variables (i.e. causal assumptions), frequently based on single datasets, all the time! We hope that by incorporating DAGs into this process, we epidemiologists will communicate assumptions more clearly to colleagues and will be able to show how we revise these assumptions during covariable selection. An important point here is presenting all plausible DAGs consistent with the observed patterns, as flagged by the reviewer below and by Dr Shrier. We have now clarified that researchers should present all such plausible DAGs and have also explained in the Discussion (p.23) that the revision of the DAGs should ultimately be an iterative process between research teams, with other groups looking at the DAGs with their own data. We hope that these revisions address the reviewer’s question.

My second major concern is about their procedure of starting with only a single preferred working DAG. This is conflict (or confusing to a reader like me) with their suggestion of “The researcher should also list the main uncertainties when preparing the DAG”. I will suggest that instead of preparing only a single prior DAG, the researchers should prepare a group of prior DAGs each presenting different prior assumptions. Then they start with the simplest DAG (i.e., the one contains the least number of covariates in the minimally sufficient variable set or contains the covariates that can be best measured) or the most preferred (if any) DAG.
The three main reasons to do so are: 1) DAGs are the best means to explicitly present prior assumptions, 2) It is not uncommon that two or more DAGs are equally good or possible, and 3) the results from empirical data should be used to evaluate only within this group of prior DAGs, thus it may help prevent the problem of researchers trying to revise DAG to fit the observed data. Point number two is particularly troubling me. Based on the proposed approach, users will report the results based on two DAGs, the working (prior) DAG and the revised DAG (based on data). Why limited to only these two sets of causal assumptions if all prior assumptions are equally plausible?

We appreciated this suggestion, which was similar to a point raised by Dr. Shrier, and have revised the manuscript to take this into account. Our intention in the original manuscript was to recommend preparing a set of plausible prior DAGs showing uncertainties, and then to review the DAG from within this set as described in the example. However, the reviewer’s comment highlighted that this was not clear. We have therefore revised the text to explicitly recommend drawing up a group of prior plausible DAGs showing uncertainties; starting the analysis with a preferred prior DAG (in general the most plausible DAG); and presenting all plausible prior DAGs consistent with the observed patterns as final DAGs with accompanying adjusted estimates. We have also strengthened the language around the risks of revising a DAG a posteriori to fit the data (p.13) based on the reviewer’s feedback.

My third main concern is the title of the manuscript, which is “Combining Directed Acyclic Graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology”. I feel it is not reflecting their objectives accurately as the main objective of the proposed approach was to use data to assist in reviewing the prior DAG.

We are not comfortable with this assessment of the objective of the proposed approach but certainly appreciate the comment as it showed the need to clarify the objective in the text. We have therefore revised the final paragraph of the Background (p.5) to restate the objectives of the approach. The other revisions to the manuscript, notably the clarifications around presenting adjusted estimates backed up by a group of plausible prior and final DAGs, should also clarify this point.

The authors ambitiously tried to convey some very complex concepts to the audiences in this manuscript. But who are their target audiences? Those who are expert in DAGs or with only little knowledge? The current manuscript does not suit both groups of audiences well. This comment is especially for the subsection of DAGs and minimally sufficient adjustment variable sets in Methods. The authors may leave the whole subsection out by targeting only those whom have sufficient knowledge of DAGs or they should provide more details so the readers can follow the contents better. For example, by adding “A<-B->C represents a backdoor path from A to C through B” and using this example to explain what children and parents are would make the manuscript more readable. In addition, for their example of A->B<-C, explain that B is the collider. Given the complexity of this manuscript, I will suggest the authors to target audiences who already have sufficient knowledge of DAGs and focus on their approaches, however. Similar comment also applied to the appendix, particularly the
appended table, which I found to be difficult to follow. It might help if the authors could provide more detailed description of each column title in the table.

This comment was very welcome. As suggested, we decided to delete this subsection as we felt it was not realistic to provide a comprehensive introduction to DAGs whilst keeping the article to a readable length. We also decided it was not necessary, given the number of excellent introductions available in the literature. We now simply state that we assume the reader has a basic understanding of DAGs and have added additional background references of different technical levels (p.6). Based on the reviewer’s feedback, we have also clarified the headings and steps in the table in the appendix.

The following are additional minor comments:
1. Abstract; Background: remove quotations for expert-knowledge.

This change has been made (p.2).

2. Abstract; Background: change “change-in-coefficient” to “change-in-estimate”.

This correction has been made (p.2).

3. Please use the same numbering: such as i. in page 6 or (i) in abstract.
We have revised numbering to be of the i. format in the abstract (p.2).

4. I don’t understand the wordings describing the two conditions for a sufficient variable set to adjust for confounding in pages 6 and 7. Please revise the wording.

We preferred to keep the wording for these two conditions as they closely follow the formal graphical criteria proposed by Greenland et al. in their seminal article on causal diagrams in epidemiological research. However, we have added sentences to explain the second point and to provide a simple summary of what the criteria mean (p.6).

5. Page 7, Using minimally sufficient adjustment sets to compare a DAG with data: check the wording of the first sentence.

We have rewritten this sentence (p.8).

6. Page 8: are not mediators (or ancestors or descendants)...or colliders (or descendants): what are the ancestors and descendants referred to?

We have revised the text to clarify this (p.8). It now reads “are not mediators (or ancestors or descendants of mediators)…or colliders (or descendants of colliders)…”.

7. Page 9, Defining a “meaningful” change: 10% change is not an obvious “meaningful” threshold at all it is (for most) only a conventional value people use. Why adding quotations
to meaningful?

We have removed the quotation marks around “meaningful”. Our intention in putting the quotation marks was indeed to highlight that it was an arbitrary value. We have also expanded the discussion around how to choose a meaningful threshold (p.11).

8. Page 12: remove the parentheses for “Following [39], we define C∗ as the measured variable, and UC as representing all factors affecting measurement of C.” Make the c in UC subscript.

This change has been made (p.16).

9. Page 14, Empirical example, spell PKD out

This has been done (p.19).

10. Page 14: we also show a 10% change in the estimate of what? RD?

We have revised the wording to clarify that the 10% change is in the RD (p.19).
**Reviewer: Elizabeth Ogburn**

This paper provides an informal way to refine a hypothesized DAG to better represent data. I don’t feel that the authors make a sufficiently strong case for the usefulness of their proposal, given that more formal and rigorous methods exist to learn and test DAG structure. I believe that the authors’ intention is to improve upon the very simple, informal change-in-coefficients tests already in wide use for variable selection, but if this is the case then I feel that the authors should make a case for this project as compared to more sophisticated methods.

**Major Compulsory Revisions**

1. The introduction should acknowledge the existence of the DAG structure learning literature (e.g. the PC algorithm, Spirtes et al. 2000), even if the authors deem this material too technical to report in detail.

We were grateful for this comment and similar comments from Dr Shpitser which flagged the need to address DAG-discovery approaches in more detail in the manuscript. We have revised the Background to acknowledge DAG-discovery algorithms (p.4), as suggested, and also discuss DAG-discovery algorithms in more detail in the Discussion (p.26).

2. There are easy-to-implement and easy-to-understand tests of conditional independence that can verify d-separation criteria directly. Why not propose using these instead of change-in-coefficients tests to detect the presence or absence of arrows on a DAG?

Please see the response to Dr Shpitser’s feedback on the same issue below.

3. The material introducing DAGs is vague. I think that the definitions of concepts like of blocked and unblocked paths, compatibility, and faithfulness can be made precise and rigorous without taking up much more space.

We appreciated this comment on the introductory material, on which Dr Weng also gave feedback. As mentioned in the response to Dr Weng above, we decided to remove the section introducing DAGs as we agreed with the suggestion of targeting readers with knowledge of DAGs. The section now simply refers readers to other articles and book chapters which introduce DAGs in epidemiology.

4. It is not clear what is meant by partial conditioning in the last sentence of the second paragraph on p. 6: conditioning on the descendant of a collider will unblock a path through that collider; conditioning on the descendant of a non-collider confounder may not even partially control for the confounder. Similar remarks apply to the section on measurement error beginning on p. 12: Suppose C2 were comprised of two components (two distinct biological mechanisms), one of which affected A while the other determined C2*. Then adjusting for C2* would be extraneous. Furthermore, it is known that adjusting for mismeasured confounders can increase bias (e.g. Brenner 1993, Ogburn & VanderWeele 2012).
We thank the reviewer, Dr Fleischer, and Dr Shpitser for all pointing out the lack of clarity in this sentence. As noted, we have removed the introductory section on DAGs, which included this sentence, in the revised manuscript. We are not sure to have correctly understood the reviewer’s second point about distinct biological mechanisms in C2. If these mechanisms are distinct (independent?) and have separate effects on A and C2*, should they be shown as separate nodes? We were very grateful for pointing us to the articles about the conditions when adjustment on mismeasured confounders can increase confounding: we have changed the text to acknowledge this and to include these references (p. 16).

5. The limitations of the work are not clearly stated. I think that some limitations are:
   a. The change-in-coefficients approach to variable selection is ad hoc and therefore using it to test DAG structure is also ad hoc.
   b. The proposal in this article is to use change-in-coefficients analysis to test hypotheses about DAG structure, but these are not well-defined statistical tests with any of the desirable properties of consistency, validity, etc.
   c. Any test of the presence or absence of a particular arrow on a DAG is sensitive to the particular configuration of the assumed DAG. For example, suppose that Figure 2 was hypothesized to represent data for which there was in truth an arrow C3 # C5. Then a test of whether controlling for C5 changes the estimated effect of A on Y provides no evidence for the arrow from C5 to Y. But, having left C3 # C5 off the DAG, an investigator would conclude that a change in the effect estimate is evidence of C5 # Y.

Based on this helpful feedback, we have expanded the discussion of the limitations of the method (p. 27). This paragraph includes all of the above points. We would like to thank the reviewer for the example she gives in point c. for Figure 2 and have, in fact, used this in a new paragraph in the Results covering the presentation of more than one final DAG (p. 18). We recognize that the DAG configuration is sensitive to starting assumptions; the revisions to the manuscript clarify how the proposed approach handles this. In particular, we have clarified that researchers should prepare several prior DAGs and present several final DAGs if needed, thereby communicating uncertainties to other research teams. Please also see our response to Dr Shrier’s feedback on this point.

6. The discussion of “instrument strengths” at the bottom of p. 13 confused me, because C2 and C3 are not instruments in the standard sense.

We thank the reviewer, Dr Shrier, and Dr Fleischer for picking up on the need to explain instrument-like variables in more detail. As noted above, we have added text (p. 17) to clarify when a variable is instrument-like, to explain how confounders can be instrument-like even though they are not instrumental variables in the standard sense, and to explain why C2 and C3 in Figure 8 have different instrument strengths.

Minor Essential Revisions
1. Insofar as the DAGs discussed in this paper reflect expert knowledge, they are causal
DAGs. Therefore the last sentence of the first paragraph of the METHODS section is misleading.

Based on this comment and the comments from Dr Shrier and Dr Shpitser, we have revised this paragraph (p.6) and the language throughout the manuscript to address causal DAGs

Discretionary Revisions
1. I wonder if the discussion of c-equivalence on p. 17 belongs in the introduction, since it provides the theoretical foundation of the paper.

We agree that not mentioning the theoretical foundations of the paper in the Background was an oversight. We have now added a sentence on c-equivalence and recent work on collapsibility of different estimates over different DAG structures to this section (p.5). We prefer to keep the more detailed discussion of c-equivalence in the Discussion to avoid making the Background too technical for some readers.

2. I had a hard time following point (ii) at the top of p. 7.

As noted in the response to Dr Weng, we have added a sentence to clarify the meaning of point ii.. We preferred to keep the wording as it follows the criteria as presented by Greenland et al. in the reference.
**Reviewer: Ilya Shpitser**

**Major Compulsory Revisions**

The authors need to clearly distinguish causal and statistical concepts. "Confounding bias", and "causal effect", are causal concepts, "collapsibility", "measure of association" are statistical concepts. More on this in the detailed comments.

Please see response in the supplementary comments below.

The author's method is a variant of a causal discovery method. The authors need to explain why existing DAG causal discovery methods are not suitable for their problem (many of them can incorporate prior knowledge).

Please see response in the supplementary comments below.

**Minor Essential Revisions**

"Conditioning on a variable reverses its status. Conditioning on a child or a parent of a non-collider variable at least partially conditions on the non-collider; conditioning on the child of a collider at least partially conditions on the collider."

I think I see what the authors are trying to say here, but I think the current phrasing is misleading. I think what the authors meant to say is that conditioning on a variable reverses its status with respect to blocking collider/non-collider triples. Moreover, conditioning on a descendant of a collider also opens the collider, while conditioning on a parent or a child of a non-collider does not close the non-collider (for this reason I found the last last sentence in the quote confusing).

Please see response in the supplementary comments below.

**Supplementary comments**

I have two main issues with this draft. First, I am having a hard time distinguish ing between statistical and causal issues in this manuscript. Comments relating to this will be in section (1). Second, I am confused by the authors willing to revise the graph using change in estimate information based on selecting different confounders. Comments relating to this will be in section (2). Minor issues will be in section (3).

**Section (1).**

----------

On page 1, the authors say:

"Based on the relationships laid out in a DAG, researchers can predict how a collapsible association estimator (e.g. risk ratio or risk difference) for an association of interest should change when adjusted on different variable sets. Predicted and observed patterns can then be compared to detect inconsistencies and so guide adjustment-variable selection. Based on..."
this, the proposed approach involves (i) drawing up a plausible background-knowledge DAG; (ii) identifying a minimally sufficient variable set to control for bias affecting the association of interest.

This is confusing. Do the authors mean bias due to confounding? If so, why would this bias affect "the association of interest." Generally, confounding bias affects causal effects, not associative relationships.

We did indeed mean bias due to confounding and thank the reviewer for picking up on the lack of clarity in the language used. Based on this comment and the feedback from Dr Shrier and Dr Ogburn on causal issues, we have changed the Methods (p.6) to state that the article addresses causal DAGs and have revised the rest of the manuscript to refer to effect estimates rather than associations. We have therefore kept references to confounding.

Further on page 5:
Further approach does not pretend to identify, "true" DAGs (if such objects even exist) but rather tries to produce a well-performing working DAG for a given research question, given the data at hand. To be pragmatic, the approach focuses on an exposure-outcome relationship of interest and uses regression models and the change-in-estimate procedure familiar to epidemiologists. It also emphasizes the added value of DAGs in communicating assumptions and uncertainties in an analysis.

Again, I am not sure it makes sense to talk about confounding unless the "true" DAG is known (or at least a small set of structurally similar DAGs where a single variable may serve as a confounder in all of them).

This is an interesting and challenging comment. The short response is that we deleted references to the “truth” of a DAG when clarifying the objectives of the approach (p.5), following Dr Weng’s feedback on this above. The longer response is that we are not sure if it makes sense to talk about confounding unless a true DAG is known either. We are comfortable referring to confounding in the manuscript, even if the proposed approach does not hope to produce a single true DAG. Epidemiologists routinely talk about confounding when using other variable-selection methods without knowing the true underlying DAG. This does not make it right, of course, but it does show that there is a shared language which refers to confounding without needing knowledge of a true DAG. Further, in the proposed approach, even if the true DAG is unknown, using DAGs which are close to but not the same as a true DAG will generally help reduce confounding. This fits with epidemiologists’ way of referring to confounding by degrees (“minimize confounding”, etc.). We hope that this will often be the case since the approach, as discussed below, fits best in areas like clinical epidemiology where background knowledge places reasonable limits on the set of plausible prior DAGs. In addition, as raised by the reviewer, we do expect the approach to produce several DAGs with the same single variable or set of variables serving as confounders.

Further on page 6:
"DAGs are a graphical description of the joint probability distribution of a set of random variables, showing marginal and conditional (in)dependencies between variables [3, 14, 15–16]. Although standard [1, 17], arrows do not have to have a causal meaning [18] and so we refer to conditional associations in this article; however, arrow direction is probably most easily handled by considering causal directions."

DAGs can either represent statistical or causal models (or sometimes both!). The authors seem to refer to statistical DAGs at the beginning of p. 6, but then talk about selecting variables to adjust for confounding. Confounding is a causal concept and only makes sense in causal DAGs. I think the authors are confusing causal and statistical DAGs.

The relationship is as follows: causal models represented by a DAG G induce a joint distribution which lies in the statistical DAG model associated with G (that is the distribution Markov factorizes according to G, or alternatively d-separation is sound for the independences in the distribution). However, the reverse is of course not true -- a joint distribution which lies in the statistical DAG model associated with G may well not have come from a causal model represented by G, but from a causal model associated with some other graph G1, or even from no causal model at all! In this kind of distribution, the notion of 'adjusting for confounding' using G as a guide is incoherent.

We valued this feedback. As noted above, we have revised this paragraph to state that the article addresses causal DAGs and have updated the language throughout the manuscript accordingly. Our main reason for introducing the notion of statistical association in the original draft was to cover non-manipulable exposures but Dr Shrier’s comments helpfully clarified this point.

Section (2).
------------

The authors describe their method based on changes in the association estimate as the adjustment set changes on pages 8-9.

I am confused by their procedure. It seems to me that if there is uncertainty about the DAG, then the natural thing to do is to employ standard methods for causal discovery from data (that is learning the structure of the graph using the faithfulness assumption). There are many algorithms in the literature for this task, with lots of nice theoretical properties. Examples include:

The PC algorithm (Spirtes, Glymour, and Scheines)
The LiNGAM algorithm (Shimizu et al.)
The GES algorithm (Chickering)
etc.

It may be that the author's algorithm is superior for their particular task (perhaps if the graph is mostly known, with a few uncertain edges). However, the authors do not seem to provide an
evaluation of their method vs competing methods currently in the literature. It also seems to me that the threshold the authors employ for revisiting the DAG structure is ad hoc (vs something like a statistical test of conditional independence employed by the PC algorithm, or a principled scoring procedure like the GES algorithm).

We thank the reviewer and Dr Ogburn for highlighting the need to address DAG-discovery algorithms further in the manuscript. As the reviewer indicated, we do indeed consider the proposed approach to be most appropriate for data with reasonable background knowledge about variable relationships, as in clinical epidemiology. We had stated this in the original text (p.26); but, to emphasize this important point, we now repeat it in the Abstract (p.3), the paragraph discussing DAG-discovery algorithms (pp.26-27) and in the Conclusion (p.29). We have also expanded the discussion of the ad hoc nature of the threshold in the limitations section and have expanded the rationale proposed for choosing a threshold (p.11) based on the reviewer’s, Dr Ogburn’s, and Dr Shrier’s comments.

We feel that a head-to-head comparison of DAG-discovery algorithms is out of scope for this first article. For information, however, we note that our attempts to use these algorithms in the past have had mixed results. For example, running the PC algorithm on the data for the empirical example gave an empty adjustment set without background knowledge and only Age with background knowledge (temporal ordering of variables). The FCI algorithm produced similar results. These did not fit well with clinical knowledge about variable relationships and did not help understand the large change in estimate when adjusting on Type of assistance. However, we can see value at this point in time from using these algorithms as a complementary part of preparing prior DAGs and have revised the manuscript to suggest this (p.26).

The authors say in the conclusion on p. 20:
"A researcher wishing to explore the full DAG could apply a DAG-discovery algorithm (e.g. TETRAD [16]), but such algorithmic approaches which exclude background knowledge are controversial [51] and have not yet crossed over into applied epidemiologic research."

There is quite a bit of work on incorporating background knowledge into causal discovery.

We thank the reviewer for picking this up. The sentence has been corrected (p.26).

Section (3).
--------------

p. 6:
"Conditioning on a variable reverses its status. Conditioning on a child or a parent of a non-collider variable at least partially conditions on the non-collider; conditioning on the child of a collider at least partially conditions on the collider."

I think I see what the authors are trying to say here, but I think the current phrasing is misleading. I think what the authors meant to say is that conditioning on a variable reverses
its status with respect to blocking collider/non-collider triples. Moreover, conditioning on a
descendant of a collider also opens the collider, while conditioning on a parent or a child of a
non-collider does not close the non-collider (for this reason I found the last last sentence in
the quote confusing).

We are grateful to the reviewer, Dr Fleischer and Dr Ogburn for pointing out the problems
with these sentences. We did indeed mean a change in the open/blocking status of
colliders/non-colliders as summarized by the reviewer. However, as outlined in the response
to Dr Weng, we have deleted the introductory material on DAGs (including these sentences),
so as to focus on readers who already have knowledge of DAGs.