Responses to CASAC Questions on the PM PA from Consultant Dr. Constantin Aliferis

Questions from Dr. Cox

Question 2*: On the same topic of clear definitions, does the discussion of the BenMAP-CE beta coefficients in the PA and underlying documentation (described as typically representing "the percent change in a given adverse health impact per unit of pollution") unambiguously specify which of the following concepts the coefficients represent?*

a. Beta estimates the percent change in the conditional expected observed value of the health impact associated with a unit change in the observed value of the pollution variable. (This might be called the regression interpretation of the beta values.)

b. Beta estimates the percent change in the mean value of the health impact variable caused by a manipulation or intervention that changes the value of the pollution variable by 1 unit (e.g., increasing concentrations at all times and locations by 1 microgram per cubic meter above what they otherwise would have been), while holding the values of all other variables fixed at the values they had before the intervention. (As concrete examples, the values of lagged daily high and low temperatures, humidity, and co-morbidities such as asthma in the weeks before a death would not be affected by the hypothetical (or counterfactual) change in the pollution variable for a beta coefficient that addresses the health impact on mortality risk of a change in pollution.) (This might be called the natural direct effect interpretation of beta.)

c. Beta estimates the percent change in the mean value of the health impact caused by a manipulation or intervention that changes the value of the pollution variable by 1 unit (e.g., increasing concentrations at all times and locations by 1 microgram per cubic meter above what they otherwise would have been), while allowing the values all other variables to change in response to the changes in pollution. (As concrete examples, the joint distribution of the values of lagged daily high and low temperatures, humidity, and co-morbidities such as asthma could be changed by conditioning on counterfactual changes in the pollution variable.) (This might be called the total direct effect interpretation of beta.)

d. Beta estimates the percent change in the mean value of the health impact caused by a manipulation or intervention that changes the value of the pollution variable by 1 unit (e.g., increasing concentrations at all times and locations by 1 microgram per cubic meter above what they otherwise would have been), while holding the values of all other variables (e.g., co-exposures, co-morbidities, potential confounders or modifiers such as weather variables) fixed at the values they would be expected to have after the intervention.

e. Beta means something different from any of the above (e.g., a controlled direct effect).

Do all of the beta values in Table C-1 refer to the same one of these concepts, or might they refer to different ones (or is the answer not clear)?

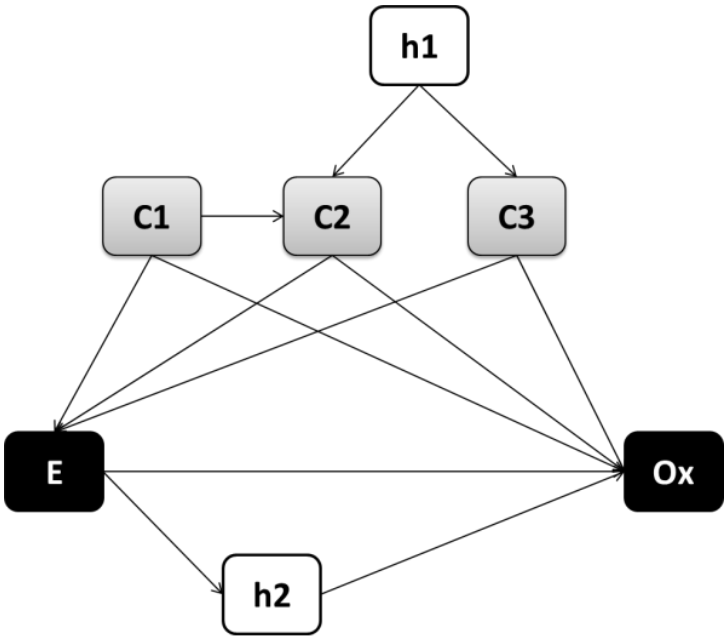
Response:

Providing causal interpretations to results of regression and other correlative studies in the absence of specific causal assumptions, constraints or background information, is tricky if not outright infeasible. Here how I read such results in the scientific and technical literature:

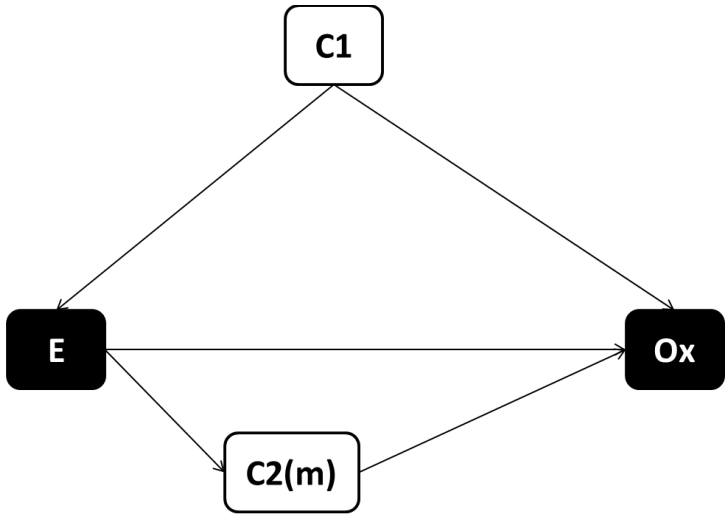
a: Yes, provided that we qualify by stating that all other covariate's values are held constant and that this beta value is strictly model-specific (i.e., lacks any invariant/inherent biomedical meaning in the sense that changing the model will change the beta).

b, c, d, e: Unclear - can be either yes or no depending on assumptions about the nature of the covariates inside and outside the model.

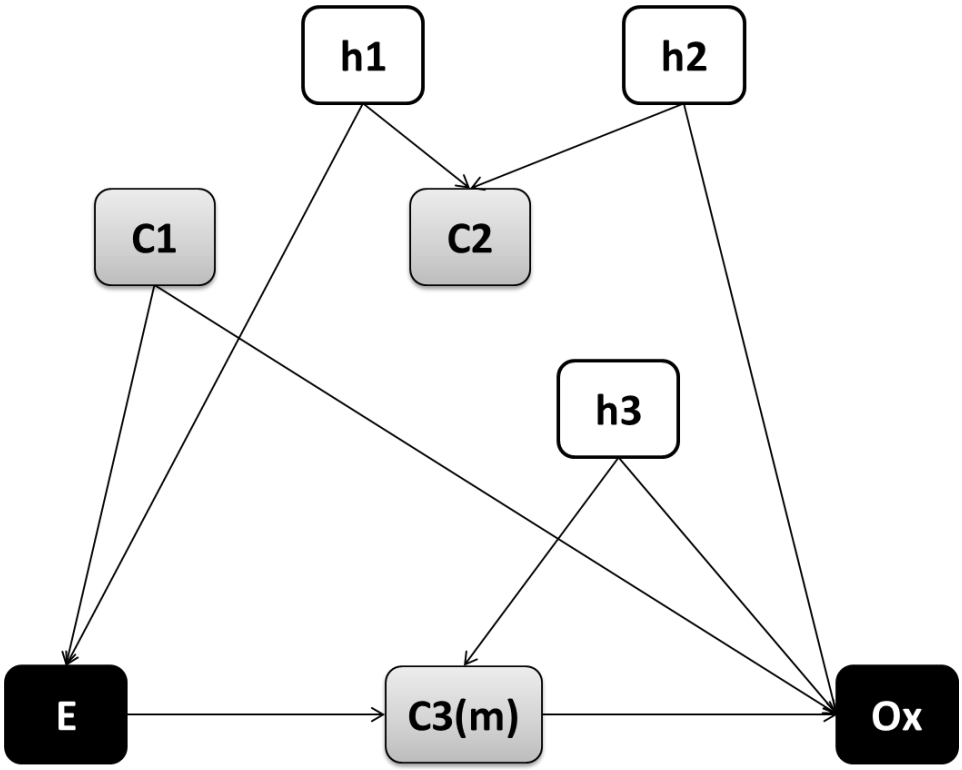
Understanding the last statement is key in clearing, in my opinion, much of the causal confusion in this and similar settings. Let's consider the following three examples (statements made hold under standard assumptions of causal modeling and for the vast majority of (i.e., Faithfull) distributions):



Example 1: this figure depicts an example true causal structure underlying some data. E= exposure, Ox=outcome, "C" variables are measured covariates, "h" variables are latent (unmeasured). If the regression model incorporates C1, C2, C3 as covariates, then the beta of E corresponds to the total causal effect of manipulating E (ie via both paths E→Ox and E→h2→ Ox) on Ox. This interpretation requires that no other unmeasured confounder exists that is not blocked by a measured covariate in the model. h2 is taken to change as a result of manipulating E. h1, C1, C2, C3 have their natural joint distribution.



Example 2: As before but I have marked C2 with "m" to point that it is mediating part of the causal influence of E on Ox. Here the beta of E corresponds to the direct causal effect of manipulating E (ie via path E→Ox only). This interpretation requires that no unmeasured confounder exists that is not blocked by a measured covariate in the model. C2 is taken to not change as a result of manipulating E. C1 has its natural joint distribution.



Example 3: Here the regression model incorporates C1, C2, C3(m) as covariates and the beta of E does not have a proper causal meaning of any kind. Specifically the indirect effect of E via path e→C3(m)→ Ox is blocked by C3(m) in the model but the (confounding) path E→C3(m)←h3→Ox opens; the confounding path E←C1→Ox is blocked; using C2 as covariate opens a second biasing path via latent variables h1, h2 (E←h1→C2←h2→Ox). Hence no unbiased estimate of causal effect of E on Ox is provided, either direct or indirect. This analysis assumes that no other unmeasured confounder exists

that will mitigate the biases (such an event would be extremely unlikely anyway). C3(m) is taken to not change as a result of manipulating E. h1, h2, h3, C1, have their natural joint distribution.

<u>These examples demonstrate a number of important facts of general significance:</u>

- For the regression analysis to produce a valid causal estimate (of any kind: total, direct, indirect, or other conceivable variants not usually discussed in the epidemiologic literature), a number of causal constraints (or assumptions) involving the measured and unmeasured variables must hold.

- These assumptions involve the following questions:
    - Are common causes of E and Ox sufficient for blocking all confounding, measured?
    - Are paths involving common causes of E, Ox blocked or opened (or both) by measured variables?
    - Are mediators of E to Px measured?
    - Are common causes of E and mediators or mediators and Ox measured?

- If and only if we can guarantee the correct answers to these questions we can produce reliable causal effect estimates for E.

- On the basis of this information we can also decide which measured variables can and should be covariates. For example in the graph 2, we can elect to omit C2(m) from the model in order to estimate total causal effect of E on Ox, or to include it, in order to obtain the direct causal effect only.

- It does not matter whether we use Poisson regression, Cox regression, Logistic Regression, GLM etc in the sense that they all depend on the right choice of covariates (and are easily chosen on the basis of distribution and research design) *.

- As a matter of fact, what modern causal inference methods do is that they create a causal graph that correctly captures the underlying causality. By using this graph the modeler can then decide which variables to use as covariates in order to obtain valid causal effect estimates.

- In addition, the causal discovery algorithms reveal which effects can be estimated without experiments and which cannot. For example, inference of the structure E←h→Ox and E→Ox informs us that we cannot estimate the causal effect of E and it should not be attempted because it cannot be done without experiments and/or additional assumptions.

- *Caution however: models that do not use conditioning (eg many machine learning and statistical models that use regularization, will systematically produce wrong causal estimates even with the right set of controlling covariates).

Question 4*: The PA and BenMAP-CE documentation repeatedly refers to relative risks, odds ratios, attributable risks, and regression coefficients as bases for quantifying causal effects of changes in air pollution on changes in population health responses. Some epidemiology methodologists and experts in causal analysis have distinguished between association and causation (and between "seeing" and "doing") and have warned that relative risks, odds ratios, attributable risks, regression coefficients, and related concepts (e.g., population attributable fractions, probabilities of causation, etiologic fractions)*

address associations but not causation (including causal effects of interventions), because measures of statistical association do not address how changing one variable would change another (e.g., Pearl J, 2009 Causal inference in statistics: An overview. Statistics Surveys 3: 96-146; Greenland and Robins 2000, Epidemiology, justice, and the probability of causation; Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol. 2002 Apr;31(2):422-9.)

Do the beta coefficients in the PA overcome these methodological objections to using relative risks, regression coefficients, and related measures of association to predict (or simulate) effects of interventions? If so, how were they overcome? If not, does this imply that the simulations in the PA are not necessarily reliable or valid predictors of the real-world effects on public health of reducing PM2.5? Why or why not?

Response:

The previous analysis (question #2) applies to answer this question. In brief:

- Valid causal interpretation of the association measures requires that all the right covariates and only those have been used in the model.
- The right covariates can be found in a number of ways, some of which are stronger, and some weaker.
- Stronger (less prone to error) approaches include, experiments and/or consistent causal discovery algorithms.
- Weaker (more prone to error) approaches include intuition, ad hoc synthesis of literature, application of "causal criteria" (Hill, Koch, Surgeon General etc), using non-causal predictive modeling to identify covariates, using clustering, etc.

Question 6*: More generally, do the Δy/Δx values calculated by BenMAP-CE have valid interpretations as causal impacts on y of interventions that change x? (If x represents daily ice cream consumption in a population and y represents daily cases of heat stroke, would the slope of an estimated regression line or C-R line relating them, β or Δy/Δx, necessarily provide valid predictions or simulations for how an intervention that changes ice cream consumption by 1 unit would change daily cases of heat stroke? Assuming the answer is no, how is the methodology of beta coefficients in Appendix C of the PA essentially different from this example?)*

Response:

The predictions would be causally valid, for a total causal effect for instance, if and only if a set of covariates was included in the model that would eliminate all confounding (by blocking all relevant paths) between ice cream consumption and heat stroke due to common causes of the two.

Essentially then the appraisal work done by the causal assessment methodology amounts, at least implicitly, to ensuring that this condition is likely enforced. It would be in my opinion much better if stronger methods for enforcing such constraints that would ensure causal interpretation, were put in practice. Having said as much, I would not venture to dismiss the coefficients as uninformative or entirely unreliable, however (please see my response to Dr Frampton below for more details).

***Question 16**: How can techniques of formal causal modeling and analysis best be applied to improve the clarity of definitions and communication and scientific soundness of simulations, inferences, causal interpretations, generalizations, and policy-relevant conclusions in the PA? Please comment on whether any aspects of the following (or other) causal model formalisms can substantially improve the clarity and scientific soundness of the analyses and simulations in the PA: causal graph and DAG methods, conditional independence tests, intervention and interrupted time series analyses, other quasi-experimental methods, Wiener-Granger causality and transfer entropy, causal dynamic Bayesian networks (DBNs), other information-theoretic and graph methods, Simon-Iwasaki causal ordering , non-parametric structural equations models, mediation analysis.*

Response:

The two most important steps toward progress, in my assessment, are:

A. Apply valid (theoretically consistent under well-defined and broad assumptions) and scalable causal structure discovery methods to infer the underlying qualitative causality. Then use the results to drive the choice of confounding or mediator controlling covariates in quantitative models of choice.

B. Use Big Data designs where as many relevant variables are induced *at the same time* in the analysis. It is very hard to capture the causal interactions if the variables are not measured together and included in the analysis and is also very hard to combine partial causal results from separate, small-scope studies.

Questions from Dr. Frampton

***Question 3**: Dr. Cox, Chair of CASAC, has introduced concepts of causality determination that differ from the framework established by the EPA in prior ISAs, used in the current PM ISA, and reiterated in section 3.2 of the current PA. This was a topic of discussion during CASAC's recent review of the PM ISA. Please opine on the adequacy of the causality analysis framework currently used by the EPA, and whether and how the concepts espoused by Dr. Cox should, or should not, be incorporated into the NAAQS causality framework. Also please comment on the implications, of any changes in the causality framework that you would recommend, for the analyses and conclusions in this current PA. For background, see the following documents:*

1) Preamble to the Integrated Science Assessments, November 2015, pages 18 to 25 (https://yosemite.epa.gov/sab/sabproduct.nsf/78476291901FF5AE8525835F00634158/$File/ISA_PREAMBLE_FINAL2015.pdf);

2) Follow-up Questions for John Vandenberg from Dr. Cox, letter of 12/17/2018 (https://yosemite.epa.gov/sab/sabproduct.nsf/B28289529495F929852583660057A3EE/$File/Follow-up+questions+for+John+Vandenberg-rev.pdf);

3) Responses from John Vandenberg, letter of 2/20/2019 (https://yosemite.epa.gov/sab/sabproduct.nsf/B48131F413362439852583A7005FDFDA/$File/JVandenberg+response+to+TCox+ltr+of+121718.pdf);

4) CASAC letter to the Administrator, April 11, 2019 - Dr. Cox comments, pages A8 to A27 (https://yosemite.epa.gov/sab/sabproduct.nsf/264cb1227d55e02c85257402007446a4/6CBCBBC3025E13B4852583D90047B352/$File/EPA-CASAC-19-002+.pdf).

Response:

As a conceptual framework underlying my answer please review my responses to Dr. Cox since they are foundational to your question as well (I will not repeat for parsimony).

The causality analysis framework currently used by the EPA is, in my assessment, a reasonable framework given the limitations of causal discovery methodologies available until recently, the fact that it is very hard to conduct complex manipulations (policy) involving the environment, and that cause-effect horizons are lengthy.

Operationally its defining features are the following:

(a) Controlling for confounding using some form of regression.
(b) Choice of plausible confounders on the basis of domain theory and/or synthesis of prior studies.
(c) Application of causality criteria similar to the ones proposed by Hill, Koch, the Surgeon General and others.
(d) Creating a hierarchy of evidentiary support where results are placed.

The weaknesses of this framework (as also indicated in my responses to Dr Cox) are:

(a) Errors in the choice of confounders translate to errors in causal effect estimates. This is <u>the primary weaknesses.</u>
(b) Small-scale (narrow variables scope) studies produce partial results with large variance on effects as a consequence of Simpson's paradox.
(c) Because assumptions are not stated explicitly it is hard to state/understand/share/differentiate effects of mediators, total, indirect, direct, or combinatorial effects.
(d) The causality criteria, are in general highly heuristic and it is trivial to show that in an infinity of causal situations they do not hold. A very prominent book on formal causal discovery refers to such criteria as "indefensible" (from a technical perspective, to be clear) and conducts a thorough (and thoroughly convincing) analysis of this claim (Spirtes et al. Causation, Prediction and Search, 2000).
(e) Moreover, it is not known currently what is the heuristic value of applying such criteria in this domain (and others). For example, when these criteria are applied, what is the probability that qualitative correct inferences are made about direct and indirect causal edges? What is the probability that quantitative causal effect estimates are within epsilon from the true value (for some acceptable epsilon)? What are correct ways and incorrect ones to apply the criteria? How tolerant are they in misapplication situations? And so on.
(f) Because the evidentiary pyramid, a heuristic by itself, is built upon several component methods that are themselves heuristic (as explained), the EPA may be facing a "house of cards" situation where errors in several stages of the process accumulate uncontrollably.

Does these criticisms imply that:

(i) EPA should immediately abandon this framework?
(ii) Prior results using the framework are invalid and policies depending on these results must be reversed or abandoned?
(iii) Every method that has some uncertainties and unknowns is unscientific and useless?

In my assessment, the answers are emphatically "no" to all three questions. The main reasons are as follows:

- First, we know that imperfect causal inference methods often produce useful results. Characteristic example is the identification of smoking as causing cancer and CVD (one of the all-time triumphs of epidemiology and public health policy).
- Second, in data science several mitigation factors that are often in play allow imperfect methods to perform well (classic example: bias-variance error decomposition allows highly unlikely and simplistic models outperform in small samples more realistic but complex models).
- Third, recent empirical evidence suggests that correlative studies have good overall performance. Illustrative massive study to that effect:
  Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials.Cochrane Database of Systematic Reviews 2014, Issue 4. Art. No.: MR000034.DOI: 10.1002/14651858.MR000034.pub2.
- Fourth, from common sense we know that the majority of our practical, functional knowledge about the world comes neither from experiments, nor from formal designed studies and analysis. Yet, it is highly useful and accepted by all (e.g., we all know that crossing the street when the traffic light is red is a bad idea; no experiments have been conducted and no cohorts were created and analyzed to that issue however).

Balance is needed: the above considerations do not mean that we should perpetuate an imperfect system of causal inference when new methods and tools have recently become available that give improved options for better understanding of environmental risk factors. As I mentioned above, recent methods for large scale causal inference are now available and when combined with big data designs can yield extraordinarily powerful results.

Such techniques have transformed for example cancer genomics and oncology and are beginning to do the same across several health sciences/disciplines. Indicatively in my group alone we have used such methods for discovery of causal knowledge, predictive modeling, treatment personalization and many other applications in many medical problems: predicting risk of death from community acquired pneumonia, predicting risk of neonatal sepsis at the ICU, determining which patients with ovarian cancer will benefit from bevacizumab, cancer diagnosis and outcome prediction (several types), factors determining regression of atherosclerosis, factors determining progression of osteoarthritis, genetic factors for rheumatoid arthritis causation and prediction, psoriasis diagnosis using microbiomic sequencing signatures, diagnosis of pre-clinical acute viral infections, differential diagnosis of stroke types and stroke-like syndromes, modeling of physician decision making and guideline compliance, predicting lab measurements in a hospital setting, predicting and understanding childhood PTSD, predicting suicidal thoughts and behaviors, modeling temporal brain connectivity, studying the mediating role of sRNA for longevity, as well as in several studies of mining bibliographies, text, citations, and the health WWW. Many others have contributed a large body of work with significant and highly novel results in numerous diseases and discovery settings.

Questions from Dr. Packham

Question 3: *Can you describe the research efforts necessary to provide the required information? To what extent has the needed research already been done, or started? For example, are there crucial experiments or research initiatives that could clarify the shape of the PM2.5-chronic inflammation causal dose-response relationships at relevant exposure concentrations? Are there specific data analyses (e.g., testing for confounding by weather variables over more days prior to mortality) that could clarify the causal interpretation of epidemiological associations relied on in the draft PA to simulate effects of interventions?*

Response:

A critical opportunity exists in the analysis of additional data forms in rich contexts (i.e., high-dimensional or "Big Data") that should be pursued using the latest, scalable and consistent causal discovery algorithms. I provide 3 illustrative examples of types of studies that have potential to make a positive difference:

a. Analysis of full medical record data that are now routinely mapped to common (and thus interoperable) data models and geo-located across top-tier institutions in the US; overlay and link geo-location to environmental data. This design can both test and generate hypotheses.
b. Analysis of total causal effects (and information content) of environment versus non-environmental factors for various outcomes on a large scale. For example studying pan-omic mediators and co-determinants, medications, occupational data, etc.
c. Systematically study the effectiveness of newer and/vs older causal methods by validating them against real-life effects of interventions against their predicted responses.