THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Public Interest Comment[1] on

The Environmental Protection Agency's Proposed Rule:

National Ambient Air Quality Standards for Ozone

Docket ID No. EPA–HQ–OAR–2008–0699;

RIN: 2060–AP38

March 17, 2015

Louis Anthony (Tony) Cox, Jr.[2]

The George Washington University Regulatory Studies Center

The George Washington University Regulatory Studies Center works to improve regulatory policy through research, education, and outreach. As part of its mission, the Center conducts careful and independent analyses to assess rulemaking proposals from the perspective of the public interest. This comment on the Environmental Protection Agency's proposed rule revising National Ambient Air Quality Standards (NAAQS) for ozone does not represent the views of any particular affected party or special interest, but is designed to evaluate the effect of EPA's proposal on its statutory mandate to "protect public health."

Introduction

The Clean Air Act directs the EPA Administrator to set "primary," or health-based, NAAQS at levels that are "requisite to protect the public health … allowing an

[1] This comment reflects the views of the author, and does not represent an official position of the GW Regulatory Studies Center or the George Washington University. The Center's policy on research integrity is available at http://research.columbian.gwu.edu/regulatorystudies/research/integrity.

[2] Tony Cox is President of Cox Associates, LLC, and a visiting scholar with the GW Regulatory Studies Center.

adequate margin of safety,"[3] based on "air quality criteria [that] shall accurately reflect the latest scientific knowledge useful in indicating the kind and extent of all identifiable effects on public health or welfare which may be expected from the presence of such pollutant in the ambient air, in varying quantities." §108(a)(2) It further requires the Administrator to set "secondary" (welfare-based) standards based on these criteria at a level "requisite to protect the public welfare from any known or anticipated adverse effects."[4] (This comment focuses exclusively on the primary standard.)

Amendments to the Clean Air Act in 1977 (P.L. 95-95) required the Administrator to conduct a "thorough review of the criteria…and promulgate such new standards as may be appropriate," at least every five years. In setting primary standards, EPA may not consider the costs of achieving the standard.[5]

In this notice, EPA proposes to determine that the current O_3 primary standard of 0.075 ppm "is not requisite to protect public health within an adequate margin of safety"; thus, "it should be revised to provide increased public health protection."[6] EPA proposes to revise the O_3 standard to "within the range of 0.065 ppm to 0.070 ppm" in order to increase protection of public health including for populations deemed "at risk" such as children, older persons, and those with health problems.

EPA's reasoning and recommendations are presented in several key documents, including the following.

- Proposed Rule: National Ambient Air Quality Standards for Ozone, *79 Federal Register 75233 -75411*
- Final Report: Health Risk and Exposure Assessment for Ozone, *August 2014*
- Final Report: Integrated Science Assessment of Ozone and Related Photochemical Oxidants, *2013* (hereafter "ISA")

[3] The Clean Air Act, 42 U.S.C. § 7408 (b)(1) Available at: http://www.ehso.com/ehso.php?URL=http%3A%2F%2Fwww.epa.gov/oar/caa/caa113.txt

[4] The Clean Air Act, 42 U.S.C. § 7408 (b)(2) Available at: http://www.ehso.com/ehso.php?URL=http%3A%2F%2Fwww.epa.gov/oar/caa/caa113.txt

[5] *Whitman v. American Trucking Associations, Inc.*, 531 U.S. 457 (2001) 99-1426.175 F.3d 1027 and 195 F.3d 4, affirmed in part, reversed in part, and remanded.

[6] 79 FR 75236

Executive Summary

EPA's proposed determination that existing ozone NAAQS are not requisite to protect public health with an adequate margin of safety is not justified by the evidence it presents. Nor do EPA's predictions that further reductions in ozone standards will cause future public health benefits follow from sound and reliable scientific methods of causal analysis and prediction. The U.S. EPA document entitled *Health Risk and Exposure Assessment for Ozone, Final (July, 2014): Executive Summary* states in its first paragraph that "The health effects evaluated in this HREA are based on the findings of the O_3 ISA (U.S. EPA, 2013) that short term O_3 exposures are causally related to respiratory effects, and likely causally related to cardiovascular effects, and that long term O3 exposures are likely causally related to respiratory effects." The introduction also states that "The results of the HREA are developed to inform the O_3 Policy Assessment (PA) in considering the adequacy of the existing O_3 standards, and potential risk reductions associated with potential alternative levels of the standard."

Thus, EPA positions causality and its implications for potential risk reductions as leading considerations in its proposed decision to reduce existing standards for ozone. The specific meaning of causality to which EPA refers is that it would allow readers of the policy assessment to consider "potential risk reductions associated with potential alternative levels of the standard." We assume that EPA means "caused by" rather than "associated with" here, and that they mean that their assessment and causality determination will allow interested policy makers to understand how different reductions in the ozone standard would (probably) affect future risks to human health and other endpoints of interest. This crucial claim appears to be mistaken and misleading. EPA has presented no validated causal modeling that would enable correct prediction (and uncertainty characterization) of the effects (if any) that future changes in O_3 levels might cause in future changes in health effects. Rather, EPA uses measures and models of historical associations between O_3 levels and adverse health effects as if they were known to be entirely causal. This does not provide a sound or reliable basis for risk assessment (Appendix A; Dominici et al., 2014).

EPA follows a well-developed but unreliable quantitative risk assessment (QRA) process, critically discussed in the next section and in Appendix A, that interprets statistical associations as being causal based on weight-of-evidence (WoE) considerations. As discussed in detail later, WoE considerations are not logically or practically adequate for this purpose (e.g., Goodman et al., 2013; Morabia, 2013; Rhomberg et al., 2013; Appendix A). Table 1 of EPA's ozone ISA presents these

considerations. They incorporate several formal logical fallacies, such the *ex post ergo propter hoc fallacy* that "Evidence of a temporal sequence between the introduction of an agent, and appearance of the effect, constitutes another argument in favor of causality." There is no theoretical or practical reason to suppose that EPA's causal judgments based on these WoE considerations are factually correct. Experts, like other people, typically have high confidence in their own judgments, even when these lack objective validity (Kahneman, 2011). But subjective confidence in subjective judgments should not be used in place of sound, objective scientific methods. To do so, as in EPA's risk assessment for ozone, replaces sound science with potentially arbitrary, biased, and mistaken judgments.

To obtain quantitative risk and benefit estimates, EPA applies a new quantitative model (The "MSS model") of the relation between ozone levels and decrements in lung function. This model makes predictions that depend sensitively on assumptions that are known to be incorrect (e.g., that individual variability is normally distributed) but that are made anyway for the sake of convenience. There are several key uncertainties about the model assumptions and conclusions, none of which has been quantified.

EPA also develops an approach to quantitative risk assessment based on epidemiological data in section 7.1.2 that makes the crucial error of treating the *slope* of a curve (i.e., the change in y divided by change in x) as if it showed the future changes in the variable on the y axis (health effects) that would be *caused* by changes in the variable on the x axis (exposure). This conceptual error invalidates EPA's quantitative risk and benefit predictions; it is a completely invalid interpretation of the slope of the model curve fit to past data (see Appendix A and Rothman et al., 2012).

A simple analogy may help to make this technical point clear. Dividing car accidents per year in a population by pounds of potatoes consumed per year in that population would produce a positive "slope factor" (i.e., ratio) linking these two quantities. One could meaningfully discuss the change in car accidents per unit change in potato consumption when discussing the slope of a regression curve fit to their historical values. But it would be utterly mistaken to interpret this as implying that reducing future potato consumption would *cause* a reduction in future car accidents. Figure 2 of Appendix A shows a more realistic example. EPA's risk assessment makes essentially this conceptual error, confusing the changes used to calculate slopes of exposure-response curves with the changes in health effects that causal mechanisms might cause if exposure were changed.

EPA's use of associational methods to drive key causal conclusions ignores a consensus in disciplines outside epidemiology that associational methods are unreliable, logically unsound, and inappropriate for drawing causal inferences (e.g., Dominici et al., 2014; Rothman et al., 2012; Cox and Popken, 2015). EPA's judgment that model-based associations are causal also conflicts with quantitative causal analyses of historical data. Studies that have applied formal quantitative causal analysis methods have found no detectable human health benefits caused by past reductions in ambient ozone levels (e.g., Moore et al., 2012; Cox and Popken, 2015). These studies provide no objective reason to expect that further reductions in the ozone standard will cause future human health benefits.[7]

The absence of any impact of relatively large reductions in ozone concentrations on public health outcomes in previous causal analyses suggests that there may be an exposure below which diseases do not occur, and that this threshold may be well above the current standard. If this is true, a NAAQS set higher than current levels would meet the statutory standard. This is consistent with proposed causal mechanisms involving inflammation of the lung, although such thresholds would not be apparent in population data unless errors in exposure estimates are properly modeled (Cox, 2011, 2012). EPA considers and rejects the hypothesis of an exposure threshold for adverse effects that is above current standards, but does not quantify exposure uncertainties and their effects on detection and estimation of thresholds. Such quantitative uncertainty analysis is essential for correct inferences about thresholds.

In summary, EPA's quantitative risk estimate (QRA) provides no legitimate reason to believe that the proposed action is "requisite to protect public health" or that reducing the ozone standard further will cause any public health benefits. The QRA's model-based projections to the contrary are known to rely on mistaken assumptions (for the MSS model) and mistaken interpretations of curve-fitting (for the epidemiological risk assessment in Section 7). Past data on human health before and after reductions in ozone do not reveal any such causal impacts. Given EPA's information and the unquantified model uncertainty that remains, there is no sound technical basis for asserting with confidence, based on the models and analyses in EPA's ozone risk assessment, that an ozone standard of 65 ppb would be any more protective than 70 ppb, or that 80 ppb is less protective than 60 ppb. To the contrary, available data suggest that further reductions in ozone levels will make no difference

[7] EPA's risk assessment cites a 2008 paper by Moore et al. suggesting that ozone causes harm to human health, based on untestable modeling assumptions, but does not cite the 2012 paper that superseded it, which found no causal relation between them.

to public health, just as recent past reductions in ozone have had no detectable causal impact on improving public health.

Critical Review of EPA Ozone Health Risk Assessment

EPA's approach to "causal determination" reflects a half century of departure of epidemiology from sound scientific and statistical methods of causal modeling and analysis. Starting in the 1960s, epidemiology took a turn away from other sciences and engineering disciplines by developing a unique, unproven, and logically unsound approach to assessing causation; this approach, now incorporated into weight-of-evidence systems, lacks justification by underlying principles or by demonstrated logical or practical validity (Morabia, 2013).

Following an influential essay by Hill (1965), epidemiologists began relying on their own personal judgments and consideration of various aspects of association to guess whether statistical associations might be causal. Regulatory risk assessments started substituting consensus opinions based on these guesses for more objective and accurate methods of assessing causality. This judgment-based approach, which forms the main basis for EPA's ozone risk assessment, has never been shown to necessarily or usually produce correct results (Morabia, 2013). To the contrary, empirical research on the accuracy of expert judgments has found that they are usually biased and mistaken (Kahneman, 2011). Biases toward false positives (e.g., believing that proposed actions will cause benefits that they turn out not to cause) predominate (Lehrer, 2012; Ioannadis, 2005, Lehrer, 2012, Sarewitz, 2012, Ottenbacher, 1998; Imberger et al., 2011, *The Economist*, 2013).

Comment 1: Associational methods are unreliable guides to valid causal inference

EPA experts may (and demonstrably sometimes do) consider that it "could not be more clear" that observed associations between air pollution levels and adverse health effects are causal (e.g., Harvard School of Public Health, 2002) even though more thorough and objective quantitative causal analyses would reveal that they are not – that they only reflect coincident historical trends or other non-causal explanations (e.g., HEI, 2013). In general, substituting qualitative expert judgments about whether associations are probably causal (e.g., using the "Aspects to aid in judging causality" listed in Table 1 of EPA's ozone ISA) for rigorous quantitative causal analysis (e.g., using the techniques listed in Table 1 of Appendix A) lacks justification (Morabia, 2013). It is part of the sociology of applied science that motivated reasoning, confirmation bias, wishful thinking, over-confidence in one's

own judgments, and other biases (Kahneman, 2011) can and do routinely produce confident but mistaken beliefs that causal relations have been discovered when in fact there is no objective evidence of causation; as a result, projected benefits from costly actions, based on the mistaken expert judgments, often fail to materialize after the costly actions have been taken (Lehrer, 2012; Ioannadis, 2005, Sarewitz, 2012, Ottenbacher, 1998; Imberger et al., 2011, *The Economist*, 2013).

More fundamentally, association-based methods (e.g., regression modeling, WoE judgments), relied on throughout EPA's ozone risk assessment, are not in general reliable guides to the objective truth; for example, they can enable modelers to find either statistically "significant" positive exposure-response associations or statistically "significant" negative ones, depending on their modeling choices (Dominici et al., 2014). Associational methods are not adequate or valid for allowing accurate predictions of the effects that future changes in exposures would cause in future health risks. Appendix A and its references develop this point in more detail.

Comment 2: EPA's ozone risk assessment depends on associational methods.

That EPA's ozone risk assessment depends heavily on conflating associations with causation is well illustrated by the following passages:

> "**6.2.3.2 Asthma**
>
> In reference to epidemiologic studies, the ISA states that '[t]he evidence supporting associations between short-term increases in ambient O_3 concentration and increases in respiratory symptoms in children with asthma is derived mostly from examination of 1-h max, 8-h max, or 8-h avg O_3 concentrations and a large body of single-region or single-city studies. The few available U.S. multicity studies produced less consistent associations.' (ISA, p. 6-101 to 6-102). 'Although recent studies contributed mixed evidence, the collective body of evidence supports associations between increases in ambient O_3 concentration and increased asthma medication use in children' (ISA, p. 6-109)." pp 6-7 to 6-8
>
> "**Mortality Effects**
>
> Recent multicity studies and a multicontinent study have reported associations between short-term O_3 exposure and mortality, expanding upon evidence available in the last review (see Section 6.6). These recent studies reported consistent positive associations between short-term O_3 exposure and total (nonaccidental) mortality, with associations being stronger during

the warm season, when O_3 concentrations were higher. They also observed associations between O_3 exposure and cardiovascular and respiratory mortality. These recent studies also examined previously identified areas of uncertainty in the O_3-mortality relationship, and provided additional evidence supporting an association between short-term O_3 exposure and mortality. As a result, the current body of evidence indicates that there **is likely to be a causal relationship between short-term exposures to O3 and total mortality.**"

"**6.6 Mortality**

6.6.1 Summary of Findings from 2006 O_3 AQCD

The 2006 O_3 AQCD (U.S. EPA, 2006b) reviewed a large number of time-series studies… The association between short-term O_3 exposure and mortality was substantiated by a collection of meta-analyses and international multicity studies. … Overall, the 2006 O_3 AQCD identified robust associations between various measures of daily ambient O_3 concentrations and all-cause mortality, with additional evidence for associations with cardiovascular mortality, which could not be readily explained by confounding due to time, weather, or copollutants. … Collectively, the 2006 O_3 AQCD concluded that "the overall body of evidence is highly suggestive that O_3 directly or indirectly contributes to non-accidental and cardiopulmonary-related mortality."

6.6.2 Associations of Mortality and Short-Term O3 Exposure

Recent studies that examined the association between short-term O_3 exposure and mortality further confirmed the associations reported in the 2006 O_3 AQCD. New multicontinent and multicity studies reported consistent positive associations between short-term O_3 exposure and all-cause mortality in all-year analyses, with additional evidence for larger mortality risk estimates during the warm or summer months (Figure 6-27 [and Table 6-42]). These associations were reported across a range of ambient O_3 concentrations that were in some cases quite low."

"**6.6.3 Summary and Causal Determination**

The evaluation of new multicity studies that examined the association between short term O_3 exposure and mortality found evidence which supports the conclusions of the 2006 O_3 AQCD. These new studies reported

> consistent positive associations between short-term O_3 exposure and all-cause (nonaccidental) mortality, with associations persisting or increasing in magnitude during the warm season, and provide additional support for associations between O_3 exposure and cardiovascular and respiratory mortality."
>
> "**7.1.2 Calculating Ozone-Related Health Effects Incidence**
>
> The C-R functions used in the risk assessment are empirically estimated associations between average ambient concentrations of O_3 and the health endpoints of interest (e.g., mortality, hospital admissions (HA), emergency department (ED) visits). … Given a C-R function of the form shown in equation (1) and a particular difference in ambient O_3 levels, Δx, the RR associated with that difference in ambient O_3, denoted as $RR\Delta x$, is equal to $e^{\beta\Delta x}$. The difference in health effects incidence, Δy, corresponding to a given difference in ambient O_3 levels, Δx, can then be calculated based on this $RR\Delta x$… These health impact equations are the key equations that combine air quality information, C-R function information, and baseline health effects incidence information to estimate ambient O_3 health risk... "

It is clear from such passages that EPA's risk assessment makes extensive use of associations, and that it repeatedly misinterprets evidence of associations as evidence for causation. This fundamental error leads to EPA's "Causal Determination" that reducing current ambient ozone levels would improve human health.

EPA's approach to quantitative risk assessment based on epidemiological data in section 7.1.2 also mistakenly conflates association and causation, completely confusing (a) "The difference in health effects incidence, Δy, corresponding to a given difference in ambient O_3 levels, Δx" as determined by the slope of a curve fit to *past levels* of ozone and health effects incidence rates with (b) The future *change* in health effects incidence that would be caused by a future change in ambient O_3 levels.

This confusion invalidates EPA's resulting quantitative benefits assessment. As discussed in the above example of potato consumption and car accidents and as elaborated further in Appendix A (illustrated with an example of baby aspirin and heart attack risk, and in the context of Figure 2 showing a linear slope for historical data plotting cardiovascular mortality against cell phone ownership), the *slope* of an exposure-response regression curve is an entirely distinct concept from the *causal impact* of changes in exposure on changes in response. The two have no necessary relation to each other. They need not even have the same sign, e.g., a positive

exposure-response slope may describe historical data even if the causal relation between exposure and response is negative, so that reducing exposure would cause increased future risk of response.

EPA's use of associational methods to drive key causal conclusions ignores a consensus in disciplines outside epidemiology that associational methods are unreliable, logically unsound, and inappropriate for drawing causal inferences (Dominici et al., 2014; Rothman et al., 2012; Cox and Popken, 2015). As illustrated in the above passages, EPA repeatedly moves straight from reports of associations to conclusions about causation without applying any formal quantitative causal analyses or objective tests to check the validity of their conclusions (e.g., using the methods in Hernan and Robins, 2011, 2015 and in the references cited in Table 1 of Appendix A).

Comment 3: EPA has not correctly "established" or "determined" causation or demonstrated that reducing ozone standards would cause human health benefits

Although EPA presents no objective causal analysis showing that reducing current ozone standards and levels will cause (or in recent decades has caused) any human health benefits, its QRA nonetheless projects human health benefits to be achieved by proposed future reductions in ozone. To accomplish this, the ozone risk assessment applies models of how future human health benefits might change as ozone levels changed *if* there were a causal relation between them and *if* that causal relation were correctly described by the assumed models. Uncertainties about these two crucial conditions are not quantified.

No detectable causal relation between ozone reductions and human health benefits has been found in careful examinations of historical data (e.g., HEI, 2010; Moore et al., 2013; Cox and Poken, 2015), casting doubt on the validity of EPA"s assumption that such a causal relation will hold in future. As discussed in the following comments, there is little reason to have confidence that EPA's models used to quantify projected future benefits provide even approximately correct descriptions of available data.

Comment 4: EPA's risk assessment models and their conclusions and predictions depend on implausible, unverified, and inaccurate modeling assumptions.

As explained in section 6.2.4 of the risk assessment, "In this review, EPA is investigating the use of a new model that estimates FEV1 responses for individuals

associated with short-term exposures to O_3 (McDonnell, Stewart, and Smith, 2007; McDonnell, Stewart, and Smith, 2010; McDonnell et al., 2012). This is a fundamentally different approach than the previous approach, for which the E-R function is at a population level, not an individual level. This model was developed using controlled human exposure data…" The new model that EPA is investigating involves a parameter alpha2 that is assumed to more than quadruple on one's 18th birthday, then more than double on one's 36th birthday, and finally plummet to zero on one's 55th birthday (Table 6-2, p. 6-13). Such jumps do not seem plausible.

EPA then selects two *ad hoc* functional forms (out of an infinite number of possibilities) and claims that this incorporates model uncertainty: "EPA considered both linear and logistic functional forms in estimating the E-R relationship and chose a 90 percent logistic/10 percent piecewise-linear split using a Bayesian Markov Chain Monte Carlo approach. This Bayesian estimation approach incorporates both model uncertainty and uncertainty due to sampling variability." A more sober assessment is that the correct functional forms are unknown, and that neither of the two considered forms has been shown to be even approximately correct; this enormous model uncertainty has not been correctly accounted for or quantitatively characterized in the analysis, results, or statement of conclusions (e.g., using Bayesian Model Averaging, as described in Samet and Bodurow, 2008).

Likewise, EPA states (Section 6.5.1) that "Clearly the intra-individual variability Var(å) in the MSS model is a key parameter and is influential in predicting the proportions of the population with FEV1 decrements > 10 and 15%. The assumption that the distribution of this term is Gaussian is convenient for fitting the model, but is not accurate. The extent to which this mis-specification affects the estimates of the parameters of the MSS model and its predictions is not clear."

Comment 5: EPA has not quantified crucial model uncertainties. Therefore, confidence intervals calculated assuming that the models used are correct are misleadingly narrow and EPA has provided policy makers with no basis for confident predictions about how different changes in the ozone standard would probably affect public health.

As acknowledged in EPA's risk assessment (Section 6.5.1), "Although the model does not have good predictive ability for individuals (psuedo-R2 0.28), it does better at predicting the proportion of individuals with FEV1 decrements. 10, 15, and 20% (psuedo-R2s of 0.78, 0.74, 0.68) (McDonnell et al., 2012). The clinical studies that these model estimates are based on were conducted with young adult volunteers rather than randomly selected individuals, so it may be that selection bias has

influenced the model parameter estimates. The parameter estimates are not very precise, partly as the result of correlations between the parameter estimates… The MSS model is also sensitive to the exposure concentrations, but we have not quantified that sensitivity. … We are unable to properly estimate the true sensitivities or quantitatively assess the uncertainty of the MSS model. … As discussed in Section 6.5.3 below, there are uncertainties in extrapolating the MSS model down to age 5 from the age range of 18 to 35 to which the model was fit. …[T]he uncertainty of the extension to children of the MSS model could be substantial."

Section 6.5.7 adds that "EPA staff have identified key sources of uncertainty with respect to the lung function risk estimates. These are: the physiological model in APEX for ventilation rates, the O_3 exposures estimated by APEX, the MSS model applied to ages 18 to 35, and extrapolation of the MSS model to children ages 5 to 18. … At this time we do not have quantitative estimates of uncertainty for any of these."

Comment 6: EPA has not demonstrated that its recommended reductions in the current ozone standard will cause public health benefits.

EPA's ozone risk assessment does not provide quantitative risk models and uncertainty analyses adequate to support confident or accurate predictions about the future public health consequences of different standards. In this respect, its risk assessment is crucially incomplete. Its key conclusions and recommendations are based only on expert judgments and unverified assumptions. They are not supported by causal analyses of historical data, which show no detectable human health benefits caused by substantial reductions in ozone levels (e.g., HEI, 2103; Moore et al., 2013; Cox and Popken, 2015).

In summary, EPA is using a new model that takes "a fundamentally different approach than the previous approach," and that knowingly adopts convenient but inaccurate modeling assumptions (mis-specified models), to predict substantial future health benefits from further ozone reductions. No such benefits caused by past reductions in ozone levels have been found in careful examinations of past data (HEI, 2010; Moore et al., 2013; Cox and Poken, 2015). Since the new model is known not to be valid (it is mis-specified for convenience) and key uncertainties about it have not been quantified, there is no legitimate reason for EPA or policy makers to trust its predictions that further ozone reductions will cause substantial health benefits, particularly given that empirical observation does not support causal connections between past reductions in ozone levels and these health benefits.

As a result of its use of associational models and methods and unverified and incorrect modeling assumptions and interpretations, *EPA has not established causality* for the health effects that it attributes to ambient levels of ozone. Its claims to have done so – including its "Causal determinations" – are based on applying expert judgments, which are known to be highly fallible (Kahneman, 2011), to associational studies, which are known to be unreliable and inadequate for drawing valid causal inferences (Dominici et al., 2014; Cox and Popken, 2015; Appendix A). These qualitative subjective judgments are contradicted by more objective and reproducible quantitative analyses that have found no human health benefits caused by past reductions in ozone, and that have shown that health benefits initially attributed to ozone reductions were in fact just artifacts of flawed statistical methodology (e.g., HEI, 2010; Moore et al., 2013; Cox and Poken, 2015). Hence, the risks that EPA's assessment attributes to current ambient levels of ozone are unsupported by sound, objective, and reliable analysis or by past experience.

Furthermore, EPA's ozone risk assessment has not demonstrated that these public health risks, even if they exist, would be reduced at all by reducing the current ozone standard. (It *assumes* this in fitting models to data in Section 6 and in misinterpreting curve-fitting as if it were causal modeling in Section 7, as previously discussed. But an assumption is not a demonstration.)

The risk reductions that EPA projects from the proposed reductions in ozone standards are no more than consequences of unsupported modeling assumptions. These projections, which might seem to have been refuted by past experience, are not supported by any analyses showing that past ozone reductions have caused the benefits that EPA's models would have predicted for them, or explaining why those projected benefits have not been found in practice. Rather, they substitute optimistic judgments from selected experts that benefits will occur for objective scientific analysis and factual evidence that they have not occurred in places where ambient levels of ozone have already been decreased substantially in recent decades (e.g., HEI, 2010; Moore et al., 2013; Cox and Poken, 2015; HEI, 2010).

Comment 7: EPA's ozone risk assessment can and should be improved by applying readily available formal quantitative methods of causal analysis and uncertainty characterization.

Quantitative methods for sound causal analysis and characterization of remaining uncertainties, pioneered mainly in disciplines outside epidemiology, have progressed rapidly since the 1960s. They are now readily available and widely applied in social statistics, econometrics, engineering, neuroscience, artificial intelligence, machine

learning, and statistics, among others and are increasingly being incorporated into the best practices of modern epidemiology (Rothman et al., 2012).

Appendix A outlines both the pitfalls of association-based methods, and constructive methods for more objectively and correctly assessing and quantifying causal relations from data (see e.g., Table 1 of Appendix A). A key idea is that causation is *not* primarily about statistical associations between historical levels of variables, such as exposure and health effects. Rather, it is primarily about how changes in inputs (such as exposure) propagate through a network of validated causal mechanisms to cause resulting changes in outputs (such as health effects). Appendix A discusses methods to determine, document, and quantify such genuine causal influences.

The following steps would make EPA's risk assessment for ozone more sound and useful.

1. Formally test whether there is any significant statistical association between past changes in ambient levels of ozone and changes in risks of mortality or adverse health effects after accounting for modeling biases and uncertainties (e.g., by using Bayesian Model Averaging or nonparametric methods).

2. Formally test whether past changes in ozone exposures have preceded and helped to explain the changes in public health effects that EPA's risk assessment claims they will cause. Technical methods for conducting such tests (e.g., change-point analyses, intervention analyses, Granger causality tests) are readily available and should be used.

3. Formally test whether any health effects attributed to past changes in ozone levels can be made conditionally independent of those ozone levels by conditioning on other variables (e.g., daily temperature). If not, as revealed by conditional independence tests, this would provide legitimate evidence for possible causation.

4. Present and then test explicit causal graph models for how changes in ambient ozone levels propagate along hypothesized causal paths (possibly involving changes in reactive oxygen species (ROS) and lung inflammation) to produce changes in mortality or other health effects. As mentioned by Samet and Bodurow (2008) in a report cited by EPA, one way to do this is to "… [D]raw a set of causal graphs, each of which represents a particular causal hypothesis, and then consider evidence insofar as it favors one or more of these hypotheses and related graphs over the others."

5. Use the causal graphs to identify, test, and if possible refute, non-causal explanations for statistical relations among observed variables (including exposures, health effects, and any intermediate variables, modifying factors, and confounders).

6. Show that causal mechanisms postulated in the QRA modeling hold for realistic changes in ambient concentrations, and that they exhibit stable, uniform, law-like behavior. Ideally, there should be no large unexplained heterogeneity in estimated input-output (e.g., E-R or C-R) relations, since heterogeneity should be explained by differences in input variables such as individual covariates (rather than by differences in the causal mechanism or laws through which changes in inputs propagate to produce changes in outcomes).

7. Explain why past substantial reductions in ozone levels have not led to the predicted (and often claimed) health benefits predicted by modeling assumptions and associational modeling (e.g., HEI, 2010). The failure of predicted benefits to materialize in the past is a key reason for suspecting that they may not materialize in the future, and that EPA's risk analysis modeling methods and assumptions are overly optimistic and lack predictive validity. Addressing this gap between predictions and reality is crucial for developing trustworthy causal models of how changes in ozone exposures will affect public health risks.

The following sections expand on the preceding points in the context of specific passages from the EPA risk assessment for ozone and the source documents on which it relies.

Critical Comments on EPA's Causal Determination Framework

The comments that follow are based on the principle that sound quantitative risk assessment (QRA) must be based on sound analysis of causal relations between proposed risk management actions, such as reducing the current ozone standard, and their probable consequences, such as changes in health effects. Appendix A reviews principles for how causal analysis for quantitative risk assessment *should* (and should not) be done. This covers not only pitfalls to avoid, based largely on concerns about over-reliance on expert judgments and under-reliance on more objective methods of causal analysis and modeling at IARC, EPA, OSHA, FDA, and other agencies; but also principles for valid causal analysis, inference, and quantitative risk assessment (QRA), drawn from a large technical literature on causal analysis.

EPA's "Aspects to Consider in Making Causal Judgments" Do Not Accurately Reveal Causation

Table 1 of the ozone ISA summarizes "aspects" of association that are frequently considered in making expert judgments about whether associations might be causal. EPA used them as a primary basis for its causal determinations, principle conclusions on causality, and predictions that reducing ozone standards further will cause future health benefits. These "aspects" of association are reviewed here, along with critical comments.

1. **Strength:** It is often asserted that the larger is a statistical association between exposure and response (i.e., adverse health effects), the more likely it is to be causal. Table 1 of EPA's ISA document states that "The finding of large, precise risks increases confidence that the association is not likely due to chance, bias, or other factors." This belief is unjustified. It is a form of the logical fallacy that confuses correlation and causation, modified to suggest that *strong* correlation constitutes evidence for causation. As reviewed in Appendix A (see e.g., the discussion of Figure 2), a strong association may (and often does) simply reflect strong confounding, strong modeling biases, or strong spurious correlation between causally unrelated historical trends. Nor does a stronger association necessarily imply greater probability of causality; indeed, stronger reported associations may be *less* likely to be causal, as they are more likely to reflect strong modeling biases (e.g., Ioannidis, 2005).

The practical importance of modeling choices as explanations for reported significant positive associations for ozone and health effects was highlighted in a recent study by Moore et al. (2012), who carried out a causal analysis "with the aim of estimating the extent to which reductions in ambient ozone concentrations were associated with measurable health benefits, specifically on the proportion of asthma-related hospital discharges." Unlike previous analyses that made untestable modeling assumptions, this causal analysis found no significant effects of ozone reduction on health benefits, prompting the authors to speculate that "Thus, in this ozone study, significant results from the [earlier] analysis may be a consequence of the approach taken and not solely based on the information in the data."

In EPA's ozone risk assessment, the extent to which significant associations are consequences of the modeling approaches taken, including unverified and possibly mistaken modeling assumptions, rather than reflecting information in the data, is not quantified. Thus, policy-makers have no way to know whether any of the key

conclusions and recommendations is more than a consequence of untested modeling assumptions.

Likewise, Cox and Popken (2015) applied a different causal analysis method (Granger causality testing) to time series data on ozone concentrations and mortality rates in hundreds of counties in the United States. They found a clear positive exposure-response association in regression models, consistent with EPA's findings, but no evidence that the association was causal rather than coincidental (i.e., both ozone levels and mortality rates, especially cardiovascular mortality rates, have been declining, but not because one causes the other).

Again, since EPA's ozone risk assessment treats association as causal without applying any formal quantitative tests of this assumption, it is impossible for policy makers to know whether any of the causal impacts of reducing ozone that EPA projects will actually occur. However, the findings from Moore et al. (2012) and Cox and Popken (2015) indicate that past reductions in ozone have not caused any detectable health benefits for asthma-related hospitalization or mortality risk reductions, respectively. This calls into question EPA's central assumptions that associations imply causation and that future health benefits can be achieved by reducing the current ozone standard.

For ozone, EPA and others have found that pollutant concentrations and adverse health effects are often *negatively* associated, unless modeling constraints are imposed or modeling adjustments are made to force them to be positive (Powell et al., 2012). A finding of a strong positive association made when other possibilities are eliminated via modeling provides no evidence for a true causal relation.

For example, EPA's risk assessment notes in 6.6.2.1 that "Because O_3 peaks in the summer and mortality peaks in the winter, not adjusting or not sufficiently adjusting for the seasonal trend would result in an apparent negative association between the O_3 and mortality time-series. …However, it should be noted, the majority of studies in the literature that examined the mortality effects of short-term O_3 exposure, particularly the multicity studies, used 7 or 8 df/year to adjust for seasonal trends, and in both methods a positive association was observed between O_3 exposure and mortality." Again, beginning with a negative correlation and then "sufficiently adjusting" it to make it positive raises the possibility that reported positive associations reflect the willingness of modelers to make adjustments as needed to produce them.

This does not deny the desirability and importance of adjusting for seasonality (although adjusting for daily temperatures might be more important and reduce

modeling biases due to insufficient flexibility of seasonal adjustments). But such adjustments create a strong requirement to quantify and report biases due to multiple testing (i.e., searching through multiple possible model specifications, e.g., using different lags and degrees of freedom for smoothers, until the desired result is obtained) and model specification errors. EPA's ozone risk assessment, and the key studies on which it depends (e.g., the ozone ISA) do not quantify the effects of these biases on results. Key model uncertainties are not quantified. Thus, policy makers cannot determine from EPA's risk assessment to what extent reported associations and recommendations simply reflect mistaken modeling assumptions (e.g., model selection biases and misspecification errors).

When model diagnostics are not presented and model uncertainties are not quantified, as in EPA's ozone risk assessment, it is always possible to create strong positive associations using misspecified statistical model (see Appendix A). For example, strong positive associations can be created by misspecified linear regression modeling assumptions even in purely random data where it is known that there is no causal relation (and no true associations, apart from that created by modeling) between exposure and response variables. Appendix A makes this point via a simple example using a linear regression model that is misspecified to have a zero intercept term. However, the point that modeling assumptions can create strong positive associations independent of the true empirical relation – positive, zero, or negative – between exposure and risk is much more general.

In summary, the finding of a strong association does not necessarily or usually give any support to the hypothesis that the association is causal, rather than a product of strong modeling assumptions, biases, confounding, or coincident trends.

2. **Consistency:** A second unwarranted belief is that a positive association that is consistently reproduced by different researchers is more likely to be causal than an association that has not been reproduced. As EPA states in multiple documents, including Table 1 of their ozone ISA, "The reproducibility of findings constitutes one of the strongest arguments for causality." However, consistently reproduced findings may simply reflect consistent errors and biases in modeling, or consistent failures to fully account for confounders.

For ozone and mortality, there is no consistency in exposure-response associations. As noted in EPA's ozone risk assessment (section 6.6.2.3), "Evaluation of the O_3-mortality C-R relationship is not straightforward because the evidence from multicity studies (using log-linear models) suggests that O_3-mortality associations are highly heterogeneous across regions." For older adults, "The inconsistent epidemiologic

findings for older adults parallel observations from controlled human exposure studies (Section 6.2.1.1)." For asthma, "The few available U.S. multicity studies produced less consistent associations (ISA, p. 6-101 to 6-102)." Figure 6-4 shows that different studies of campers disagree even about the signs of associations (each with 95% confidence, which suggests that the alleged 95% confidence intervals are not trustworthy, as opposite signs of association should not be observed this frequently if they were correct).

Moreover, as already discussed, there is no consistency of exposure-response relations across studies and models. Thus, Moore et al. (2012) showed that statistically significant health benefits from ozone reduction are found in models that make unverifiable assumptions, but not in models that use only realistic, empirically tested assumptions. Similarly, Powell et al. (2012) comment on the inconsistency of associations across studies (some find negative associations and others positive ones, all of which are claimed to be statistically significant) and advocate constraining studies to force a consistent finding of positive associations rather than negative ones. (Roberts (2004) recommends similar measures for similar reasons.) Clearly, consistency of positive associations that is enforced when necessary by discarding equally valid findings of negative associations cannot be correctly construed as providing any evidence for causation, let alone being regarded as "one of the strongest arguments for causality."

3. **Specificity:** Table 1 of EPA's ozone ISA report states that "Evidence linking a specific outcome to an exposure can provide a strong argument for causation." The link between asbestos exposure and mesothelioma risk is a common example. For ozone at and below currently permitted levels, no such specific outcomes are claimed, so this aspect of association does not apply. However, it is worth noting that confounding can lead to specificity of an exposure-response association even without causality.

4. **Temporality:** EPA asserts that "Evidence of a temporal sequence between the introduction of an agent, and appearance of the effect, constitutes another argument in favor of causality." However, this argument is unsound. It is a form of the *ex post* logical fallacy. Although EPA is correct that causes precede their effects, sound inference about causation requires much more of temporal order, including the following.

    a. To count as evidence that exposure causes increased risk of adverse health effects, the temporal sequence should enable health effects to be predicted better with knowledge of the exposure time series than without

it. This condition was not found to hold for ozone and mortality risks when it was formally tested using Granger causality and conditional independence tests (Cox and Popken, 2015).

b. The temporal sequence should not be due to coincident historical trends. "History" is recognized as a standard threat to valid causal inference in quasi-experiments (see Appendix A and references therein). EPA experts who have offered confident expert judgments about the causality of air pollution health effects associations while ignoring the need to refute history as an explanation have thereby reached incorrect conclusions (HEI, 2013).

c. The temporal sequence should not be due to natural variability in time series (perhaps combined with selection of time windows by modelers to give an appearance of temporality). For ozone, a well-publicized claim of a causal impact is "the reduction in the rates of childhood asthma events during the 1996 Summer Olympics in Atlanta, Georgia, due to a reduction in local motor vehicle traffic" (Buka et al., 2006). Belief in this reduction as evidence of a clear causal effect between pollution levels and asthma risk was spread by a 2001 article in the *Journal of the American Medical Association*, which reported that "During the Olympic Games, the number of asthma acute care events decreased 41.6% (4.23 vs 2.47 daily events) in the Georgia Medicaid claims file," coincident with significant reductions in ozone and other pollutants (Friedman et al., 2001).

   Unfortunately, the investigators did not formally test whether the observed reduction in asthma cases is unusual for the relevant time of year. When other investigators compared fluctuations in asthma acute care events during the 1996 Atlanta Olympics to fluctuations over the same period in other years, they "did not find significant reductions in the number of emergency department visits for respiratory or cardiovascular health outcomes in adults or children" (HEI, 2010). Although "They confirmed that Atlanta experienced a significant decline in ozone concentrations of 20% to 30% during the Olympic Games, with less pronounced decreases in concentrations of carbon monoxide, PM_{10}, and nitrogen dioxide," these significant reductions in pollutant levels were not matched by any detectable reduction in health risks: "In their primary analyses, which were adjusted for seasonal trends in air pollutant concentrations and health outcomes during the years before and after the

Olympic Games, the investigators did not find significant reductions in the number of emergency department visits for respiratory or cardiovascular health outcomes in adults or children." In fact, the only significant effect was that "relative risk estimates for the longer time series were actually suggestive of increased ED [emergency department] visits during the Olympic Games," primarily for Chronic Obstructive Pulmonary Disease (COPD) patients (HEI, 2010).

In short, although the reduction in air pollution levels during the Olympics was real and substantial (although perhaps coincidental, as similar reductions occurred at the same time in much of the region), the only significant change noted in adverse health effects, compared to other years, was a modest *increase* in emergency department admissions. The widely publicized "reduction in the rates of childhood asthma events during the 1996 Summer Olympics in Atlanta, Georgia, due to a reduction in local motor vehicle traffic" (Buka et al., 2006) appears to be an artifact of poor statistics – specifically, failing to quantify natural seasonal variability in event rates before interpreting a significant (over 40%) reduction as evidence of a causal impact.

5. **Biological gradient:** Table 1 of EPA's ISA asserts that "A well-characterized exposure-response relationship (e.g., increasing effects associated with greater exposure) strongly suggests cause and effect, especially when such relationships are also observed for duration of exposure (e.g., increasing effects observed following longer exposure times)." This is fallacious, insofar as model misspecification or a strong confounder can create a positive association between exposure (concentration and duration) and effect in the absence of any causal relation between them, as explained further in Appendix A (see e.g., the discussion of Figure 1).

6. **Biological plausibility:** For biological plausibility, Table 1 of EPA's ISA states that "A proposed mechanistic linking between an effect and exposure to the agent is an important source of support for causality, especially when data establishing the existence and functioning of those mechanistic links are available." Of course, a proposal is not by itself "an important source of support for causality" if the proposal is wrong and the proposed mechanisms do not operate in reality. Nor do "data establishing the existence and functioning of those mechanistic links" provide support for causality unless the mechanism actually operate in the setting of interest. For example, proposing a vague explanation ("hand waving") that refers to experimental data with no known

relevance to real public health changes should not be considered "an important source of support for causality." More generally, many mechanistic proposals suggested by *in vitro* findings, high-concentration experiments, or other experiments carried out under artificial conditions may prove to be mistaken; they should not be counted as evidence for causation. Similarly, proposed mechanisms that seem plausible only because of ignorance of relevant biology and pathology should not be construed as evidence for causation.

For ozone, EPA further explains that "The O_3-induced lung function decrements consistently demonstrated in controlled human exposure studies (Section 6.2.1.1) provide biological plausibility for the epidemiologic evidence consistently linking short-term increases in ambient O_3 concentration with lung function decrements in diverse populations." This claim is problematic, in so far as both "The O_3-induced lung function decrements consistently demonstrated in controlled human exposure studies" and the "epidemiologic evidence consistently linking short-term increases in ambient O_3 concentration with lung function decrements" are in fact far from consistent; see the discussion of consistency above.

7. **Experimental evidence:** In EPA's view, "Strong evidence for causality can be provided through 'natural experiments' when a change in exposure is found to result in a change in occurrence or frequency of health or welfare effects." For ozone, such natural experiments have shown that relatively large changes in exposure (e.g., 40% reductions) have resulted in no detectable reductions in the occurrence or frequency of health effects (e.g., HEI, 2010 for the Atlanta Olympics; Cox and Popken, 2015 for U.S. counties; Moore et al., 2013 for hospitalization). It is perhaps fair to wonder whether these natural experiments should be construed as providing strong evidence *against* EPA's hypothesis that future (relatively small) reductions in ozone will cause public health benefits. EPA's risk assessment and ISA document are silent on this issue. No discussion or explanation is offered for why projected public health benefits caused by ozone reductions are not found in past data. This is an area in which research might provide answers to inform future policy. For example, if the explanation for the apparent lack of effects of previous ozone reductions on public health is that there are two main types of people in the population, insensitive or healthy people whose health does not change when ozone exposures are reduced; and very sensitive responders who continue to respond just as frequently and in the same ways even when ozone levels are reduced, then this understanding might also lead to better predictions about the public health effects (or lack of them) to be expected from future reductions in ozone levels. Understanding historical data better, and being able to explain the results of quantitative causal analyses of past

data from natural experiments, seems to be a useful prerequisite for confident forecasting about the probable effects of future policy actions.

8. **Coherence.** Table 1 of EPA's ozone ISA also explains that "An inference of causality from one line of evidence (e.g., epidemiologic, controlled human exposure [clinical], or animal studies) may be strengthened by other lines of evidence that support a cause-and-effect interpretation of the association." However, combining lines of evidence none of which suggests that reducing ozone has caused public health benefits does not strengthen an inference of causality. Conversely, a single line of evidence (e.g., natural experiments showing no detectable effects of past ozone reductions on public health) suffices to refute any number of lines of evidence that mistakenly predicted that clear public health benefits should have been seen.

9. **Analogy:** The ozone ISA's table of "Aspects [of association] to aid in judging causality" concludes with these observations: "Structure activity relationships and information on the agent's structural analogs can provide insight into whether an association is causal. Similarly, information on mode of action for a chemical, as one of many structural analogs, can inform decisions regarding likely causality."

   For ozone, the ISA presents extensive discussions of whether a mode of action involving increased reactive oxygen species (ROS) and inflammation in the lung might be plausible. However, as detailed in previous biomathematical analyses (Cox, 2011, 2012), inflammation-mediated lung diseases should be expected to have exposure thresholds below which they are not caused. Epidemiological models that do not explicitly and adequately model errors in exposure estimates, including EPA's epidemiological models for ozone health effects in Section 7, smear out such thresholds, making them appear to be smooth dose-response relations. (They also tend to inflate estimated no-observed-adverse-effects levels, so that estimated thresholds may be far higher than true ones.)

   EPA's ISA presents experimental results on lung responses to ozone in human volunteers. These results constitute overwhelming evidence that the subjects were alive, and hence had lungs that responded normally to normal challenges. But they do not analyze thresholds in the context of exposure estimation errors. Hence, they do not confront the fact that the proposed inflammation-mediated mechanisms of adverse health effects are analogous to those for other much-studied lung irritants for which exposure thresholds exist which must be

exceeded (for both concentration and duration) to produce lung pathologies or diseases (*ibid*).

In short, analogy suggests that ozone probably has no adverse health effects at concentrations and durations below certain thresholds. The absence of any impact of relatively large reductions in ozone concentrations on public health outcomes in causal analyses, previously discussed, might suggest that such a threshold may be well above the current standard.

In summary, although we disagree with EPA that subjective qualitative reflections on aspects of association should ever be used in place of rigorous causal analysis and hypothesis testing as a basis for making claims about causality in exposure-response or concentration-response relations, we also consider that the "Aspects [of association] to aid in judging causality" listed in Table 1 of the ozone ISA tend to support the view that there is no evidence of a causal relation. That a reasonable application of these aspects might lead to a finding of no evidence of causation at least as easily as to a determination of causality, as suggested by the preceding comments, highlights the subjectivity and unreliability of such WoE-based judgments and determinations.

Critique of EPA's Risk Assessment Based on Other Criteria

EPA's ozone risk assessment may also be evaluated according to the checklist for sound QRAs offered in the last part of Appendix A as an alternative to the aspect listed in Table 1 of the ozone ISA. The results are as follows.

Does EPA's ozone QRA show that changes in exposures precede the changes in health effects that they are said to cause?

Answer: No. The QRA presents no results of any formal quantitative causal analyses (e.g., change-point analyses, intervention analyses, or Granger causality tests) showing that changes in ambient O_3 exposure concentrations precede, and help to predict and explain, subsequent changes in public health effects. For mortality risks, specifically, it appears that O_3 exposures are significantly positively *associated* with both all-cause (AC) and cardiovascular disease (CVD) mortality rates in historical data for U.S. counties, but the association does not appear to be *causal*, insofar as "A causal relation between pollutant concentrations and AC or CVD mortality rates cannot be inferred from these historical data, although a statistical association between them is well supported. There were no significant positive associations between changes in $PM_{2.5}$ or O_3 levels and corresponding changes in disease

mortality rates between 2000 and 2010, nor for shorter time intervals of 1 to 3 years." (Cox and Popken, 2015)

Does the QRA demonstrate that health effects cannot be made conditionally independent of exposure by conditioning on other variables (especially, potential confounders)?

Answer: No. No conditional independence tests are presented. However, EPA's risk assessment does discuss confounding, as follows.

> "**6.6.2.1 Confounding**
>
> Recent epidemiologic studies examined potential confounders of the O_3-mortality relationship. These studies specifically focused on whether PM and its constituents or seasonal trends confounded the association between short-term O_3 exposure and mortality."
>
> "**Confounding by Seasonal Trend**
>
> The APHENA study (Katsouyanni et al., 2009), mentioned above, also conducted extensive sensitivity analyses to identify the appropriate: (1) smoothing method and basis functions to estimate smooth functions of time in city-specific models; and (2) degrees of freedom to be used in the smooth functions of time, to adjust for seasonal trends. Because O_3 peaks in the summer and mortality peaks in the winter, not adjusting or not sufficiently adjusting for the seasonal trend would result in an apparent negative association between the O_3 and mortality time-series. Katsouyanni et al. (2009) examined the effect of the extent of smoothing for seasonal trends by using models with 3 df/year, 8 df/year (the choice for their main model), 12 df/year, and df/year selected using the sum of absolute values of partial autocorrelation function of the model residuals (PACF) (i.e., choosing the degrees of freedom that minimizes positive and negative autocorrelations in the residuals). Table 6-46 presents the results of the degrees of freedom analysis using alternative methods to calculate a combined estimate: the Berkey et al. (1998) meta-regression and the two level normal independent sampling estimation (TLNISE) hierarchical method. The results show that the methods used to combine single-city estimates did not influence the overall results, and that neither 3 df/year nor choosing the df/year by minimizing the sum of absolute values of PACF of regression residuals was sufficient to adjust for the seasonal negative relationship between O_3 and mortality. However, it should be noted, the majority of studies in the

> literature that examined the mortality effects of short-term O_3 exposure, particularly the multicity studies, used 7 or 8 df/year to adjust for seasonal trends, and in both methods a positive association was observed between O_3 exposure and mortality."

This analysis is limited by the fact that it adjusts for only two confounders ($PM_{2.5}$ and seasonal effects), and not for many others (e.g., daily temperatures). Moreover, it shows that EPA readily discovers a significant *negative* association between O_3 and mortality, but then applies various statistical adjustments to get a positive association instead – but without adjusting for the resulting multiple testing and model selection biases. In this context, and in the absence of any conditional independence tests, it is not clear what causal significance, if any, the positive associations that EPA's manipulations eventually produce might have. (Adjusting for confounding by season, month, temperature, etc. is certainly sensible, but whether any association between O_3 and adverse health effects remains after conditioning on all relevant confounders has not been answered, since conditional independence test results are not reported.)

Does the QRA present and test explicit causal graph models, showing the results of formal statistical tests of the causal hypotheses implied by the structure of the model (i.e., which variables point into which others, as in Figures 1 or 3)?

Answer: No

Have non-causal explanations for statistical relations among observed variables (including exposures, health effects, and any intermediate variables, modifying factors, and confounders) been identified and refuted using well-conducted and reported statistical tests?

Answer: No. As noted in EPA's ozone ISA (p. 6-220), "[M]ultiple uncertainties remain regarding the O_3-mortality relationship including: the extent of residual confounding by copollutants; factors that modify the O_3-mortality association; the appropriate lag structure for identifying O_3-mortality effects (e.g., single-day lags versus distributed lag model); the shape of the O_3-mortality C-R function and whether a threshold exists; and the identification of populations at-risk to O_3-related health effects." (Section 6.6.1). The hypothesis that modeling choices, uncertainties, and selection biases, multiple testing bias, and specification errors explain reported E-R and C-R associations has not been refuted. Neither has the hypothesis that confounders explain them.

Have any causal mechanisms postulated in the QRA modeling been demonstrated to exhibit stable, uniform, law-like behavior, so that there is no substantial unexplained heterogeneity in estimated input-output (e.g., E-R or C-R) relations?

Answer: No. To the contrary, EPA's risk assessment notes (6.6.1) that "The studies evaluated found some evidence for heterogeneity in O_3 mortality risk estimates across cities and studies" and (6.6.2.3) "Evaluation of the O_3-mortality C-R relationship is not straightforward because the evidence from multicity studies (using log-linear models) suggests that O_3-mortality associations are highly heterogeneous across regions." Thus, there is considerable unexplained heterogeneity, indicating that EPA has not yet identified stable causal relations that could be used to make valid predictions.

In summary, EPA's QRA for ozone meets none of the criteria identified in Section 1 for a sound risk assessment. As a result, we recommend that additional work be done to close the gaps just identified (e.g., perform and report conditional independence tests, Granger causality tests, causal graph modeling, correction for modeling biases and confounding (including residual confounding) by temperature and other major confounders, and resolution of heterogeneity. Doing so will provide the foundation for a QRA that produces trustworthy results rather than unjustified opinions as a basis for future regulation of O_3.

The following section discusses the above points in further detail in the context of specific passages from the EPA risk assessment for ozone and the source documents on which it relies.

Comments on Passages from EPA's Ozone Integrated Science Assessment

EPA's O_3 risk assessment relies heavily on the O_3 ISA (Integrated Science Assessment),[8] which explains its approach to determining causality as follows. We present the approach using quotes from the ISA, interspersed with comments:

> *"Using the causal framework described in the following section, EPA scientists consider aspects such as strength, consistency, coherence, and biological plausibility of the evidence, and develop causality determinations*

[8] http://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=247492#Download

on the nature of the relationships. Causality determinations often entail an iterative process of review and evaluation of the evidence." (p. lvi)

COMMENT: The words "evidence" and "relationships" here refer primarily to *statistical associations* in *historical* data. They do *not* refer to validated causal relations, mechanisms, or laws that could legitimately be used to predict how future changes in O_3 exposure will change future health effects. The assumption that past statistical associations are the same as future causal relations lacks justification.

COMMENT: The concept of "causality determination" is flawed. It is not usually the case that a statistical association is either causal or not, and hence the presumption that there is a binary truth ("causal" or "not causal") is incorrect. Rather, statistical associations may be partly due to biases, partly to confounding, partly to modeling choices and errors, partly to coincident historical trends, partly to genuine causality, partly to neglected errors in exposure estimates, and so forth. What needs to be determined is not whether a statistical relation is causal, but what fraction of it (if any) is causal. This, EPA has neither done, nor attempted to do. Yet, it is the only thing that matters for making useful predictions about how future changes in exposure will change future health effects (if at all). Making a "determination" that a statistical association is "causal" (implicitly meaning entirely causal) and then treating it as being entirely causal without quantifying or bounding the fractions that are due to biases, confounding, coincident trends, and other non-causal sources of association, leaves uncompleted the task of informing policy makers correctly about the change in health effects that would be caused by future changes in health effects. By relying on qualitative "causality determination" and then implicitly treating associations as being 100% causal for quantitative purposes, EPA has committed a logical error and produced misleading (exaggerated, upper-bound) estimates of the health consequences of changing future O_3 exposures.

COMMENT: Notwithstanding common practice in epidemiology (but not in other fields, such as engineering or neuroscience or econometrics, where causal relations are more usually determined by objective data analysis and explicit modeling and hypothesis-testing, rather than by subjective opinions), the consideration of "aspects such as strength, consistency, coherence, and biological plausibility of the evidence" does not provide the information that is logically necessary to develop valid "causality determinations."

The weight-of-evidence considerations used by EPA (and many others) have no known validity as guides to causality (Morabia, 2013). They are not necessary or sufficient for establishing causality, or for making it plausible. They do not, for

example rule out strong confounders with time delays, or coincident (but not causal) historical trends in exposure and response, or modeling choices and biases, or many other possible non-causal explanations for strong, consistent, coherent associations that might also seem biologically plausible (especially in the absence of well-explicated, tested, and validated causal mechanisms explaining how ambient levels of exposure would create increased health effects).

On the other hand, the information that is most essential for drawing valid causal inferences—showing how changes in exposures propagate through causal mechanisms to increase mortality or morbidity rates (Freedman, 2004)—is entirely missing from these "aspects" of association. Therefore, when EPA writes that "EPA scientists consider aspects such as strength, consistency, coherence, and biological plausibility of the evidence, and develop causality determinations," they describe a process in which EPA scientists make determinations about causality that cannot be validly derived from the aspects of the data that they are asked to consider. There is no reason to regard the resulting opinions as useful guides to objective, factual truth about causality.

> *"EPA has developed a consistent and transparent basis for integration of scientific evidence and evaluation of the causal nature of air pollution-related health or welfare effects for use in developing ISAs. The framework described below establishes uniform language concerning causality and brings more specificity to the findings."*

COMMENT: EPA's "basis for integration of scientific evidence and evaluation of the causal nature of air pollution-related health or welfare effects for use in developing ISAs" is here asserted to be "consistent and transparent." But it is not. It depends essentially on the subjective judgments (via the "causality determination") of EPA-selected experts, none of whom is necessarily expert in the methodology of valid causal inference and hypothesis-testing.

The opinions of experts are subject to the same pitfalls, such as motivated reasoning, groupthink, confirmation bias, narrow framing, overconfidence (possibly ameliorated by calibration training), and so forth as the opinions of other people (Kahneman, 2011). They are not "transparent" in the sense of using independently observable and verifiable calculations to move from publicly available data to conclusions. Rather, they amount to selected experts using inscrutable judgment processes to render "causality determinations" that EPA then declares to be "consistent and transparent," although other experts (especially those trained in causal analysis) might reach vastly different conclusions.

Indeed, EPA's own experts have sometimes reached vastly different conclusions about health effects of pollutants when offering subjective judgments about causality without fear of contradiction (e.g., Harvard School of Public Health, 2002) and when re-doing their work more carefully to satisfy the rudiments of objective causal analysis, such as comparing the outcomes for relevant treatment and control groups (e.g., HEI, 2013). It is therefore misleading for EPA to characterize as "consistent and transparent" its practice of using expert judgments to obtain "causality determinations" that cannot necessarily be derived by objective data analysis of data (transparency) or replicated by others (consistency) or made consistent with any more objective, data-driven causal analysis.

> *"This standardized language was drawn from sources across the federal government and wider scientific community, especially the National Academy of Sciences (NAS) Institute of Medicine (IOM) document, Improving the Presumptive Disability Decision-Making Process for Veterans (Samet and Bodurow, 2008), a comprehensive report on evaluating causality."*

COMMENT: The Samet and Bodurow (2008) report is not a "comprehensive report on evaluating causality." It does not provide a comprehensive discussion of most methods of valid causal inference (e.g., structural equation modeling, Simon causal ordering, compositional causal modeling, causal graph models, Granger causality tests, quasi-experiment design and analysis, regression discontinuity and instrumental variables approaches, etc.) A comprehensive treatment of evaluating causality was not part of its scope or intent.

> *"This framework… describes the kinds of scientific evidence used in establishing a general causal relationship between exposure and health effects;"*

COMMENT: The term "general causal relationship" is not carefully defined. From context, it appears to mean "an opinion about whether historical statistical associations between exposure and response have certain aspects, listed in Table 1 of the ISA." It is not clear why such "general causal relationships" are emphasized in EPA's risk assessment, as they do not necessarily (or probably) predict how changes in future exposures will change future health effects. As discussed previously, the aspects of association in Table 1 of the ozone ISA are neither necessary nor sufficient to establish causation.

COMMENT: The framework does not describe "the kinds of scientific evidence used in establishing a general causal relationship between exposure and health

effects,"whatever that means. Rather, it presents a framework for inviting experts to opine about whether associations are causal without doing any formal or objective causal analysis needed to reach valid conclusions. It encourages selected experts to substitute their own subjective opinions or preconceptions for objective factual or scientific analyses of causation by asking them to base their opinions on considerations of selected "aspects" statistical associations (e.g., strength and consistency) that do not establish causation and that have no known or proven validity for informing speculations about causality (e.g., Morabia, 2013). The framework neither requires nor uses any formal or objective quantitative causal analyses, causal modeling, or causal hypothesis-testing to check whether the expert's subjective judgments are consistent with data or might be factually correct.

COMMENT: The cited report by Samet and Bodurow (2008) explicitly rejects EPA's claim here that consideration of aspects of statistical associations provide a sound basis for "establishing a general causal relationship between exposure and health effects." To the contrary, their report states that "Because a statistical association between exposure and disease does not prove causation, plausible alternative hypotheses must be eliminated by careful statistical adjustment and/or consideration of all relevant scientific knowledge. Epidemiologic studies that show an association after such adjustment, for example through multiple regression or instrumental variable estimation, and that are reasonably free of bias and further confounding, provide evidence but not proof of causation."[9] Nowhere does this report support EPA's claim that the types of evidence considered by EPA suffice to establish a causal relationship (general or otherwise) between exposure and health effects. Thus, EPA cites as an authority for establishing causation based on expert consideration of associations a source that explicitly states that epidemiologic studies showing associations do not prove causation.

> *"This framework… characterizes the process for integration and evaluation of evidence necessary to reach a conclusion about the existence of a causal relationship;"*

COMMENT: The framework does not "characteriz[e] the process for integration and evaluation of evidence necessary to reach a conclusion about the existence of a causal relationship." For example, it says nothing about how to evaluate evidence for or against propagation of changes in exposure through networks of causal mechanisms to produce changes in health outcomes, although consideration of such propagation of changes are necessary to reach valid conclusions about whether a causal relationship exists (Freedman, 2004).

[9] www.nap.edu/openbook.php?record_id=11908&page=173

"This framework… identifies issues and approaches related to uncertainty;"

COMMENT: We agree that the cited report by Samet and Bodurow (2008) does an excellent job of identifying some key issues and approaches related to uncertainty. Specifically, this report states that:

"The uncertainty about the correct causal model involves uncertainty about whether exposure in fact causes disease at all, about the set of confounders that are associated with exposure and cause disease, about whether there is reverse causation, about what are the correct parametric forms of the relations of the exposure and confounders with outcome, and about whether there are other forms of bias affecting the evidence. One currently used method for making this uncertainty clear is to draw a set of causal graphs, each of which represents a particular causal hypothesis, and then consider evidence insofar as it favors one or more of these hypotheses and related graphs over the others. We explain this approach in more detail in Appendix J.

Uncertainty about the model is not just limited to the qualitative causal structure; however, it also involves uncertainty about the parametric form of the model specified, the variables included, whether or not measurement error is modeled, and so on. When mechanistic knowledge exists, this sort of uncertainty is mitigated. Nevertheless, model uncertainty is perhaps the most important level of uncertainty."[10]

However, EPA's risk assessment and uncertainty analysis do not incorporate or respond to these insights. They do not quantify or characterize model uncertainty about "about whether exposure in fact causes disease at all." They do not draw causal graph models or test causal hypotheses. They do not refute non-causal explanations for E-R and C-R statistical associations. In short, they cite a relatively sophisticated discussion of uncertainty and then do not follow the good advice and useful insights that it offers.

"This framework… provides a framework for classifying and characterizing the weight of evidence in support of a general causal relationship.

[10] www.nap.edu/openbook.php?record_id=11908&page=170

Approaches to assessing the separate and combined lines of evidence (e.g., epidemiologic, controlled human exposure, and animal toxicological studies) have been formulated by a number of regulatory and science agencies, including the IOM of the NAS (Samet and Bodurow, 2008), International Agency for Research on Cancer (IARC, 2006), U.S. EPA (2005), and Centers for Disease Control andPrevention (CDC, 2004). Causal inference criteria have also been described for ecological effects evidence (U.S. EPA, 1998; Fox, 1991). These formalized approaches offer guidance for assessing causality. The frameworks are similar in nature, although adapted to different purposes, and have proven effective in providing a uniform structure and language for causal determinations."

COMMENT: The term "general causal relationship" is not carefully defined, as previously discussed.

COMMENT: Samet and Bodurow and the other references actually do not "provid[e] a framework for classifying and characterizing the weight of evidence in support of a general causal relationship." They provide a framework for classifying and characterizing opinions about some aspects of statistical associations that have no justification (other than long and widespread usage in epidemiology, though not in other fields) for being used to inform expert opinions and speculations about causation (Appendix A; Morabia, 2013).

As stated in the Samet and Bodurow report, "We begin by discussing the problem of integrating the evidence from multiple epidemiologic studies. We then describe a framework for combining epidemiologic and other evidence into a single *quantitative* judgment about the strength of causation. Next we discuss *qualitative* frameworks that have been used by expert committees for categorizing the overall strength of evidence for or against a causal claim. And lastly we propose a qualitative framework for causal reference to be used in the presumptive disability decision-making process."[11]

The quantitative judgment about strength of causation referred to is meta-analysis. This fails when the individual studies being integrated do not provide valid evidence about causation. The "*qualitative* frameworks that have been used by expert committees for categorizing the overall strength of evidence for or against a causal claim" do not identify the *fraction* of a statistical association that is causal (if any). Thus, however appropriate or inappropriate they might be "in the presumptive

[11] www.nap.edu/openbook.php?record_id=11908&page=175

disability decision-making process" that is the explicit focus of the Samet and Bodurow report, they are not appropriate for risk assessment of future health effects – the use that EPA makes of them.

Likewise, the proposed "qualitative framework for causal reference to be used in the presumptive disability decision-making process" may or may not be appropriate for this purpose, but it is certainly not justified for the purpose of predicting how future changes in ozone exposures will change health risks (if at all). EPA has used a *presumptive* framework intended for a different purpose (e.g. retrospective compensation awards), not a *predictive* framework for giving valid predictions to policy makers of the likely health consequences of future changes in exposure levels. The framework on which EPA has relied has not been validated or shown to be useful for the purpose to which they apply it, i.e., predictive risk modeling rather than presumptive, retrospective decision-making about disability awards.

APPENDIX A. Principles of Causal Analysis and Inference for Quantitative Risk Assessment

Some Limitations of Traditional Epidemiological Measures for Causal Inference

Uncertainty about Whether Associations are Causal

Epidemiology has a set of well-developed traditional methods and measures for quantifying associations between observed quantities, especially regression model coefficients and relative risk (RR) ratios (e.g., the ratio of disease rates for exposed and unexposed populations) and quantities derived from them. These include population attributable risks (PARs) and population attributable fractions (PAFs) for the fraction of disease or mortality cases attributable to a specific cause; global burden of disease estimates; etiologic fractions and probability-of-causation calculations; and estimated concentration-response slope factors for exposure-response relations (Rothman et al., 2012). The key idea of all these measures is to observe whether more-exposed people suffer adverse consequences at a higher rate than less-exposed people and, if so, to attribute the excess risks in the more-exposed group to a causal impact of exposure.

Conventional statistical methods for quantifying uncertainty about measures of association, such as confidence intervals and p-values for RR, PAF, or regression coefficients in logistic regression, Cox Proportional Hazards, or other parametric or semi-parametric regression models, are ubiquitously used to show how firmly the data can be used to reject the null hypothesis of independence (no association) between exposures and adverse health responses. In addition, model diagnostics (such as plots of residuals and tests of model assumptions) can reveal whether modeling assumptions appear to be satisfied; more commonly, though less informatively, goodness-of-fit measures are reported to build confidence that the models used do not give conspicuously poor descriptions of the data, at least as far as the goodness-of-fit test can determine.

The main limitation of these techniques is that they only address associations, rather than causation, and hence they cannot quantify the fraction or number of illnesses or mortalities per year that would be prevented by reducing or eliminating specific exposures. Unfortunately, as many methodologists have noted and warned against, PAF and probability of causation, as well as regression coefficients, are commonly misinterpreted as doing precisely this (e.g., Rothman et al., 2012). Large epidemiological initiatives, such as the World Health Organization's Global Burden

of Disease studies, make heavy use of association-based methods that are treated as if they were causal relations. This has become a very common mistake in contemporary epidemiological practice. It undermines the validity, credibility, and practical value of many (some have argued most) causal claims now being published using traditional epidemiological methods (Nuzzo, 2014; Sarewitz, 2012; Lehrer, 2012). To what extent associations correspond to stable causal laws that can be relied on to predict future consequences of policy actions is beyond the power of these traditional epidemiological measures to say; doing so requires different techniques (Rothman et al., 2012).

Example: Exposure-Response Relations Depend on Modeling Choices

A 2014 article in Science (Dominic et al., 2014) noted that "There is a growing consensus in economics, political science, statistics, and other fields that the associational or regression approach to inferring causal relations – on the basis of adjustment with observable confounders – is unreliable in many settings." To illustrate this point, the authors cite estimates of the effects of total suspended particulates (TSPs) on mortality rates of adults over 50, in which significantly positive associations (regression coefficients) are reported in some regression models that did not adjust for confounders such as age and sex, but significantly negative associations are reported in other regression models that did adjust for confounders by including them as explanatory variables.

The authors note that the sign, as well as the magnitude, of reported exposure concentration-response (C-R) relations depends on details of modeling choices about which variables to include as explanatory variables in the regression models. Thus, the quantitative results of risk assessments presented to policy makers as showing the expected reductions in mortality risk per unit decrease in pollution concentrations actually reflect specific modeling choices, rather than reliable causal relations that usefully predict how (or whether) reductions in exposure concentrations would reduce risks.

A distinction from econometrics between structural equations and reduced-form equations (Klein, 1974) is helpful in understanding how different epidemiologists can estimate exposure concentration-response regression coefficients with opposite signs from the same data. The following highly simplified hypothetical example illustrates the key idea. Suppose that cumulative exposure to a chemical increases in proportion to age, and that the risk of disease (the age-specific hazard rate, i.e., the probability of being first diagnosed with a disease or adverse medical condition at each age) also increases with age. Finally, suppose that the effect of exposure at any age is to decrease risk.

These causal relations are shown via the following two *structural equations*, which have the explicit causal interpretation that an increase in the variable on the right side causes an increase in the variable on the left side to make the equation hold (e.g., increasing age increases cumulative exposure and disease risk, but increasing exposure decreases risk at any age):

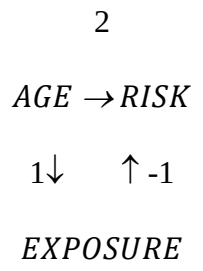
SEM Equations

$$EXPOSURE = AGE$$

$$RISK = 2 * AGE - EXPOSURE$$

These two structural equations together constitute a simple *structural equation model* (SEM) that can be diagrammed as in Figure 1.

Figure 1. SEM Causal Graph Model



In this diagram, each variable depends causally only on the variables that point into it, as revealed by the SEM equations. The weights on the arrows (the coefficients in the SEM equations) show how the average value of the variable at the arrow's head will change if the variable at its tail is increased by one unit. By contrast to such causal SEM models, what is called a *reduced-form model* is obtained by using the first SEM equation, EXPOSURE = AGE,to substitute EXPOSURE for AGE in the second SEM equation, RISK = 2*AGE – EXPOSURE, to obtain the following:

Reduced-form Equation

$$RISK = EXPOSURE$$

This reduced-form model is a valid *descriptive statistical* model: it reveals that in communities with higher exposure levels, risk should be expected to be greater. But it is not a valid *causal* model: a prediction that reducing exposure would cause a reduction in risk would be mistaken. The reduced-form equation is not a structural equation, so it cannot be used to predict correctly how changing the right side would cause the left side to change. The coefficient of EXPOSURE in the linear regression model relating exposure to risk is +1 in the reduced-form model, but is -1 in the SEM

model, showing how different investigators may reach opposite conclusions about the sign of "the" exposure-response coefficient based on whether or not they condition on age (or, equivalently, on whether they use structural or reduced-form regression equations).

In current epidemiological practice, the distinction between structural and reduced-form equations is often not clearly drawn. Regression coefficients of various signs and magnitudes, as well as various measures of association based on relative risk ratios, are all presented to policy makers as if they had valid causal interpretations and therefore important implications for risk management policy-making. In air pollution health effects epidemiology, for example, it is standard practice to present regression coefficients as expected reductions in elderly mortality rates (or expected increases in life span) per unit reduction in air pollution concentrations (EPA, 2011; Fann, 2012), thereby conflating associations between historical levels (e.g., pollutant levels and mortality rates both tend to be higher on cold winter days than during the rest of the year, and both have declined in recent decades) with a causal, predictive relation that implies that future reductions in pollution would cause further future reductions in elderly mortality rates. Since such association-based studies are often unreliable indicators of causality (Dominici et al., 2014) or simply irrelevant for determining causality (Rothman et al., 2012), policy makers who wish to use reliable causal relations to inform policy decisions by considering their likely consequences must seek elsewhere.

These limitations of association-based methods have been well discussed among methodological specialists for decades (Rothman et al., 2012). Key lessons, such as that the same data set can yield either a statistically significant positive exposure-response regression coefficient or a statistically significant negative exposure-response regression coefficient, depending on the detailed modeling choices made by the investigators, are becoming increasingly appreciated by practitioners (Dominici et al., 2014). They illustrate an important type of uncertainty that arises in epidemiology, but that is less familiar in many other applied statistical settings: *uncertainty about the interpretation of regression coefficients* (or other association-based measures such as RR, PAF, etc.) as indicating causal relations vs. confounded associations vs. some of each.

This type of uncertainty cannot be addressed by presenting conventional statistical uncertainty measures such as confidence intervals, p-values, regression diagnostics, sensitivity analyses, or goodness-of-fit statistics, since the uncertainty is not about how well a model fits data, or about the estimated parameters of the model. Rather, it is about the extent to which the model is descriptive of the past vs. predictive of

different futures caused by different choices. Although this is not an uncertainty to which conventional statistical tests apply, it is crucial for the practical purpose of making model-informed risk management decisions. Policy decisions successfully increase the probabilities of desired outcomes and decrease the frequencies of undesired ones only to the extent that they act causally on drivers of the outcomes, and not based on how well the models used describe past associations.

One way to try to bridge the gap between association and causation is to ask selected experts what they think about whether or to what extent associations might be causal. However, research on the performance of expert judgments has called into question the reliability of expert judgments, specifically including judgments about causation. Such judgments typically reflect qualitative "weight of evidence" (WoE) considerations about the strength, consistency (do multiple independent researchers find the claimed associations?), specificity, coherence (are associations of exposure with multiple health endpoints mutually consistent with each other and with the hypothesis of causality?) temporality (do hypothesized causes precede their hypothesized effects?), gradient (are larger exposures associated with larger risks?), and biological plausibility of statistical associations and the quality of the data sources and studies supporting them.

In practice, WoE-based judgments face challenges that may render them unreliable (e.g., Goodman et al., 2013; Rhomberg et al., 2013). One difficulty is that a strong confounder (such as age in Figure 1) with delayed effects can create strong, consistent, specific, coherent, temporal associations between exposure and risk of an adverse response, with a clear gradient associating larger risks with larger exposures, all without thereby providing any evidence that exposure actually causes increased risk. Showing that an association is strong, for example, does not address whether it is causal (although many WoE systems assume that the former supports the latter).

Similarly, showing that many different investigators find the same association does not shed any light on whether the association they have found is causal, or whether it results from biases and confounders that are common across studies. Conflating causal and associational concepts, such as evidence for the strength of an association and evidence for causality of the association, makes assessments of causality in epidemiology idiosyncratic and of uncertain trustworthiness (Morabia, 2013; Höfler, 2005) compared to methods used in other fields. Most experts in epidemiology have been trained to treat various aspects of association as evidence for causation, even though they are not, and this undermines the trustworthiness of expert judgments about causation based on WoE considerations (Morabia, 2013).

In addition, experts are sometimes asked to judge the *probability* that an association is causal (e.g., EPA, 2006). This makes little sense. It neglects the fact that an association may be partly causal and partly not (e.g., due to confounding or modeling biases or coincident historical trends). For example, if exposure does increase risk, but is also confounded by age, then asking for the probability that the regression coefficient relating exposure to risk is causal overlooks the realistic possibility that it reflects both a causal component and confounding, so that the probability that it is partly causal might be 1 and the probability that it is completely causal might be 0. (A more useful question to pose to experts might be what *fraction* of the association is causal.)

Common non-causal sources of statistical associations include model selection and multiple testing biases, model specification errors, unmodeled errors in explanatory variables in multivariate models, biases due to data selection and coding (e.g., dichotomizing or categorizing continuous variables such as age, which can lead to residual confounding), and coincident historical trends (which can induce statistically significant-appearing associations between statistically independent random walks – a phenomenon sometimes dubbed spurious regression) (Rothman et al., 2012; Cox et al., 2013).

Finally, qualitative subjective judgments and ratings used in many WoE systems are subject to well-documented psychological biases. These include confirmation bias (seeing what we expect to see and discounting or ignoring what challenges our preconceptions); motivated reasoning (finding what it benefits us to find and believing what it pays us to believe); and over-confidence (not sufficiently doubting, questioning, or testing our own beliefs, and hence not seeking potentially disconfirming information that might cause us to revise what we believe) (Kahneman, 2011; Sarewitz, 2012).

Despite these limitations, there is much of potential value in several WoE considerations (Rhomberg et al., 2013), especially consistency, specificity, and temporality of associations, if they can be freed from the context of qualitative subjective judgments and used as part of a more objective, quantitative, data-driven approach to inferring probable causation. This possibility is discussed in the following sections.

Some Limitations of Quantitative Risk Assessment (QRA) for Health Effects of Exposures

In the United States, regulatory risk assessments of the health effects attributed to exposures are often conducted according to the following six-step process, which corresponds closely to the process followed by EPA in its ozone risk assessment:

1. Establish that there is a statistically significant positive association between one or more measures of exposure (e.g., cumulative exposure) and one or more measures of risk (e.g., age-specific mortality risk), e.g., using meta-analyses of multiple individual studies that report such associations.

2. Estimate the quantitative magnitude of the association between exposure and risk, e.g., via a significant positive regression coefficient.

3. Judge whether the association is (probably) causal, based on expert evaluation of weight-of-evidence (WoE) considerations such as the strength, consistency, gradient, temporality, biological plausibility, etc. of the association.

4. Predict the reduction in risk for a proposed reduction in exposure, e.g., by applying the regression coefficient using the formula $\Delta Risk = K * \Delta Exposure$, where $\Delta Risk$ is the estimated reduction in risk caused by a reduction in exposure ($\Delta Exposure$), and K is the regression coefficient ("potency") linking the exposure and risk metrics.

5. Show that alternative models (within one or more classes of models) do not provide a significantly better fit to the data than the model (typically, a best-fitting model in some class of models) selected to project risk reductions caused by exposure reductions. Although many hypotheses about model form might be considered for purposes of sensitivity analysis, using a best-fitting model on which to base conclusions is standard practice.

6. Use uncertainty intervals (e.g., 95% confidence intervals) to quantify uncertainties about the likely reduction in risk caused by a proposed reduction in exposures. Part of responsible risk assessment and communication is to characterize uncertainties about the conclusions. A 95% confidence interval for risk reduction indicates a range, due to sampling variability, of risk reductions that might be expected (with 95% statistical confidence) to include the actual reduction in risk in response to a proposed reduction in exposures, if the selected model and underlying assumptions are all correct. Other sources of potential errors, such as errors in exposure estimates, uncertainty about model form,

possible confounders (e.g., smoking), possible selection bias in the data, etc. should also be discussed.

Despite their widespread acceptance, these six steps produce quantitative risk assessment (QRA) numbers that have no known or necessary relation to the risks actually caused by exposure, largely because they do not adequately model causation and uncertainties. Key challenges at each step are as follows.

- In Step 1, a statistically positive ER association can always be created by a combination of large sample size and incorrect model specifications, as well as by other possible explanations already mentioned such as coincident historical trends, data selection biases, model selection biases, and confounding. For example, fitting the line $Risk = K * Exposure$ to any data set consisting of positive values of *Exposure* and corresponding *Risk* estimates will always produce a positive estimate of the potency parameter *K*, even if *Exposure* has no effect on *Risk* (or has a negative effect). If the sample size is large enough, this positive association will always be statistically significantly greater than zero.

  Other, less obviously biased, modeling choices, specification errors, and selection biases can likewise generate statistically "significant" positive associations under the implicit assumption that the selected model describes the data approximately correctly, whether or not there is any true association in the data (e.g., discoverable using "model-free" non-parametric tests). As emphasized by Dominici et al. (2014), both significant positive and significant negative ER relations, depending on modeling choices. Therefore, unless model diagnostics (e.g., plots of residuals) are presented (Greenland, 1989; Maldonado and Greenland, 1993) that show that a model provides an appropriate description of the data to which it is applied, there is no way for readers of reported results to know whether a reported significant positive association based on regression modeling or other models has resulted from a real pattern in the data, or from fitting a mis-specified model to the data, or perhaps from some other reason such as residual confounding by a categorized continuous confounder. Yet, too often, QRAs and the key studies on which they are based present no model diagnostics, and instead offer only sensitivity analyses and goodness-of-fit measures that do not sharply test modeling assumptions and that lack the power to reveal important model specification errors.

- Step 2, quantification of regression of coefficients or other measures of association, is irrelevant for making causal predictions if the association in Step 1 is not causal. (See discussion of Figure 2.)

- Step 3, judging whether an association is causal, invites experts to substitute their own beliefs, opinions, or judgments for rigorous causal analysis. There is no reason to believe that the resulting opinions will necessarily or usually be useful guides to the truth. Experts (like other people) are notoriously poor at correctly judging causation (Kahneman, 2011). Moreover, the associations that they are asked to form causal judgments about typically do not contain the essential information needed to draw valid inferences about causation, so that there is no legitimate basis for offering judgments about whether they are causal. That experts are able and willing to offer opinions on many questions, including nonsensical ones (such as how old is the current king of France, or what is the probability that an association is causal, the first of which mistakenly assumes that there is a current king of France and the second of which mistakenly assumes that an association is either entirely causal or not), does not make their judgments useful or trustworthy.

  Although expert elicitation procedures today typically train experts on calibration and try to reduce over-confidence (Hora, 2007), they usually do not provide training on valid causal inference. As a result, experts often make confident causal judgments (e.g., Harvard School of Public Health, 2002) that are statistically naïve (e.g., ignoring comparisons to control groups) and that subsequently prove to be blatantly false (e.g., HEI, 2013). This holds not only for health effects of exposures, but also in other application domains, such as expert opinions about political, military, and world events.

  When expert judgments and predictions have been evaluated in hindsight, after the truth is learned, they turn out to be, on average, slightly worse than random guesses, with the more famous experts tending to be less accurate (but to have more interesting and plausible-sounding rationales for their predictions) (Kahneman, 2011, citing prior work by Tetlock). This puts a considerable burden of proof on risk assessors who use expert judgments in place of more objective, data-driven methods to show that expert judgments about causation are trustworthy.

- Step 4 is not valid: the slope of an exposure-response (ER) regression relation does *not* necessarily predict how (or whether) future responses would change if future exposure were changed, as discussed further for the example in Figure 2. Valid predictions about reductions in health risks caused by reductions in exposure require a different type of study design and a different type of analysis – *causal* study design and *causal* analysis – than the association-based approach in steps 1-6.

- Step 5, selection of a "best" model, understates model uncertainty. It leads to predictions and confidence intervals that do not account for the fact that even the "best" model (based on goodness-of-fit or other criteria) is almost certain to be wrong when there is a lot of model uncertainty. Model ensemble methods, such as Bayesian model averaging (BMA), consider the results from many (e.g., thousands or tens of thousands) of plausible models to avoid making model-based predictions overly dependent on the selection of one or a few models that are probably wrong. They typically lead to more accurate predictions, wider confidence intervals, and fewer false positives than the traditional approach of selecting a single "best" model (Hoeting et al., 1999). They avoid the fundamental error of treating one or a few finally selected models as if they were known to be correct, for purposes of calculating confidence intervals and significance levels.

- Step 6, presentation of uncertainty intervals around best estimates for the risk caused by exposure, or the risk that can be removed by reducing exposure, is misleading highly misleading in the usual case where the presented interval ignores model uncertainty and where no discrete probability (a positive fraction greater than zero) is presented for the discrete possibility that reducing exposure would not reduce risk because there is no causal relation between them.

  As just discussed, ignoring model uncertainty by selecting a single model biases conclusions toward false-positive findings (since confidence intervals that would include "no effect" if model uncertainty were accounted for are incorrectly narrowed) (e.g., Piegorsch, 2013; Swartz et al., 2001; Viallefont et al., 2001). Methods for overcoming this bias by including multiple plausible models in the calculation of results are now widely available (e.g., Piegorsch, 2013 and references therein) and should be used unless the data-generating process is understood well enough so that a single model that describes it well can be identified.

To illustrate the crucial difference between association-based inference and causal analysis and inference for exposure-response relations, imagine a population in which people with elevated risk of heart attacks tend to consume significantly more baby aspirin than people without elevated risks of heart attack. In that scenario, association-based risk assessment following steps 1-6 (with a mistaken assumption or belief that the association is causal) would lead to a confident statistical prediction that reducing the level of baby aspirin consumption would reduce heart attack risk. However, there is in fact no valid basis for such a conclusion: the association between *levels* of baby aspirin consumption and heart attack risk contains no relevant

information for predicting how *changes* in consumption would change heart attack risk.

As another example, consider the linear association between per-capita phone ownership and coronary heart disease (CHD) mortality risk shown in Figure 2.

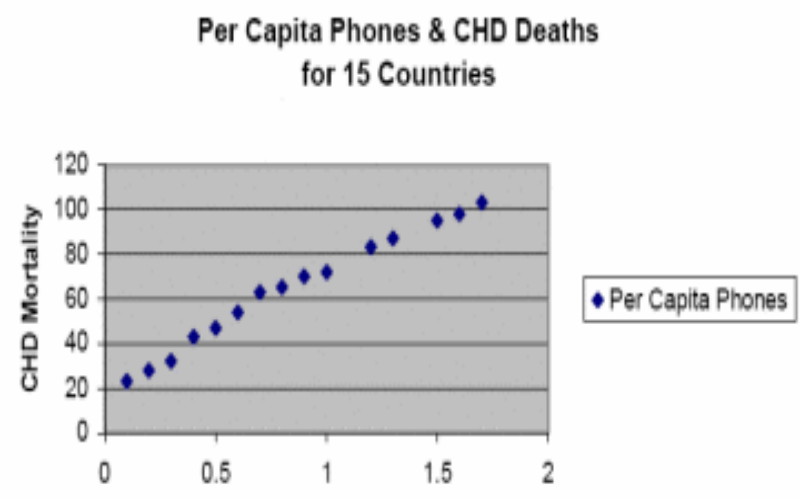


Figure 2. A strong, consistent, linear, no-threshold relation between exposure and mortality risk does not imply that reducing exposure would reduce risk.

Source: http://ocw.tufts.edu/Content/1/readings/193106

An association-based risk assessment might conclude on the basis of such evidence of a strong, linear, no threshold exposure-response (ER) relation that reducing exposure to phones would reduce CHD mortality risk. The slope of the line in Figure 2 would then give a quantitative estimate of the incremental reduction in mortality per unit reduction in per-capita phone ownership. But such causal interpretations and predictions are not valid and cannot be justified by associational data.

By contrast, statistical associations in historical data can be useful for predicting *observed* values of one variable from *observed* values of another. Indeed, predicting values (especially, past values) of one variable from past values of another is precisely what regression models are useful for. But most QRAs purport to answer a very different type of question, about future *changes* in health that would be caused by proposed future *changes* in exposure levels – not about historically *observed* levels of health based on historically *observed* levels of exposure. It is this type of causal inference which cannot legitimately be based on associations.

The statistical relation in Figure 2 could be used correctly to predict the *level* of CHD mortality risk in a population based on the *level* of phone ownership per capita. (A possible mechanism for the association might be that both are associated with a third factor or variable, such as stress levels or dietary factors, although no such explanation need be known to make an empirical association useful for predicting the level of one variable from the level of another, correlated variable.) This is a perfectly good use of statistical inference, but it implies nothing about how changing one variable would change the other.

The key point illustrated by such examples is not simply the truism that association is not causation. Rather, the key point is that the information contained in associations between historical levels of exposure and levels of adverse response is not what is needed to draw valid causal conclusions about how future changes in exposures will change future levels of risk. An implication is that expert judgments about causation based on such associations are not well founded: the associations do not carry the information needed to draw sound inferences about causation, and so neither do expert opinions or interpretations derived from them. What is required instead is information on how changes in inputs (such as exposures) propagate through causal pathways to produce changes in outputs (such as mortality rates). Without such information on propagation of changes, statistical associations are powerless to give valid causal predictions, including predicting whether reducing exposures will reduce risks, and, if so, by how much (Freedman, 2004).

A second implication is that a positive, increasing ER relation such as that in Figure 2 gives no insight into what the causal relation between exposure and response (if there is one) might look like. The causal ER relation could have a threshold, or be J-shaped (flat or negative slope at very low exposures, positive slope at higher exposures), or have a flat or even a negative slope throughout, and yet these characteristics would be obscured by the positive statistical association, which does not isolate the causal component (if any) of the empirical ER association.

This crucial methodological point—that statistical associations do not in general convey information useful for making valid causal predictions—has been well understood for over a decade by statisticians and epidemiologists specializing in technical methods for causal analysis (e.g., Greenland and Brumback, 2002; Freedman, 2004). This understanding strengthens warnings that associations do not *prove* causation (e.g., Samet and Bodurow, 2008) by showing that they need not be *informative* about causation. It is only slowly percolating through the larger epidemiological and risk analysis communities, however. Peer-reviewed published papers and authoritative reports, including those relied on in many regulatory QRAs,

still make the fundamental mistake of reinterpreting empirical exposure-response (ER) associations between historical levels of exposure and response *as if* they were causal relations useful for predicting how changes in exposures would change responses. This confusion is unnecessary: appropriate technical methods for causal analysis and modeling are now well developed, and can be applied to the same kinds of cross-sectional and longitudinal data collected for association-based studies. Table 1 summarizes some of the most useful study designs and methods for valid causal analysis and modeling of causal exposure-response relations.

Particularly harmful is that many authoritative sources propagate the delusion, peculiar to epidemiology (Morabia, 2013), that somehow associations alone can suffice to establish causation, if only they are qualified with enough laudatory adjectives (the association is strong, consistent, plausible, etc.) This is a mistake. For example, Samet and Bodurow (2008) present a table that begins as follows:

TABLE 8-2 IOM Categorization from the Executive Summary of *Gulf War and Health, Volume 1: Depleted Uranium, Pyridostigmine Bromide, Sarin, Vaccines*

Sufficient Evidence of a Causal Relationship	Evidence is sufficient to conclude that a causal relationship exists between the exposure to a specific agent and a health outcome in humans. The evidence fulfills the criteria for sufficient evidence of an association (below) and satisfies several of the criteria used to assess causality: strength of association, dose-response relationship, consistency of association, temporal relationship, specificity of association, and biological plausibility.
Sufficient Evidence of an Association	Evidence is sufficient to conclude that there is a positive association. That is, a positive association has been observed between an exposure to a specific agent and a health outcome in human studies in which chance, bias, and confounding could be ruled out with reasonable confidence.

Source: www.nap.edu/openbook.php?record_id=11908&page=186

It should be clear that the characterization of "Sufficient evidence of a causal relationship" offered here is wrong: as suggested by counter-examples such as Figure 2 (and recognizing that "biological plausibility" is often in the eye of the beholder, with more associations perhaps seeming plausible to those who know least), it is easy for statistical associations to satisfy all of these conditions, and yet not be causal.

Table 1. Some formal methods for modeling and testing causal hypotheses

Method and References	Basic Idea	Appropriate study design
Conditional independence tests (Freedman, 2004, Friedman and Goldszmidt, 1998)	Is hypothesized effect (e.g., lung cancer) statistically independent of hypothesized cause (e.g., exposure to crystalline silica), given values of other variables (e.g., education and income)? If so, this undermines causal interpretation.	Cross-sectional data Can also be applied to multi-period data (e.g., in dynamic Bayesian networks)
Panel data analysis (Angrist and Pischke, 2009, Stebbings, 1976)	Are changes in exposures followed by changes in the effects that they are hypothesized to help cause? If not, this undermines causal interpretation; if so, this strengthens causal interpretation. Example: Are changes in crystalline silica exposure levels in different quarries followed (but not preceded) by corresponding changes in respiratory mortality rates?	Panel data study: Collect a sequence of observations on same subjects or units over time
Granger causality test (Eichler and Didelez, 2010)	Does the history of the hypothesized cause improve ability to predict the future of the hypothesized effect? If so, this strengthens causal interpretation; otherwise, it undermines causal interpretation. Example: Can lung cancer mortality rates in different occupational groups be predicted better from time series histories of crystalline silica levels and mortality rates than from the time series history of mortality rates alone?	Time series data on hypothesized causes and effects
Quasi-experimental design and analysis (Campbell and Stanley, 1966)	Can control groups and other comparisons refute alternative (non-causal) explanations for observed associations between hypothesized causes and effects? For example, can coincident trends and regression to the mean be refuted as possible explanations? If so, this strengthens causal interpretation.	Longitudinal observational data on subjects exposed and not exposed to interventions that change the hypothesized cause(s) of effects.
Intervention analysis, change point analysis (Helfenstein, 1991; Gilmour et al., 2006)	Does the best-fitting model of the observed data change significantly at or following the time of an intervention? If so, this strengthens causal interpretation.	Time series observations on hypothesized effects, and knowledge of

	Do the quantitative changes in hypothesized causes predict and explain the subsequently observed quantitative changes in hypothesized effects? If so, this strengthens causal interpretation. Example: Did lung disease mortality rates fall significantly faster or sooner in workplaces that reduced exposures more or earlier than in workplaces that did not?	timing of intervention(s) Quantitative time series data for hypothesized causes and effects
Counterfactual and potential outcome models (Robins et al. 2000; Moore et al., 2012)	Do exposed individuals have significantly different response probabilities than they would have had if they had not been exposed? Example: Do workers have lower mortality risk after historical exposure reductions than they would have had otherwise?	Cross-sectional and/or longitudinal data, with selection biases and feedback among variables allowed
Causal network models of change propagation (Hack et al., 2010, Dash and Druzdzel, 2008)	Do changes in exposures (or other causes) create a cascade of changes through a network of causal mechanisms (represented by equations), resulting in changes in the effect variables?	Observations of variables in a dynamic system out of equilibrium
Negative controls (for exposures or for effects) (Lipsitch et al., 2010)	Do exposures predict health effects better than they predict effects that cannot be caused by exposures (e.g., reductions in traumatic injuries)?	Observational studies

Although they are not adequate for valid causal inference or QRA, statistical and epidemiological methods based on associations (e.g., relative risks, odds ratios, logistic regression of Cox proportional hazards coefficients, etc.) are far from useless. Although they usually cannot deliver valid causal predictions about the effects on responses caused by changes in exposures—the main requirement for sound risk assessment and risk management—they can, when carefully implemented, support valid inferences about the expected past historical values (or conditional distributions) of some quantities, such as mortality rates, given the measured values of other historical quantities, such as exposure levels. The technology for drawing statistical inferences about the likely values of some variables, given the values of others (e.g., based on maximum-likelihood, best-fit, or conditional probability distributions or expected values), is now extremely well developed and sophisticated, but it should not be confused with the quite different technology of causal analysis and modeling needed for purposes of valid QRA (e.g., Table 1).

Association-Based QRAs commonly produce false-positive results.

In practice, QRAs based on some or all of steps 1-6, routinely produce a large excess of false-positive errors, both in published results and in confident public assertions about the health benefits expected from various interventions. That is, they predict beneficial effects from interventions that turn out to reflect motivated reasoning, confirmation bias and other biases by investigators, and that do not hold in reality (e.g., Lehrer, 2012, Sarewitz, 2012, Ottenbacher, 1998; Imberger et al., 2011). Scientists with subject matter expertise in health effects epidemiology for specific chemicals are not necessarily or usually also experts in causal analysis and valid causal interpretation of data, and their causal conclusions are often mistaken, with a pronounced bias toward declaring and publishing findings of "significant" effects where none actually exists (false positives).

This has led some commentators to worry that "science is failing us," due largely to widely publicized but false beliefs about causation (Lehrer, 2012); and that, in recent times, "Most published research findings are wrong" (Ioannadis, 2005), with the most sensational and publicized claims being most likely to be wrong. These important policy-relevant limitations of association-based risk assessments and epidemiological studies have been increasingly well recognized by specialists in recent decades (e.g., Ottenbacher, 1998), but have only recently started to attract much attention in the mainstream press (e.g., Lehrer, 2012; *The Economist*, October, 2013). To overcome them, it is probably crucial to pivot from reliance on expert judgments about causation to greater reliance on more objective methods, such as those in Table 1.

Association-Based QRA does not answer critical questions for risk management.

The numbers produced by the regulatory risk assessment framework using steps 1-6 also fail to address key questions that should be answered in informing any responsible risk management decision or regulatory action. Among them are the following.

- *What fraction of the association-based ER relations is causal?* As already discussed, ER associations are unlikely to be entirely causal, since some fraction of association probably results from coincident historical trends (e.g., both pollution levels and disease rates have declined in recent decades, apart from any causal relation between them), confounding (e.g., both pollution levels and elderly mortality rates may be elevated on extremely hot summer days and extremely cold winter days, apart from any causal relation between them), and

biases (e.g., excess false positives in published results, as just discussed). It is therefore important to quantify how large is the causal fraction, and how sure we can be about the answer. Steps 1-6 leave this unaddressed, usually asking experts instead (in Step 4) to opine about whether associations are causal.

- What is the probability that the health benefits caused by a proposed reduction in exposures will be less than the lowest estimated values? Focusing on "best-fit" models (often within a few parametric classes, none of which necessarily describes the true but unknown ER relation) does not answer the key question of how wrong the predictions are likely to be. Might the true benefits be less than 10% as large as the smallest estimated ones? How likely is this? How likely is it that the true benefits are zero, or less than 1% as large as those estimated?

  There is no way to answer such decision-relevant questions based solely on modeling assumptions and choices having uncertain validity, along with confidence intervals calculated using those assumptions. (By contrast, a well-executed QRA using the methods in Table 1 would provide estimates or bounds for the posterior probability of different effects sizes caused by a proposed reduction in exposure.) Too many studies are silent about the probability that their best predictions are mistaken, and about the potential size of the mistakes. Yet, such information is crucial for rational risk management (e.g., working within the expected utility model of rational choice among alternative actions with uncertain consequences).

- What is the probability that reducing exposures below the current permissible exposure level will not yield any incremental human health benefits?

- Is the totality of available evidence more consistent with the null hypothesis that future reductions in exposures will not reduce risk, or with the alternative hypothesis that they will? This is similar to the previous question, but focuses on comparing the evidence for each of the two disjoint hypotheses, rather than on assessing their probabilities. (Technically, this question might be addressed via a likelihood ratio calculation rather than via a Bayesian posterior probability calculation.)

  QRAs often focus primarily on building a case for the alternative hypothesis, and on (mis)using associational data to develop estimates of the change in health risk that would be caused by a further reduction in exposures if the association between them were entirely causal and were correctly described by the QRA's assumed models. Too often, they neglect to evaluate whether the same data are

equally or more consistent with the null hypothesis that further reductions in exposure will have no effect on human health, and that ER associations are due entirely to non-causal explanations such as data selection biases, modelling choices, coincidental but not causal historical trends in exposure and health effects, and so forth.

In summary, steps 1-6 do not answer the questions that a sound QRA should address: what are the probable consequences of alternative actions, such as making or not making a proposed reduction in exposure levels? Moreover, the numbers that they produce are based on historical associations that usually have no known predictive value and no basis for being interpreted as causal. Reinterpreting them as if they were derived from sound causal analysis and modeling is unwarranted. Using them to project reductions in future morbidity and mortality rates that would be caused by proposed reductions in exposure levels is not justified in the absence of sound causal analysis and modeling.

Thus, even if steps 1-6 are executed flawlessly, the resulting numbers do not support valid causal interpretations and predictions. As indicated in Table 1, many methods are available that address the challenges of quantifying causal relations and uncertainties about them without invoking expert judgments about the causal significance of associations. These are examined next.

Valid Methods for Establishing Probable Causation

Over the past century, the following main ideas have been developed, largely outside epidemiology, to determine whether available knowledge and data warrant an inference that exposure probably causes adverse health effects.

Causes precede their effects

If significant changes in exposures always precede and help to explain and predict subsequent significant changes in health effects, this is consistent with the hypothesis that the former cause the latter. To formally test whether this is the case, *change-point analysis* (CPA) algorithms estimate the times of changes in effects time series (e.g., http://surveillance.r-forge.r-project.org/; James and Matteson, 2014). These times can then be compared to the times at which exposures changed (e.g., due to passage of a regulation, opening or closing of a pollution source, etc.) to determine whether changes in exposures are followed by changes in effects.

Similarly, *intervention analysis* algorithms (also called interrupted time series analysis algorithms) test whether effects time series change following changes in exposures that occur at known times. Finally, *Granger causality tests* provide formal

quantitative tests of the hypothesis that the future of the health effects time series can be predicted better using the exposure time series than it can be predicted without using the exposure time series. If this is not the case, then the exposure-response histories provide no evidence that exposure is a (Granger) cause of the effects time series, no matter how strong, consistent, etc. the association between their levels over time may be.

Causes are informative about their effects

If exposure is a cause of increased disease risk, then measures of exposure and response (i.e., disease risk) should provide mutual information about each other that allows either to be predicted from the other (i.e., the conditional probability distribution for either one varies with the value of the other). Moreover, it should not be the case that the mutual information between exposure and response can be eliminated by conditioning on the values of other variables, such as confounders, if exposure is a cause of the response. This is the basis for using formal statistical *conditional independence tests* as tests of causal hypotheses: An effect should never be conditionally independent of its direct causes, given the values of other variables.

Changes in causes produce changes in effects via causal mechanisms

Perhaps the most useful and compelling valid evidence of causation (with the possible exception of well-conducted randomized control trials) consists of showing that changes in exposures propagate through a network of validated law-like equations or mechanisms to produce predictable changes in effects (e.g., health responses). For example, showing that measured changes in ambient levels of pollution produce consistent corresponding changes in lung inflammation markers, recruitment rates of alveolar macrophages and activated neutrophils to the inflamed lung, levels of enzymes released by these cell populations that degrade the alveolar wall, and resulting rates of lung tissue loss and scarring and onset of inflammation-mediated diseases, would provide compelling evidence of a causal relation between changes in exposures and changes in disease rates.

Structural equation models (SEMs) in which changes in right-hand side variables cause adjustments of left-hand side variables to restore equality provide one way of describing mechanisms for situations where the precise time course of the adjustment process is not of interest. Differential equation models, in which changes in flow rates lead to changes in the equilibrium levels of variables for different compartments provide another, complementary way to describe mechanisms when the time course of adjustment is of interest. Figure 3 presents a high-level view of the structure of a simulation model for cardiovascular disease (CVD) outcomes; these

provide still a third way to describe and model the propagation of changes through causal networks.

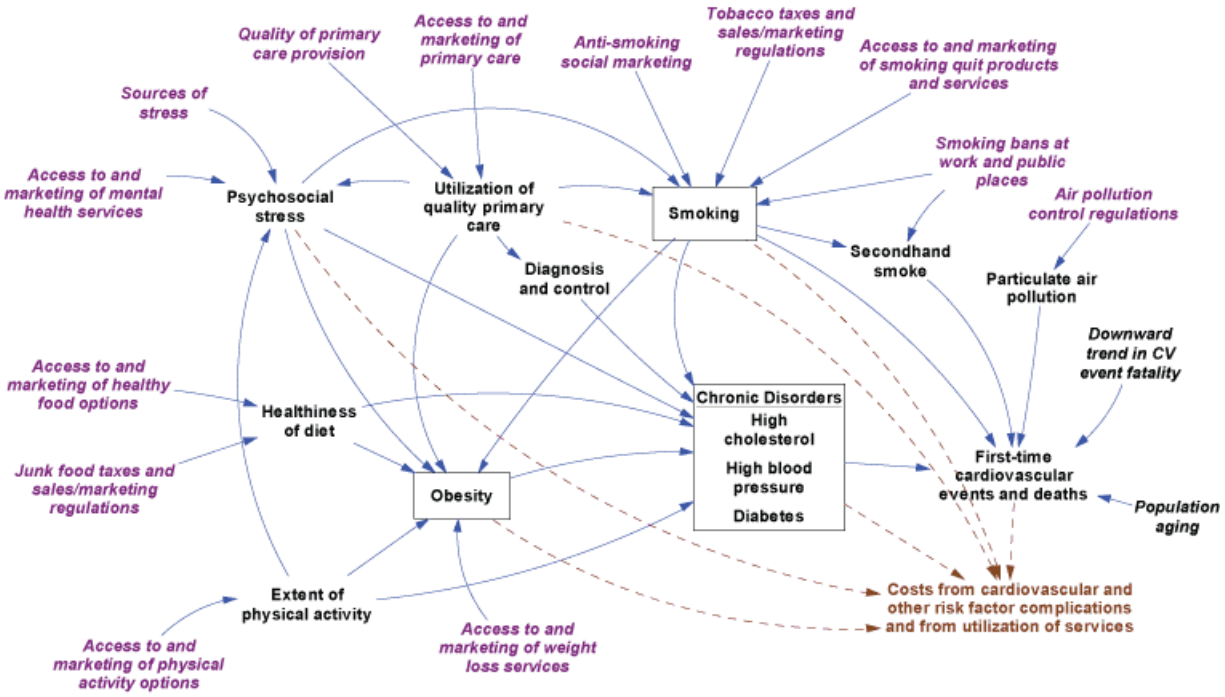


Figure 3. Simulation model for cardiovascular disease (CVD) outcomes. This diagram depicts major health conditions related to CVD and their causes. Boxes identify risk factor prevalence rates modeled as dynamic stocks. The population flows associated with these stocks — including people entering the adult population, entering the next age category, immigration, risk factor incidence, recovery, cardiovascular event survival, and death — are not shown.

Key: Blue solid arrows: causal linkages affecting risk factors and cardiovascular events and deaths.
Brown dashed arrows: influences on costs.
Purple italics: factors amenable to direct intervention.
Black italics (population aging, cardiovascular event fatality): other specified trends.
Black nonitalics: all other variables, affected by italicized variables and by each other

Source: Homer J, Milstein B, Wile K, Trogdon J, Huang P, Labarthe D, et al. Simulating and evaluating local interventions to improve cardiovascular health. Prev Chronic Dis 2010;7(1):A18. http://www.cdc.gov/pcd/issues/2010/jan/08_0231.htm. Accessed 3-11-15

For most of the past century, methods such as *path analysis* (based on SEMs) have been used to explicate causal networks of mechanisms and to provide formal tests for

hypothesized causal structures. These methods exploit the fact that, if all effects of variables on each other can be well approximated by linear regression models, then correlations between variables should be strongest when the variables are closer to each other along a causal chain than when they are more remote.

Moreover, the effect of a change in an ancestor variable on the value of a remote descendent (several nodes away along one or more causal paths) should be decomposable into the effects of the change in the ancestor variable on any intermediate variables and the effects of those changes, in turn, on the remote descendent variable. Such consistency and coherence constraints can be expressed as systems of SEM equations that can be solved to estimate the path coefficients relating changes in parent variables to changes in their children.

Summing these changes over all paths leading from exposure to response variables allows the total effect (via all paths) of a change in exposure on changes in expected responses to be estimated. Path analysis and other SEM models are particularly valuable for detecting and quantifying the effects of unmeasured ("latent") confounders based on the patterns of correlations that they induce among observed variables. Standard statistics packages and procedures, such as PROC CALIS in SAS, have made this technology available to modelers for the past four decades.

More recently, *causal graph models* (including Bayesian networks) have been developed, in which mechanisms need not be described by linear models, but may be described by nonlinear and probabilistic relations (e.g., the conditional probability distributions for the value of a node (i.e., variable) in a causal graph, given the values of the variables that point into it). These models greatly extend the flexibility and power of causal hypothesis testing and causal predictive modeling.[12]

These advances are particularly useful for characterizing uncertain causal relations. As noted by Samet and Bodurow (2008), "The uncertainty about the correct causal model involves uncertainty about whether exposure in fact causes disease at all, about the set of confounders that are associated with exposure and cause disease, about whether there is reverse causation, about what are the correct parametric forms of the relations of the exposure and confounders with outcome, and about whether there are other forms of bias affecting the evidence. One currently used method for making this uncertainty clear is to draw a set of causal graphs, each of which represents a particular causal hypothesis, and then consider evidence insofar as it favors one or more of these hypotheses and related graphs over the others."

[12] www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch25.pdf

Valid causal relations cannot be explained away by non-causal explanations

An older, but still useful, approach to causal inference from observational data developed largely in the 1960s and 1970s consists of showing that there is an association between exposure and response that cannot plausibly be explained by confounding, biases (including model and data selection biases and specification errors), or coincidence (e.g., from historical trends in exposure and response that move together but that do not reflect causation).

Quasi-experiment design and analysis approaches developed in social statistics (Campbell and Stanley, 1966) systematically enumerate potential alternative explanations for observed associations (e.g., coincident historical trends, regression to the mean) and provide statistical tests for refuting them, if indeed they can be refuted. A substantial tradition of *refutationist approaches* in epidemiology follows the same general idea, of providing evidence for causation by using data to explicitly test, and if possible refute, other explanations for E-R associations (Maclure, 1991).

As stated by Samet and Bodurow (2008), "Because a statistical association between exposure and disease does not prove causation, plausible alternative hypotheses must be eliminated by careful statistical adjustment and/or consideration of all relevant scientific knowledge. Epidemiologic studies that show an association after such adjustment, for example through multiple regression or instrumental variable estimation, and that are reasonably free of bias and further confounding, provide evidence but not proof of causation."[13]

As we have seen, these authors are overly optimistic in asserting that associations that are reasonably free of bias and confounding therefore provide evidence of causation (since, for example, the strong, statistically "significant" associations in regression models that often occur between levels of statistically independent random walks ("spurious regression") do not arise from confounding or bias, and since "In general, regression models for non-stationary variables give spurious results" due to coincident historical trends created by random processes that are not well described by regression models.[14] Nonetheless, the recommendation that "plausible alternative hypotheses must be eliminated by careful statistical adjustment and/or consideration of all relevant scientific knowledge" well expresses the refutationist point of view.

[13] www.nap.edu/openbook.php?record_id=11908&page=173

[14] www.econ.ku.dk/metrics/econometrics2_05_ii/slides/10_cointegration_2pp.pdf

Valid causal mechanisms are law-like: consistent and homogeneous

A proposed causal relation that turns out to be very heterogeneous, sometimes showing large significant positive effects and other times no effects or significant negative effects for exchangeable individuals under the same conditions, does not correspond to a law-like causal relation, and cannot be relied on to make valid causal predictions (e.g., by using mean values averaged over many heterogeneous studies). A true causal relation can be thought of as a stable relation (e.g., the conditional probability table (CPT) for a node in a Bayesian network or causal graph model) that gives the same conditional probabilities of output values whenever the input values are the same. Unexplained heterogeneity, in which the CPT appears to differ significantly when study designs are repeated by different investigators, signals that a causal mechanism has not yet been discovered, and that models and the knowledge they represent need to be further refined to discover and express predictively useful causal relations that reflect genuine causal mechanisms instead of random associations.

These methods for drawing valid causal conclusions suggest the following checklist for judging the adequacy of a quantitative health risk assessment (QRA) that claims to have identified a useful predictive causal relation between exposure concentrations and risk of adverse health effects (responses), i.e., causal exposure-response (E-R) or concentration-response (C-R) relations.

1. Does the QRA show that changes in exposures precede – and do not follow or coincide with – the changes in health effects that they are said to cause? Are results of change-point analyses, intervention analyses, and Granger causality tests presented, along with supporting data? (If effects turn out to precede their presumed causes, then unmeasured confounders or residual confounding by confounders that have been statistically "controlled for" may be at work.)

2. Does the QRA demonstrate that health effects cannot be made conditionally independent of exposure by conditioning on other variables (especially, potential confounders)? Does it present the details, data, and results of conditional independence tests showing that health effects and exposures share mutual information that cannot be explained away by any combination of confounders?

3. *Does the QRA present and test explicit causal graph models,* showing the results of formal statistical tests of the causal hypotheses implied by the structure of the model (i.e., which variables point into which others, as in

Figures 1 or 3)? Does it identify which alternative causal graph models are most consistent with available data (e.g., using the Occam's Window method of Madigan and Raftery, 1994)?[15] Most importantly, does it present clear evidence that changes in exposure propagate through the causal graph, causing successive measurable changes in the intermediate variables along hypothesized causal paths?

Such coherence, consistency, and biological plausibility demonstrated in explicit causal graph models showing how hypothesized causal mechanismsdovetail with each other to transduce changes in exposures to changes in health risks can provide compelling objective evidence of a causal relation between them, thus accomplishing what older and more problematic WoE frameworks have long sought to provide(Rhomberg et al., 2015).

4. Have non-causal explanations for statistical relations among observed variables (including exposures, health effects, and any intermediate variables, modifying factors, and confounders) been identified and refuted using well-conducted and reported statistical tests? Especially, have model diagnostics (e.g., plots of residuals and discussions of any patterns) and formal tests of modeling assumptions been presented that convincingly show that the models used appropriately describe the data to which the QRA applies them, and that claimed associations are not caused by model selection biases or specification errors, failures to model errors in exposure estimates and other explanatory variables, omitted confounders or other latent variables, uncorrected multiple testing bias, or coincident historical trends (e.g., spurious regression, if the exposure and health effects time series in longitudinal studies are not stationary)?

5. Have any causal mechanisms postulated in the QRA modeling been demonstrated to exhibit stable, uniform, law-like behavior, so that there is no substantial unexplained heterogeneity in estimated input-output (e.g., E-R or C-R) relations?

If the answers to these five questions are all yes, then the QRA has met the burden of proof of showing that the available data are consistent with a causal relation, and that other (non-causal) explanations are not plausible. The QRA can then proceed to quantify the changes in (probability distributions of) outputs, such as future health effects, that would be caused by changes in controllable inputs (e.g., exposure

[15] www.stat.cmu.edu/~fienberg/Statistics36-756/MadiganRaftery-JASA-1994.pdf

levels). The effort needed to establish valid evidence of a causal relation between historical levels of inputs and outputs by being able to answer yes to questions 1-5 pays off at this stage. Causal graph models (e.g., Bayesian networks), simulation models based on composition of validated causal mechanisms, and path diagrams and SEM models can all be used to predict quantitative changes in outputs caused by changes in inputs, e.g., changes in future health risks caused by changes in future exposure levels, given any scenario for the future values of other inputs (e.g., those with only outward-pointing arrows in Figure 3).

If the answer to any of the preceding five questions is no, then it is premature to make causal predictions based on the w done so far. Either the additional work needed to make the answers yes should be done, or results should be stated as contingent on the as-yet unproved assumption that this can eventually be done.

REFERENCES

Angrist JD, Pischke J-S. *Mostly Harmless Econometrics: An Empiricist's Companion*. 2009 Princeton University Press, Princeton, NJ.

Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004 May-Jun;66(3):411-21.

Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL, Thorlund K. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol. 2008 Sep;61(9):857-65.

Bender R, Lange S. Adjusting for multiple testing--when and how? J Clin Epidemiol. 2001 Apr;54(4):343-9

Buka I, Koranteng S, Osornio-Vargas AR. The effects of air pollution on the health of children. Paediatr Child Health. 2006 Oct;11(8):513-6.

Campbell DT, Stanley JC. Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNally, 1966.

Cami A, Wallstrom GL, Hogan WR. (2009). Measuring the effect of commuting on the performance of the Bayesian Aerosol Release Detector. BMC Med Inform DecisMak. Nov 3;9Suppl 1:S7.

Cox LA Jr, Popken DA. Has reducing fine particulate matter and ozone caused reduced mortality rates in the United States? Ann Epidemiol. 2015 Mar;25(3):162-73.

Cox LA Jr, Popken DA, Berman DW. Causal versus spurious spatial exposure-response associations in health risk analysis. Crit Rev Toxicol. 2013;43 Suppl 1:26-38

Cox LA Jr. Dose-response thresholds for progressive diseases. Dose Response. 2012;10(2):233-50.

Cox LA Jr. An exposure-response threshold for lung diseases and lung cancer caused by crystalline silica. Risk Anal. 2011 Oct;31(10):1543-60.

Dash D, Druzdzel MJ. A note on the correctness of the causal ordering algorithm. 2008. Artificial Intelligence 172:1800–1808. http://www.pitt.edu/~druzdzel/psfiles/aij08.pdf

Dash D, Druzdzel MJ. Robust independence testing for constraint-based learning of causal structure. In Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-03), pp 167-174. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 2003.

The Economist (2013). Trouble at the Lab: Scientists like to think of science as self-correcting. To an alarming degree, it is not. October 19, 2013.

Eichler M, Didelez V. On Granger causality and the effect of interventions in time series. Lifetime Data Anal. 2010 Jan;16(1):3-32.

EPA (U.S. Environmental Protection Agency). 2011. The Benefits and Costs of the Clean Air Act from 1990 to 2020. Final Report – Rev. A. Office of Air and Radiation, Washington D.C.

EPA (2006). Expanded expert judgment assessment of the concentration-response relationship between $PM_{2.5}$ exposure and mortality.

Fann N, Lamson AD, Anenberg SC, Wesson K, Risley D, Hubbell. Estimating the National Public Health Burden Associated with Exposure to Ambient $PM_{2.5}$ and Ozone. *Risk Analysis* Jan. 2012 32(1):81-95

Freedman DA. Graphical models for causation, and the identification problem. Eval Rev. 2004 Aug;28(4):267-93.

Friedman MS, Powell KE, Hutwagner L, Graham LM, Teague WG. Impact of changes in transportation and commuting behaviors during the 1996 Summer Olympic Games in Atlanta on air quality and childhood asthma. JAMA. 2001 Feb 21;285(7):897-905.

Friede T, Henderson R, Kao CF. A note on testing for intervention effects on binary responses. Methods Inf Med. 2006;45(4):435-40.

Friedman, N., & Goldszmidt, M. (1998) Learning Bayesian networks with local structure. In M.I. Jordan (Ed.), Learning in Graphical Models (pp 421-459). MIT Press. Cambridge, MA.

Gilmour S, Degenhardt L, Hall W, Day C. Using intervention time series analyses to assess the effects of imperfectly identifiable natural events: a general method and example. BMC Med Res Methodol. 2006 Apr 3;6:16.

Goodman JE, Prueitt RL, Sax SN, Bailey LA, Rhomberg LR. Evaluation of the causal framework used for setting national ambient air quality standards. Crit Rev Toxicol. 2013 Nov;43(10):829-49. doi: 10.3109/10408444.2013.837864.

Greenland S, Brumback B. An overview of relations among causal modelling methods. Int J Epidemiol. 2002 Oct;31(5):1030-7.

Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol. 1996 Dec;25(6):1107-16.

Grundmann O. (2014) The current state of bioterrorist attack surveillance and preparedness in the US. RiskManagHealthc Policy. Oct. 9;7:177-87.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics. 2009 Apr;10(2):258-74.

Hack CE, Haber LT, Maier A, Shulte P, Fowler B, Lotz WG, Savage RE Jr. A Bayesian network model for biomarker-based dose response. Risk Anal. 2010 Jul;30(7):1037-51.

Harvard School of Public Health, 2002. Press Release: "Ban On Coal Burning in Dublin Cleans the Air and Reduces Death Rates"

Health Effects Institute (HEI). 2013. Did the Irish Coal Bans Improve Air Quality and Health? HEI Update, Summer, 2013. Last Retrieved 1 February 2014.

Health Effects Institute (HEI) (2010). Impact of Improved Air Quality During the 1996 Summer Olympic Games in Atlanta on Multiple Cardiovascular and Respiratory Outcomes. HEI Research Report #148. April, 2010. Authors: Jennifer L. Peel, Mitchell Klein, W. Dana Flanders, James A. Mulholland, and Paige E. Tolbert. Health Effects Institute. Boston, MA.

Helfenstein U. (1991) The use of transfer function models, intervention analysis and related time series methods in epidemiology. Int J Epidemiol. Sep;20(3):808-15.

Hernan MA, Robins JM. (2011, 2015).Causal Inference. Chapman and Hall/VCFC Press (forthcoming). 2011 draft available at www.tc.umn.edu/~alonso/hernanrobins_v1.10.11.pdf

Hoeting, J., Madigan, D., Raftery, A., &Volinsky, C. (1999). Bayesian model averaging. *Statistical Science*, 14, 382-401.

Höfler M. The Bradford Hill considerations on causality: a counterfactual perspective. Emerg Themes Epidemiol. 2005 Nov 3;2:11.

Hora, S. Eliciting probabilities from experts." In *Advances in Decision Analysis: From Foundations to Applications*. Edited by Ward Edwards, Ralph F. Miles and Detlof von Winterfeldt, pages 129-153. Cambridge University Press, New York, 2007.

Ioannidis JPA. Why most published research findings are false. 2005. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124

Imberger G, Vejlby AD, Hansen SB, Møller AM, Wetterslev J (2011) Statistical Multiplicity in Systematic Reviews of Anaesthesia Interventions: A quantification and comparison between Cochrane and non-Cochrane reviews. PLoS ONE 6(12): e28422. doi:10.1371/journal.pone.0028422

James NA, Matteson DS. ecp: An R package for nonparametric multiple change point analysis of multivariate data. J. Statistical Software. Dec. 2014 62(7).

Joffe M, Gambhir M, Chadeau-Hyam M, Vineis P. Causal diagrams in systems epidemiology. Emerg Themes Epidemiol. 2012 Mar 19;9(1):1. doi: 10.1186/1742-7622-9-1.

Kahneman D. *Thinking Fast and Slow*. 2011. Farrar, Straus, and Giroux. New York, New York.

Kass-Hout TA, Xu Z, McMurray P, Park S, Buckeridge DL, Brownstein JS, Finelli L, Groseclose SL. Application of change point analysis to daily influenza-like illness emergency department visits. J Am Med Inform Assoc. 2012 Nov-Dec;19(6):1075-81. doi: 10.1136/amiajnl-2011-000793.

Kelly F, Armstrong B, Atkinson R, Anderson HR, Barratt B, Beevers S, Cook D, Green D, Derwent D, Mudway I, Wilkinson P; HEI Health Review Committee. The London low emission zone baseline study. Res Rep Health Eff Inst. 2011 Nov;(163):3-79.

Klein, Lawrence R. (1974). "Regression Systems of Linear Simultaneous Equations". *A Textbook of Econometrics* (Second ed.). Englewood Cliffs: Prentice-Hall. pp. 131–196. ISBN 0-13-912832-8.

Lehrer J. Trials and errors: Why science is failing us. Wired. January 28, 2012.

Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010 May;21(3):383-8.

Lu T-C, Druzdzel MJ, Leong T-Y. Causal mechanism-based model construction. In Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-00), pages 353-362, Morgan Kaufmann Publishers, Inc., San Francisco, CA, 2000

Lutter R, Abbott L, Becker R, Borgert C, Bradley A, Charnley G, Dudley S, Felsot A, Golden N, Gray G, Juberg D, Mitchell M, Rachman N, Rhomberg L, Solomon K, Sundlof S, Willett K. Improving Weight of Evidence Approaches to Chemical Evaluations. Risk Anal. 2014 Dec 16. doi: 10.1111/risa.12277.

Maclure M. Taxonomic axes of epidemiologic study designs: a refutationist perspective. J Clin Epidemiol. 1991;44(10):1045-53.

Maclure M. Multivariate refutation of aetiological hypotheses in non-experimental epidemiology. Int J Epidemiol. 1990 Dec;19(4):782-7.

Maldonado G, Greenland S. Interpreting model coefficients when the true model form is unknown. Epidemiology. 1993 Jul;4(4):310-8.

Mercer JB. Cold—an underrated risk factor for health. Environ Res. 2003 May;92(1):8-13.

Moore KL, Neugebauer R, van der Laan MJ, Tager IB. Causal inference in epidemiological studies with strong confounding. Stat Med. 2012 Feb 23. doi: 10.1002/sim.4469.

Morabia A. Hume, Mill, Hill, and the sui generis epidemiologic approach to causal inference. Am J Epidemiol. 2013 Nov 15;178(10):1526-32.

Nakahara S, Katanoda K, Ichikawa M. Onset of a declining trend in fatal motor vehicle crashes involving drunk-driving in Japan. JEpidemiol. 2013;23(3):195-204.

Nuzzo R. 2014. Scientific method: Statistical errors. P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 506, 150–152 doi:10.1038/506150a

NHS, 2012. Air pollution 'kills 13,000 a year' says study

Ottenbacher KJ (1998) Quantitative evaluation of multiplicity in epidemiology and public health research. Am J Epidemiol 147: 615–619.

Pelucchi C, Negri E, Gallus S, Boffetta P, Tramacere I, La Vecchia C. Long-term particulate matter exposure and mortality: a review of European epidemiological studies. BMC Public Health. 2009 Dec 8;9:453.

Pope, CA (2010). Accountability Studies of Air Pollution and Human Health: Where are We Now, and Where Does the Research Need to Go Next?

Piegorsch WW, An L, Wickens AA, West RW, Peña EA, Wu W. Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. Environmetrics. 2013 May 1;24(3):143-157.

Powell H, Lee D, Bowman A. Estimating constrained concentration–response functions between air pollution and health. Environmetrics. 2012 May;23(3): 228-237.

Rhomberg LR, Goodman JE, Bailey LA, Prueitt RL, Beck NB, Bevan C, Honeycutt M, Kaminski NE, Paoli G, Pottenger LH, Scherer RW, Wise KC, Becker RA. A survey of frameworks for best practices in weight-of-evidence analyses. Crit Rev Toxicol. 2013 Oct;43(9):753-84. doi: 10.3109/10408444.2013.832727.

Roberts S. Biologically plausible particulate air pollution mortality concentration-response functions. Environ Health Perspect. 2004 Mar;112(3):309-13.

Rothman KJ,Lash LL, Greenland S. (2012) Modern Epidemiology: 3rd Edition. Lippincott, Williams, & Wilkins. New York, New York.

Sarewitz D. Beware the creeping cracks of bias. Nature. 10 May 2012 485:149

Skrøvseth SO, Bellika JG, Godtliebsen F. Causality in scale space as an approach to change detection. PLoS One. 2012;7(12):e52253. doi: 10.1371/journal.pone.0052253.

Stebbings JH Jr. Panel studies of acute health effects of air pollution. II. A methodologic study of linear regression analysis of asthma panel data. Environ Res. 1978 Aug;17(1):10-32.

Swartz MD, Yu RK, Shete S. Finding factors influencing risk: comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. Stat Med. 2008 Dec 20;27(29):6158-74.

Viallefont V, Raftery AE, Richardson S. Variable selection and Bayesian model averaging in case-control studies. Stat Med. 2001 Nov 15;20(21):3215-30.