Characterizing the Long-term $PM_{2.5}$ Concentration Response Function: Comparing the Strengths and Weaknesses of Research Synthesis Approaches

Neal Fann (Environmental Protection Agency), Elisabeth A. Gilmore (University of Maryland), Katherine Walker (Health Effects Institute)*

Working Paper prepared for:

Methods for Research Synthesis:
A Cross-Disciplinary Workshop

Harvard Center for Risk Analysis

October 3, 2013

www.hcra.harvard.edu

***Corresponding author: kwalker@healtheffects.org**

Acknowledgements: Support for work on this paper was provided by Michael Cackoski who compiled the studies for the meta-analysis.

Characterizing the long-term $PM_{2.5}$ concentration response function: Comparing the strengths and weaknesses of research synthesis approaches

Neal Fann, Elisabeth Gilmore, Katherine Walker

The shape, magnitude and degree of certainty in the relationship between long-term population exposure to ambient levels of fine particulate matter ($PM_{2.5}$) and the risk of premature mortality has been one of the most intensely studied issues in environmental health. To meet the needs of decision-makers, this relationship has been estimated using several data synthesis techniques, namely systematic review, quantitative meta-analysis, expert judgment and integrated exposure – response (IER). This provides a valuable opportunity to compare and contrast alternate research synthesis approaches in a policy-relevant context. Here, we evaluate these methods as applied to the $PM_{2.5}$ concentration – response (C-R) function along a set of criteria and draw broader lessons to judge the factors that could make an approach more or less suited to informing policy decisions. We consider the trade-offs associated with the validity, the transparency, the suitability to the policy problem and the accessibility of each approach. We also compare the results of a new meta-analysis for the C-R function to the other quantitative estimates for a United States Environmental Protection Agency (USEPA) regulatory analysis. We find that the observed trade-offs between the criteria are largely a function of the state of knowledge about the C-R function. For example, the larger expenditure of time and resources as well as the more challenging interpretation of the results of an expert elicitation could be justified due to the absence of policy relevant evidence at the time of analysis. We also find that these techniques are not mutually exclusive. A systematic review of the multidisciplinary evidence is an essential starting point for all methods. The five-year reviews of the literature as conducted by the USEPA highlight the growing pool of convergent literature. This supports the use of meta-analysis and IER approaches to derive valid estimates that constrain the C-R function. Ultimately, however, all of these methods require considerable judgment of scientists, individually and collectively. Finding ways for all these methods to acknowledge, appropriately elicit and examine the implications of that judgment would be an important step forward for research synthesis.

Keywords: Concentration-response function, fine particulate matter ($PM_{2.5}$), expert elicitation, meta-analysis, integrated exposure-response

1. INTRODUCTION

Regulatory impact assessments (RIAs) for rules with implications for exposure to ambient concentrations of fine particulate matter ($PM_{2.5}$) routinely estimate monetized benefits in the tens or hundreds of billions of dollars, attributable largely to reductions in the risk of premature mortality. The benefits quantified in these RIAs are among the largest estimated for any U.S regulation (Office of Management and Budget, 2013). The quantitative relationship between changes in exposure to ambient $PM_{2.5}$ and the risk of premature mortality (i.e. the concentration-response, C-R, function), and associated assumptions about the likelihood that such a relationship is causal, are key inputs to these analyses. Given the magnitude of monetized benefits associated with reductions of ambient concentrations of $PM_{2.5}$, policymakers have historically expressed a strong desire to better characterize the magnitude, functional form and the uncertainties in this C-R function. To meet the demand for this information, researchers have applied a range of alternative research synthesis approaches, including systematic review, quantitative meta-analysis, expert judgment and integrated exposure research estimates, to essentially the same policy question. In this respect, the long-term $PM_{2.5}$ all-cause mortality relationship is a unique and rich test-bed that allows us to compare the strengths and limitations of approaches to synthesizing data to inform critical policy questions.

In the first set of analyses of the benefits and costs of the Clean Air Act (CAA) conducted by the USEPA (USEPA, 1997), the Agency quantified changes in the incidence of $PM_{2.5}$-related deaths using a C-R function reported by one long-term exposure study (Pope et al., 1995). In quantifying these premature deaths, the USEPA drew upon the consensus from the Clean Air Scientific Advisory Committee (CASAC), concluding that the evidence available at that time was sufficiently compelling "to warrant an assumption of a causal relationship and derivation of quantitative estimates of a PM-related premature mortality effect" (USEPA, 1997). Since the National Ambient Air Quality Standards (NAAQS) where established in 1997, the shape, magnitude and degree of uncertainty in this relationship has been intensely studied and hotly contested. Concerns about how the Agency characterized uncertainties in the PM-mortality relationship led to at least two other approaches to quantifying the C-R function and expressing its uncertainty. Specifically, in its 2002 review

of the benefits assessment methodology, the National Research Council (NRC) encouraged the USEPA to consider more fully the impact of several uncertainties that were not effectively captured using the slope and standard error from the Pope et al. (2002) study or through sensitivity analyses (National Research Council 2002). In 2004, the USEPA undertook a project designed to elicit the expert judgments of several epidemiologists, a toxicologist and a clinician to more fully and quantitatively characterize the state of scientific knowledge about the $PM_{2.5}$-mortality C-R relationship. The results of this expert elicitation have subsequently been employed in the RIAs as alternative estimates of the benefits. More recently, under the Global Burden of Disease (GBD) project, several scientists have collaborated to develop an integrated exposure response (IER) function, which is intended to span the full range of global ambient PM concentrations and includes estimates of uncertainty (Pope et al., 2011, Burnett et al., 2013). Here, the scientists combine the relative risk estimates of several studies of exposure to diverse sources of combustion related particulate matter exposures. The IER approach is described more fully by Shin et al. (2013) in this series.

In this paper, first, we review these methods as applied to the $PM_{2.5}$ all-cause mortality C-R function and discuss how these methods provide different types of information for the analysis of USEPA rulemaking. The expert elicitation and IER approaches are well documented and offer extensive quantitative estimates. Quantitative meta-analysis, while commonly employed in the literature (e.g. Hoek et al, 2013), has not been applied to long-term $PM_{2.5}$ mortality for a USEPA benefits analysis. We perform a meta-analysis as a straw man to compare this approach against the expert elicitation and IER techniques. Second, we develop a set of criteria, namely explanatory power, credibility, accessibility and repeatability, and evaluate the factors that may make an approach more or less suited to informing policy decisions in the context of long-term $PM_{2.5}$ mortality. In particular, we consider the trade-offs of each approach with respect to the criteria. Finally, we compare the quantitative estimates of the C-R function generated by these techniques and show the implications for the avoided $PM_{2.5}$-related premature deaths due to the implementation of the Mercury and Air Toxics Standards (MATS) (USEPA, 2011) and draw some broader conclusions.

2. REVIEW OF EXISTING ANALYSIS AND TECHNIQUES FOR THE $PM_{2.5}$ C-R FUNCTION

Below we describe the four research synthesis approaches as applied to the $PM_{2.5}$ C-R function: systematic review, expert elicitation, quantitative meta-analysis, and the IER technique. We focus on the three quantitative methods.

2.1 Systematic Review

Systematic review is a first stage for the other three research synthesis techniques. This initial step requires analysts to identify and conduct a critical review of the pool of existing literature that fits predefined criteria and relevance to the policy question under consideration. The central goal is to provide a basis for assessing the quality of individual studies, their coherence, and their contribution to the overall weight of evidence or completeness of the state of knowledge on the questions of interest while minimizing the potential for bias in the choice of evidence. The primary guidance for systematic reviews comes from the Cochrane Collaboration that focuses on human health care and policy (Higgins & Green, 2011). Moher et al. (2007) find, however, that the quality of systematic reviews can vary widely (Moher et al., 2007). The conduct of systematic reviews and how the evidence is compiled and considered is discussed in other papers in this series (e.g. Rhomberg et al., 2013).

For the C-R relationships between $PM_{2.5}$ and all-cause mortality and other health outcomes, systematic reviews have been published as broad assessments of the coherence of the existing literature (Pope & Dockery, 2006). They are also regularly conducted as part of the USEPA's 5-year reviews of the ambient air quality standards, known as Integrated Science Assessment (ISA). We will focus on the ISA and do not attempt to cover all systematic reviews in this paper. The ISA has formed an important basis for the USEPA's qualitative assessment of the strength of the evidence for a causal relationship and for its selection of the studies on which it has based its primary estimates of the $PM_{2.5}$ C-R relationship and on the others that it has used for sensitivity analyses. Presently, the effect of long-term exposure to $PM_{2.5}$ and mortality is classified as "likely to be causal" with evidence sufficient to conclude that that a causal relationship is likely, but important uncertainties remain (USEPA, 2009).

2.2 Expert Elicitation

Analysts across a wide spectrum of scientific and technical disciplines rely to varying degrees on the judgment of experts. In *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*, Morgan and Henrion (1992) succinctly outlined the arguments for a more explicit and quantitative characterization of what experts know and how well they know it (see Text Box 1). The field of quantitative elicitation of experts' judgments has emerged in an effort to develop a more explicit, transparent, cognitively sound and statistically robust process for characterizing what experts think they know about a quantity or event of interest and the confidence with which they know it.

TEXT BOX I: Excerpt from Morgan and Henrion (1992)

1. A central purpose of policy research and policy analysis is to help identify the important factors and the sources of disagreement in a problem, and to help anticipate the unexpected. An explicit treatment of uncertainty forces us to think more carefully about such matters, helps us to identify which factors are most and least important, and helps us plan for contingencies or hedge our bets.

2. Increasingly we must rely on experts when we make decisions. It is often hard to be sure we understand exactly what they are telling us. It is harder still to know what to do when different experts appear to be telling us different things. If we insist that they tell us about the uncertainty of their judgments, we will be clearer about how much they think they know, and whether they really disagree.

3. Rarely is any problem solved once and for all. … The details may change but the basic problems keep coming back again and again. Sometimes we would like to use, or adapt, policy analyses that have been done in the past to help with the problems of the moment. This is much easier to do when the uncertainties of the past work have been carefully described, because then we can have greater confidence that we are using the earlier work in an appropriate way.

Formal methods for eliciting quantitative probabilistic judgments have been pioneered by Cooke (1991), Morgan and Henrion (1992) and many others (Cooke, 1991; Morgan & Henrion, 1992). Expert elicitation methods are most often used to obtain quantitative judgments about questions where data are limited. Any question posed for elicitation must be in theory knowable with the proper experiment and collection of data; in practice, however, the experiments may be too expensive, intrusive, dangerous or require too much time before a decision must be made. Each expert's judgment about the quantity of interest

must rely on the relevant evidence – typically, some kind of systematic review is conducted. Each expert also provides uncertainty estimates that are not necessarily equivalent to the empirically derived confidence intervals on an estimate, although an expert could adopt those as his or her own best estimate. Rather, an expert subjectively assesses his or her state of knowledge, given the evidence—that is, what he or she knows and with what level of confidence with which he or she knows it. Experts can and do differ in their assessments. This difference provides analysts and decision-makers with a broad assessment of the state of knowledge within the scientific community. Some investigators also calibrate how well experts perform on questions for which the answer are or will shortly become known (e.g. Alpert & Raiffa, 1969; Lichtenstein et al., 1977; Walker et al., 2001, 2003) and/or combine experts using equal weights or weights based on performance (Cooke, 1991). Others prefer to conduct a sensitivity analysis on each value separately (Morgan & Henrion, 1992).

Prior to the elicitation, the USEPA, with the advice of its Scientific Advisory Board's Subcommittee on Health and Environmental/Exposure, relied for its primary estimates on the largest and most geographically representative study of ambient $PM_{2.5}$ and mortality by Pope et al. (2002) known as the American Cancer Society (ACS) cohort (Pope et al., 2002). This produced a no-threshold, log-linear relationship with, on average, a 6% increase in all-cause mortality for each 10 $\mu g/m^3$ increase in $PM_{2.5}$ concentrations. Uncertainty was based only on the standard error of the central estimate. The Harvard 6-city (H6C) study, which showed a stronger relationship, was later included as a sensitivity analyses (Laden et al., 2006). The NRC suggested that "EPA's decision to incorporate only one source of uncertainty, the random sampling error in the estimated concentration-response function, into the probability distributions resulting from its health benefits analyses is worth reconsidering" (National Research Council, 2002). This was consistent with the Agency's own judgment that the probability distribution from Pope et al. (2002) likely gives "a misleading picture about the overall uncertainty in the estimates" and suggests "there is less uncertainty, perhaps much less, than is actually present" (USEPA, 1999).

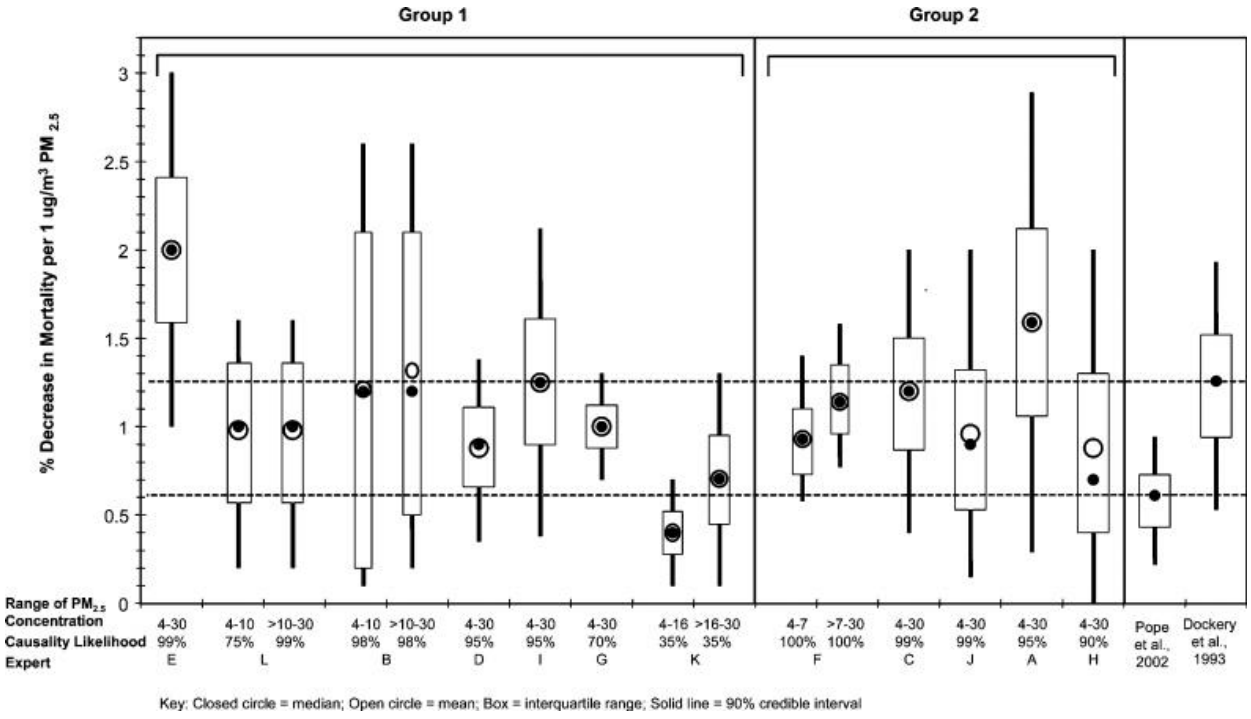
The USEPA first conducted a pilot expert elicitation study with 5 experts (USEPA, 2004, USEPA, 2005). This was followed by a full-scale elicitation of judgments from 12 U.S. and

Canadian scientists with expertise in toxicology, clinical medicine, epidemiology and biostatistics (Roman et al., 2008). A second elicitation of European experts was conducted at about the same time as the full-scale study by a separate team of investigators studying risks of the oil fires in Kuwait (Cooke et al., 2007). We focus on the USEPA study although many of the same points will apply. While there are many ways in which experts' judgments might be formally used, in the 2005 study, the experts were asked to estimate directly "the national average C-R function for adults 18 and older exposed to annual average $PM_{2.5}$ levels between 4 and 30 $\mu g/m^3$ that could be applied throughout the U.S. in a benefits analysis". This is similar to what the Pope et al. (2002) study provided but with a more complete description of the other sources of bias or uncertainty from the experts. Specifically, they were asked to estimate the "true percent change in annual, all-cause mortality in the adult U.S. population resulting from a permanent 1 $\mu g/m^3$ reduction in $PM_{2.5}$" and to characterize their uncertainty in that estimate by providing the minimum, 5^{th}, 25^{th}, 50^{th}, 75^{th}, and maximum values of the effect estimate. The experts were also asked about the shape of the C-R function including the likelihood of a population threshold, and to give their judgment about the likelihood that the observed $PM_{2.5}$ mortality relationship was causal.

We reproduce the results from the elicitation in Figure 1. Most of the experts' central estimates fell at or above the 2002 ACS median (0.6% per $\mu g/m^3$) and below the original H6C median (1.2% per $\mu g/m^3$), although the uncertainty bounds were much broader than those found in either study. The study design and elicitation is described in detail in Roman et al. 2008, and in the underlying report prepared for EPA (Industrial Economics Inc., 2006). Briefly, the elicitation protocol starts with a detailed written protocol for the elicitation interview, intended to foster a systematic consideration of key evidence and issues (e.g. likelihood of causal relationships, functional forms of the C-R relationship, likelihood of thresholds) leading up to the final judgments. The experts are identified and selected using a two-part process involving publication counts and peer-nomination. The experts are then provided with a briefing book with a broad set of scientific studies, air pollution data, and population statistics. This is followed by daylong individual interviews conducted by a team of interviewers, with access to all studies and software for plotting

and exploring experts' judgments during the interview. Pre- and post- elicitation workshops are run to familiarize participants with the evidence, process and present the results.

Figure 1: Uncertainty Distributions for $PM_{2.5}$ – mortality C-R Coefficient for annual average $PM_{2.5}$ concentrations of 4-30 $\mu g/m^3$. The box plots represent distributions as provided by the experts to the elicitation team. Experts in Group 1 gave distributions conditional on a causal relationship and separately gave probabilistic judgments about the likelihood of a causal or non-causal relationship. Experts in Group 2 preferred to give distributions that incorporated their judgments about the overall likelihood that the $PM_{2.5}$ – mortality association was causal. Therefore, the expert distributions from these two groups are not directly comparable. Reproduced from Roman et al. (2008)[***Permission requested***].



2.3 Quantitative Meta-Analysis

The quantitative meta-analysis serves to combine underlying data or risk estimates quantitatively across studies and to help identify factors that might explain heterogeneity in risk estimates among those studies (Higgins & Green, 2011). Combining the data or

results of several similar studies can increase the explanatory power, especially if some of the studies suffer from a small sample size. Likewise, meta-analysis can provide insight to the factors in the data that explain variability in the results between studies, including the possible presence of publication bias. However, the ability of meta-analysis to provide more accurate and precise estimates of the C-R function for populations can also be limited by underlying uncertainties or biases in the selected studies. Meta-analysis may or may not accompany a systematic review.

To explore the strengths and limitations of meta-analysis, we performed two illustrative quantitative random effects meta-analyses of epidemiological studies to examine the relationship between long-term exposure to $PM_{2.5}$ and the risk of premature mortality. These analyses will be based on combination or pooling of the published results (e.g. hazard ratios) and not on the underlying data. A random-effects meta-analysis assumes that "the effects being estimated in the different studies are not identical, but follow some distribution. The model represents our lack of knowledge about why real, or apparent, intervention effects differ by considering the differences as if they were random. The centre [ibid] of this distribution describes the average of the effects, while its width describes the degree of heterogeneity. The conventional choice of distribution is a normal distribution (Higgins & Green, 2011)." By contrast, a fixed effect meta-analysis assumes that the "true" effect is the same across studies, is generally considered to be less plausible in the context of air pollution epidemiology. We performed this stage of the analysis using the Comprehensive Meta-Analysis software package, version 2 (Borenstein et al., 2005).

We draw upon the literature from previous systematic reviews in two phases. First, we employ the epidemiological studies available to the experts in USEPA's 2006 elicitation (see Exhibit 3-3 of (Industrial Economics Incorporated, 2006) and (Roman et al., 2008)). This allows us to compare the results of the expert elicitation to a meta-analysis based on the same literature. Second, we include additional relevant long-term mortality studies published since the completion of the 2006 Expert Elicitation and included in Chapter 2 of the USEPA's 2009 "Integrated Health Assessment for Particulate Matter " (USEPA, 2009) and Table 5-8 in the USEPA's 2012 "Regulatory Impact Analysis for the Final Revisions to

the National Ambient Air Quality Standards for Particulate Matter" (USEPA, 2012). This second analysis allows us to evaluate the sensitivity of the first meta-analytic results to the incorporation of these additional studies. Figures 2 and 3 below summarize the results of each stage of the meta-analysis. The median pooled estimates from the two stages differ slightly. The confidence intervals, however, are narrower when we add the additional studies, showing a convergence of the estimates.

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram (Supplemental Figure 1) summarizes the steps we followed to identify studies appropriate for the quantitative meta-analysis. The three sources yielded a total of 107 articles. 101 of these studies specified $PM_{2.5}$ as the indicator; we excluded six studies specifying total suspended particulates, sulfate or PM_{10}. Of these 101 studies, 32 reported the risks of premature all-cause mortality from long-term exposure to $PM_{2.5}$; the remaining 69 studies assessed impacts other than premature mortality, assessed impacts from short-term changes in $PM_{2.5}$ or did not report an all-cause mortality risk estimate. Of the 32 studies, 29 assessed risks among adults; the remaining 3 characterized risks among children. Many of the 29 remaining studies report risk estimates from a single cohort (e.g. the ACS). To avoid over-weighting the meta-analysis toward one cohort, we selected the latest published study of each cohort. Supplemental Table 1 identifies the sets of studies incorporated into the first and second meta-analyses described above and summarizes some of the key characteristics including the cohort attributes, geographic scope, follow-up period, and hazard ratios. In certain cases, we included study results in both meta-analyses (e.g. for the ACS and H6C cohorts) to ensure that our analysis was as representative of the literature as possible. However, we also took care to select only one risk estimate from studies of each cohort, to reduce the risk of over-weighting any given cohort.

Figure 2: Pooled Estimate of Long-term All-Cause Mortality Using the Studies Available to the Experts in EPA's 2006 Elicitation

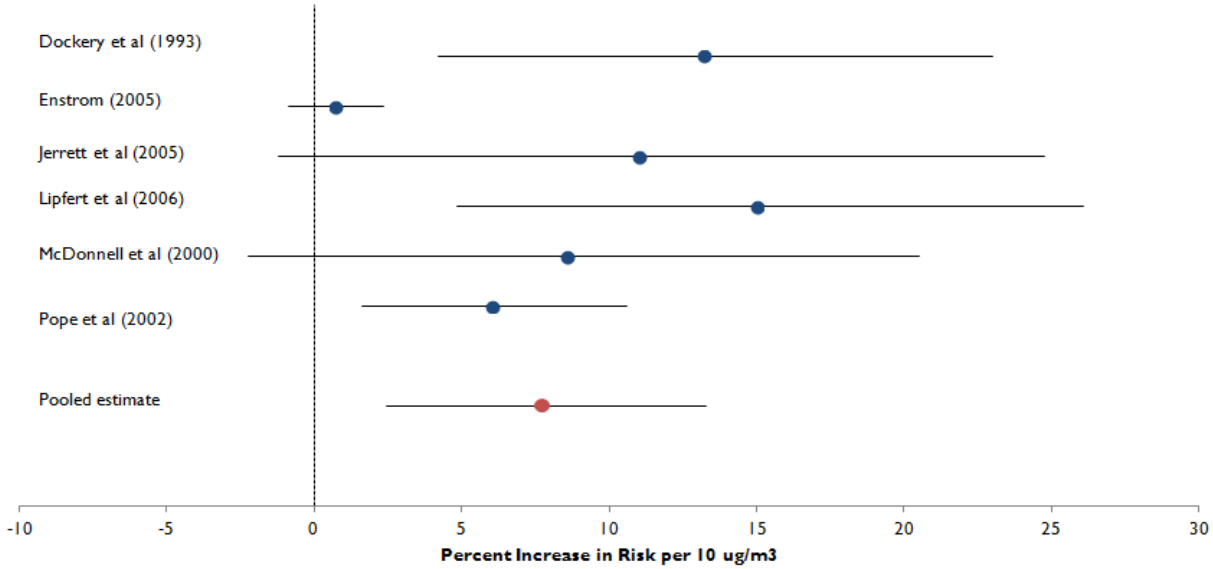
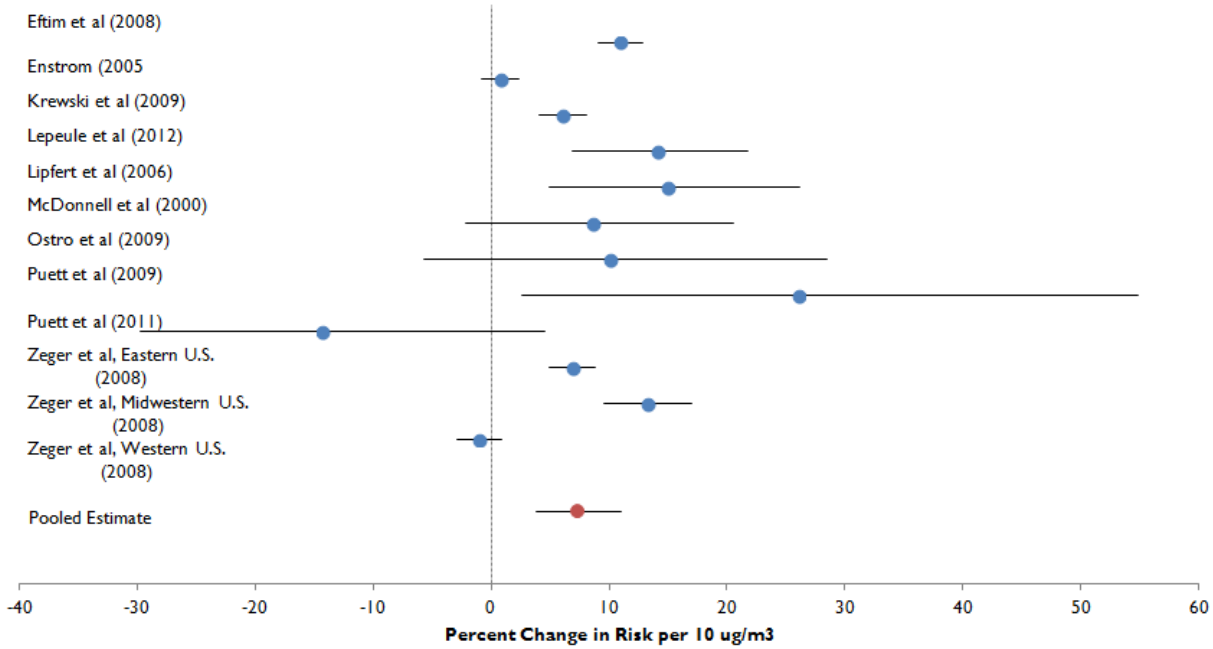


Figure 3: Pooled Estimate of Long-term All-Cause Mortality Using the Studies Available to the Experts in EPA's 2006 Elicitation and the Newer Studies



We recognize that our illustrative analysis cannot evaluate all of the strengths and weaknesses of meta-analysis, which is well covered in the extensive literature on this subject. Rather, our analysis is intended to describe some of the benefits and challenges as applied to the study of $PM_{2.5}$ and mortality. Consequently, Supplemental Table 1 compares how the various cohorts differ with respect to study period and certain population characteristics (e.g. age), but does not address various other systematic differences in individual and ecological variables that may exist among cohorts, including socioeconomic status, occupation, baseline health status and, composition of $PM_{2.5}$ particle, among others. These would need to be fully addressed in a more complete meta-analysis, though this is challenging without reanalyzing the original analytical dataset. We also question whether these variables can be adequately controlled for in a meta-analysis.

2.4 Integrated Exposure Response Assessment

Finally, we review the IER analysis presented in detail in the paper by Shin et al. (2013) in this series. This methodology brings together research on particulate exposure from sources other than ambient $PM_{2.5}$ in order to evaluate the consistency of the findings and to use the data to characterize quantitatively the shape of the C-R function over a broader range of exposures. Part of the motivation was that the linear C-R function derived from U.S. studies performed poorly at characterizing mortality risk in countries where the $PM_{2.5}$ concentrations were much higher than those observed in the U.S.

Pope et al. (2009) first brought together data on associations of cardiovascular mortality with $PM_{2.5}$ exposures from active smoking, secondhand smoking and ambient air pollution (Pope et al., 2009). The authors later applied a similar method to explore the $PM_{2.5}$ relationships with mortality from cancer, ischemic heart disease, CVD, and cardiopulmonary disease (Pope et al., 2011). More recently, the method has been adapted for the development of cause-specific mortality-$PM_{2.5}$ exposure response functions for the Global Burden of Disease (GBD) project (Pope et al., 2011; Burnett et al., 2013). That paper also offers approaches to characterizing uncertainty in the exposure response relationship over the full range of exposures. This novel method offers important advances over the reliance on ambient air pollution studies alone. By assembling evidence from a broader

range of studies, they put the results of air pollution studies into improved perspective and strengthen the overall inferences about the contributions to mortality from particulate exposures. While offering distinct improvements over previous approaches, this analysis requires a number of important assumptions to be made in selecting and adapting studies with different types of exposure to describe ambient $PM_{2.5}$.

3. EVALUATION OF THE METHODS

The charge to the investigators on this project included developing a basis for evaluating the methods. Specifically, we were asked:

1. What criteria should be used to evaluate the applicability of different research synthesis methods to particular types of problems and data?

2. What particular characteristics of the problem and data make the research synthesis method(s) you address particularly well (or poorly) suited for that context?

3. What are the strengths and limitations of the outputs provided, and the implications for their use in policy analysis?

4. What are the most important research needs, in terms of methodological development, given your findings?

We developed two approaches for evaluating the appropriateness, strengths and limitations of each approach. First, we describe both scientific and policy criteria for evaluating each method. Second, we compare the impact of the results of different methods in a hypothetical regulatory analysis.

3.1 Developing a Common Set of Criteria

Decision-makers, while perhaps familiar with certain of the basic attributes of various research synthesis approaches, are arguably not acquainted with the resources each

approach requires or the many trade-offs in reliability and accessibility associated with various techniques. To facilitate comparisons between methods and to illuminate these trade-offs, we evaluate each approach against a common set of criteria that we present in Table 1. The criteria we listed are intended to span both basic scientific attributes of sound analysis and the attributes of good analyses intended to inform policy decisions; the latter were informed in part by principles laid out in Text Box 1 (Morgan & Henrion, 1992). Some criteria are general, while others are particular to the application or policy questions for $PM_{2.5}$. Each criterion is useful to probing the types of technical and policy question each technique can answer, to assess the ability of each approach to use evidence, to describe what role analyst judgment plays, to consider the extent to which each method advances our state of knowledge regarding $PM_{2.5}$-related mortality, and finally to consider the time and financial resources required, and the accessibility of the results.

Table 1: Summary of Evaluation Criteria

How Valid are the Methods and Findings?

Characteristic	**Summary**
Interdisciplinary design and analysis	The study design reflects broad scientific input from relevant set of disciplines and a range of scientific viewpoints
Completeness	Explicit inclusion criteria identify appropriate studies or evidence to analyze to address the scientific or policy question
Analytical methods appropriate for the data	Appropriateness to data can be demonstrated with sensitivity to methodological choices
Independently verification of methods and results	Models have been validated; results have been replicated or reproduced in other settings
Independent peer review	The analysis and interpretation received rigorous peer review

How Transparent are the Methods and Results?

Characteristic	Summary
Expert/analyst selection	The experts or analysts involved are selected via an explicit set of criteria and process
Cognitive biases	Explicitly acknowledge and consider cognitive biases when selecting, analyzing, and drawing inferences about data (e.g. overconfidence, anchoring)
Hypotheses, data, models, assumptions	Identify explicitly, justify, and test sensitivity to key hypotheses, sources of data, models, and key assumptions. Describe the amount of weight assigned to various sources of evidence
Uncertainty and variability analysis	The analysis explicitly and fully characterizes key sources and extent of uncertainty and variability

How suited is the method to the policy problem?

Characteristic	Summary
Relevant exposures	The concentration-response or exposure-response function accounts for exposures experienced by the target population
Relevant health outcome	Reports quantitative estimates for the outcomes of most importance
Relevant population	The attributes of the study population(s) match those of the populations-at-risk

How suited is the method to the resources?

Characteristic	Summary
Data requirements	The approach can use readily available data
Level of resources	The amount of time and financial resources required are commensurate with the value of information generated
Repeatability/updatability	The analysis can be readily updated to reflect the latest science
Accessibility/communicability	The ease of use of results for policy analysis

3.2 Discussion of Methods

Here, we assess the extent to which each of the four approaches satisfies the criteria in Table 1 above, focusing in particular on a subset of criteria that are most relevant to each method.

3.2.1 Does the approach generate valid estimates?

Decision-makers look for research synthesis methods to yield estimates that are valid because they will ultimately use this information to select policies whose costs and benefits can be substantial. We consider both internal and external validity defined as follows: whether the necessary scientific disciplines have been involved in defining the research questions as well as the way in which data are collected and analyzed; ensuring that the data or study results used in an analysis represent an unbiased view of the universe of results; and, that the methods for synthesizing or aggregating the data are themselves appropriate for the research question. All research synthesis methods, in principle, can satisfy this requirement. However, there are differences of degree that become apparent when evaluating their applicability to the long-term C-R function for $PM_{2.5}$ mortality.

The literature linking $PM_{2.5}$ to adverse outcomes crosses many disciplinary boundaries including, and certainly not limited to, epidemiology, toxicology and biostatics. The first $PM_{2.5}$ standard in 1997 was greatly influenced by the C-R functions reported in two population level environmental epidemiological studies (Dockery et al., 1993; Pope et al., 1995). More recent efforts, however, have increasingly reflected the multidisciplinary nature of the literature. Specifically, the systematic review as conducted by the USEPA in its ISA is designed to identify literature across a broad range of technical disciplines. The ISA reviews relevant science ranging from exposure, to laboratory and clinical toxicology, to epidemiologic evidence; thus, we find that it is most multidisciplinary by design.

While the ISA can be thought of as a document reflecting scientific consensus, the other approaches aim to provide quantitative estimates. They differ on the degree that they can explicitly account for differences in the viewpoints from the different scientific communities. The expert elicitation was designed to be interdisciplinary. First, a multi-stakeholder group comprised of staff from the USEPA and the Office of Management and Budget governed the project. Further, the expert identification process was designed to find experts in exposure, epidemiology, biostatistics, toxicology and clinical medicine with the final expert group consisting of epidemiologists, a toxicologist and a clinician. In this way, the assessment benefited from the input of multiple disciplines in both the design of the questions and elicitation protocol (Roman et al., 2008). Thus, the expert elicitation was designed to allow for and to explore differences in how members of the scientific community would interpret evidence. By contrast, the analysts involved in meta-analysis and IER analyses are, as often is the case in scientific collaborations, self-selected. These individuals represent expertise that the analysts believe to be most appropriate for the data and the questions asked of it – primarily biostatisticians and epidemiologists – who work toward consensus on methodology and interpretation.

When there are broadly acknowledged sources of disagreement or uncertainty, then the expert elicitation enjoys a distinct advantage over those that represent a more singular disciplinary focus. Sources of uncertainty for the C-R function include questions about the

accuracy of exposure estimates in the relevant epidemiologic studies, factors that may modify the effect, the true shape of the C-R function, and the strength of evidence for a causal mechanism, among others. For example, an area of disagreement—or perhaps uncertainty—among air pollution scientists has been the extent to which toxicology can or should inform judgments about, or analyses of, the nature of the C-R function in human population data. Thus, if the purpose of public policy analysis is "to help identify the important factors and the sources of disagreement in a problem" and to evaluate how much these differences matter to public policy decisions, then the approach undertaken by the $PM_{2.5}$ expert elicitation is preferred. This benefit, however, is only critical if there are significant known areas of disagreement. If there is a consensus *a priori* on whose views were likely to be most expert, then those individuals' judgments should be directly elicited. To the extent that these experts are collected in the same disciplines, then the distinctions between these methods may be less important.

A second and related criterion is that of completeness, which refers to whether the approach has considered the evidence that is appropriate for the policy problem. To the extent that the analyst defines and makes available their search and inclusion criteria, any of these methods can and have been successful in satisfying this criterion. The appropriateness of the inclusion criteria with respect to the scientific validity of the findings and the ability to identify potential biases in the literature differs by method. What is considered relevant is inevitably influenced by who is involved in the discussion and how they define the relevant scientific evidence. Thus, clear inclusion and study selection criteria cannot overcome disagreements. The methods, however, differ in how and what evidence can be incorporated into the final analysis and whether any biases can be identified.

Systematic reviews like the ISA and the expert elicitation have the benefit from being able to consider a broad range of evidence from multiple disciplines. Both rely on this evidence to inform holistic judgments about the likelihood of a causal relationship and to identify those studies that the authors consider to be most informative about the concentration response relationships. In the case of the expert elicitation, each expert is allowed to draw

broadly on his or her knowledge and literature that may not be typically included in a systematic review to characterize probabilistically the nature and magnitude of the C-R relationship. For example, the expert elicitation protocol asked the experts to use evidence from passive and active smoking as a "reality check" on their own estimates of the mortality risks of $PM_{2.5}$ exposure. In addition, these methods can include "grey" literature, conference proceedings, government reports, white papers and other forms of literature not controlled by commercial publishers; this is important since there is some evidence of positive publication bias - that is, studies that report positive results are more likely to be published than those that report null results (Anderson et al., 2005; Bell et al., 2005).

By contrast, meta-analysis has a relative advantage to the expert elicitation in that it should explicitly state what types of studies it incorporates; however, the technique generally considers a smaller universe of studies because it is usually limited to the published literature. For example, meta-analytic techniques are valid only when restricted to specific exposures, populations and endpoints (e.g. it would be problematic to combine risk estimates from studies considering different causes of mortality). The process of identifying and selecting these studies is captured through the Cochrane Collaboration's PRISMA method. A PRISMA diagram describes the literature the researchers reviewed, and generally follows the same practices as a systematic review. The researchers evaluate each study against a pre-defined list of criteria, removing those studies that do not meet each criterion and clearly documenting their choice. The PRISMA diagram for this manuscript (Supplemental Figure 1) describes how we included studies that use a long-term exposure metric, were performed for North American cohorts, and report an outcome of risk for all-cause mortality. Additional techniques, including funnel plot asymmetry and "trim and fill" calculations and Begg and Egger tests can indicate the presence of publication bias due to a lack of literature identifying null effects. While these tests can indicate whether the literature the researchers identified is biased, they cannot definitively indicate whether the original literature review omitted key articles. However, the actual number of studies on which our meta-analysis can rely to characterize the C-R function is quite small relative to the universe of studies that have been published on air pollution research. Specifically, toxicological and clinical studies cannot be incorporated quantitatively into our meta-

analysis. Instead, we implicitly account for this literature when we consider the ISA causal determination. While the meta-analysis conducted for this paper may improve the estimates of the relative risk, it cannot provide evidence on whether the relationships observed are causal.

The IER can be characterized as a meta-analysis that offers an innovative step forward in the breadth of evidence that it can quantitatively consider to characterize and quantify the $PM_{2.5}$ C-R relationship. The IER analysis combined evidence from air pollution, passive smoking, and active smoking to assess the coherence of the evidence from these difference combustion sources and to describe the shape of the response function over a wider range of exposure levels than is covered by air pollution studies alone. As the authors explicitly acknowledge, their method requires a number of important assumptions that require further examination. However, by assembling and evaluating consistency of evidence from a broader range of studies and particulate exposures, they lend strength to causal inferences about the contributions to mortality from particulate exposures.

Finally, the validity of the results rests on both independent verification and peer review. Different methods are more easily verified. For example, the systematic review, meta-analysis, and IER methods clearly indicate the literature considered. The reader should be able to easily discern whether either approach has omitted important literature. Omitted literature in the ISA would affect the conclusions the document draws regarding causality. In the meta-analysis, this would affect the pooled risk estimate. In principle, meta-analyses should improve external validity over any single study estimate, as it pools risk estimates across multiple studies. However, if there are underlying systematic errors in the studies (e.g. problems with internal validity), then meta-analyses may not achieve this goal. For example, if most long-term $PM_{2.5}$ mortality studies misclassify population exposure such that they bias-low mortality risks, then the pooled estimate would also reflect this bias. The IER approach is able to incorporate more data, which could yield higher generalizability of the results. In principle, it is possible to test for whether the shape of the function is correct by running the appropriate cohort study in a part of the world that experiences these higher levels of $PM_{2.5}$. They also provide limited validation of the method

by comparing its predictions of relative risks of mortality from cardiovascular, respiratory and lung cancer with those observed in studies of exposures to high levels of total suspended particulate matter in 31 Chinese cities (Burnett et al., 2013).

Independent verification of the $PM_{2.5}$ expert judgments is more difficult because, although drawn from similar underlying literature as the other methods, the results can be interpreted as predictions of a future outcome that will not be observed for some time. In addition to the USEPA elicitation, a separate set of investigators elicited similar judgments about the C-R function from a different set of European experts - with one expert in common with the U.S. based elicitation (Cooke et al., 2007). While that study does not provide validation of the expert judgments in the Roman et al. (2008) study, the similarities between quantitative judgments suggests some agreement about the state of knowledge about the $PM_{2.5}$ mortality relationship.

3.2.2 How transparent are the methods and results?

Each method carries with it a series of assumptions—some explicit, some implicit—and understanding what they are and how they affect the results is critical to transparency. We review the degree to which the different methods allow the identification, justification and sensitivity of the key hypotheses, sources of data, models, and assumptions, and how well and fully the study characterizes the robustness and/or uncertainty in its results. We also focus on the way in which scientists are chosen or opt to participate in a study, the extent to which the study addresses potential cognitive biases. This has important implications for evaluating the validity of the results as well as for the communication and use of the results in policy applications.

All of these methods take steps to provide transparency, and all do reasonably well at identifying key judgments or assumptions about underlying conceptual frameworks, choices of data or studies, analytical methods, etc. In practice, however, there are important differences. For the expert elicitation, the experts were given a detailed interview protocol that has each expert work through a common base of evidence, a conceptual framework or system for evaluating it, and explain his or her rationale.

However, it is often difficult to know how each person selected, weighed and integrated evidence in coming to their final quantitative judgments. Experts can be challenged to explore how their views might change if key assumptions or data change, but it is harder to do than conventional sensitivity analysis except in cases where experts have well defined models in mind. The more explicit, mathematically based methods like meta-analysis and the IER analysis, to the extent that the key assumptions are laid out and evaluated, can be more transparent. Meta-analysis tends to follow a fairly well established quantitative approach, and in principle, it is straightforward to examine the sensitivity of the pooled estimate to any given study or approach. Similarly, the IER analysis clearly outlines assumptions about study selection and calculating the equivalent doses of $PM_{2.5}$ from different sources of exposure (ambient, passive and active smoking). More fundamentally, it assumes that the different sources of $PM_{2.5}$ operate via similar toxicological mechanisms so that they can be fit or explained by a common mathematical function. The sensitivity of the results to any these assumptions can be and was in many cases was tested in the IER papers (Pope et al., 2009, 2011; Burnett et al., 2013; and Shin et al., 2013 (in this series)).

We also ask whether the approach fully and explicitly characterizes the key sources, and the extent, of variability and uncertainty. Systematic review can describe to some degree the variability and uncertainty by arraying and comparing the quantitative risk estimates from individual studies, including estimates of standard error; however, the characterization of the impact of sources of uncertainty on individual or overall uncertainty in C-R function remains largely descriptive. It was partly for this reason that the NRC (2002) suggested that USEPA consider other methods, including elicitation of experts judgments, to develop a more comprehensive characterization of uncertainty in air pollution health benefits. Meta-analysis improves on the systematic review by using statistical techniques to combine the results from multiple studies, to quantifying overall uncertainty from a body of evidence and to explore possible factors that might explain the variability observed among studies. Ultimately, it can only reflect strengths and limitations of the underlying studies from which the analysis is conducted. Any biases or uncertainties (e.g. confounders, representativeness of populations studied, etc.) that are not captured by the risk estimates from those studies will not be reflected in the meta-analytic results. The

IER approach presented in this series has attempted to model uncertainty more fully and for the entire range of the C-R function. Similar to meta-analysis, it is also dependent on the strengths and limitations of the underlying studies, and the necessary assumptions made to combine them.

Even when the appropriate scientific disciplines are involved in the analysis, however, scientists may disagree on the interpretation of the same evidence. Additionally, it is not always clear whether a set of scientists represents a particular viewpoint or what the broader range of scientific opinion might be. The methods do not perform equally well on this attribute, although some of the differences that we observe among the four research synthesis methods with respect to involvement of particular scientists reflect fundamental differences in the origins of these particular applications to $PM_{2.5}$ and are not necessarily attributes of the methods themselves. The ISA and the expert elicitation are work products of the USEPA and are intended to support policy decisions. Thus, both processes make use of the well-established procedures to select the experts. The PM expert elicitation by design was intended to help decision-makers clearer about what it means "when different experts appear to be telling us different things" and ultimately whether it matters for the decision. Consequently, the selection of experts followed a specific protocol using publication counts and peer nomination for identifying experts with appropriate types of expertise who might represent a range of credible scientific opinions (see both Cooke et al., 2007 and Roman et al., 2008). The stakeholders reviewed the selection protocol; however, the implementation and selection of particular experts was intended to be independent of them. Similarly, the scientists who review the ISA as members of the USEPA Clean Air Scientific Advisory Committee (CASAC) are also chosen to ensure broad representation. There is also a public comment period that helps ensures that analysts who may not be identified or may not have time for committee service can also be heard. By contrast, the meta-analysis and IER reflect the more standard scientific process in which individual scientists or groups of scientists form collaborations based on institutional, technical, shared interests, or other factors that may not be explicit or acknowledged. Characterizing the diversity of scientific opinion is not the primary goal of these efforts, so it is likely to be expressed less within

these teams. Decision makers interested in perspective on potential diversity of scientific opinions will need to rely on systematic and peer reviews for such insights.

Finally, regardless of the experts and the selection process, it is well documented that we are all subject to a number of common cognitive biases, such as overconfidence and anchoring, which can influence the results as well as obscure judgments. Recognizing these limitations, the $PM_{2.5}$ expert elicitation explicitly trained participants in the relevant aspects of cognitive biases such as overconfidence and anchoring. The experts had multiple opportunities to revisit their estimates and account for these biases. In principle, cognitive biases could be considered in drawing inferences in a systematic review. However, the identification and training regarding these cognitive biases may be less explicit for systematic reviews, meta-analysis and IER. The statistical process strives to generate a less biased estimate. However, the researcher's choice of studies may reflect cognitive biases.

3.2.3 How suited is the method to the policy problem?

Selecting an appropriate method depends in part on how the researchers have defined and understand the policy problem. For the long-term $PM_{2.5}$ mortality C-R function, we ask whether the method can capture the following three features: the relevant range of exposures to ambient $PM_{2.5}$, the relevant health outcomes and the relevant populations that are required for the RIAs conducted the USEPA.

For systematic review, the relevant range of exposures for the USEPA is captured in the environmental epidemiological studies, published primarily in the U.S. and Canada. The systematic review process is flexible enough to consider the limited evidence for lower and potential threshold level $PM_{2.5}$ concentrations. Similarly, they systematically compile and assess information on all relevant outcomes. The ISA procedure is specifically well suited to identifying and evaluating all relevant literature and is able to match the attributes of the population at risk if there are existing studies of that population. However, this approach cannot compensate for an absence of studies, and it is not designed to yield a single quantitative estimate.

The meta-analysis pools risk estimates from studies that each account for different air pollution levels, health outcomes and population attributes, and so the final pooled risk estimate will also reflect these characteristics. While the meta-analysis cannot adjust for the attributes of this literature, measures of heterogeneity and meta-regression techniques can suggest whether certain variables help explain the size of risk estimates. For example, a meta-regression can indicate whether exposure levels accounts for some of the difference in the risk estimates across studies and locations. An important benefit of the IER approach is that it extends the type of data that can be quantitatively included in the C-R function. This method can synthesize studies with significantly higher exposure metrics and thus extend the C-R function to a wider range of ambient concentrations – specifically, those presently observed in Asia. This is the only method that can synthesize data from other studies to characterize the risk at higher $PM_{2.5}$ exposures. As long as there are available studies, IER can be used to characterize a range of mortality outcomes. Presently, Burnett et al. (2013) have developed integrated C-R functions for ischemic heart disease (IHD), cerebrovascular disease (stroke), chronic obstructive pulmonary disease (COPD), and lung cancer (LC). They also developed functions for the incidence of acute lower respiratory illness (ALRI) that can be used to estimate mortality and lost-years of healthy life in children less than 5 years old (Burnett et al., 2013).

Expert elicitation has substantial advantages when existing studies have not provided evidence that adequately addresses the policy problem. For the USEPA elicitation, the experts were conditioned on the U.S. studies and asked to consider the results to concentrations that are observed prior to, and as a result of, the relevant regulations. The experts also provided quantitative estimates of uncertainty about whether the C-R function remains linear at lower ambient concentrations despite the paucity of direct evidence, and the probability that there is a causal relationship between long-term $PM_{2.5}$ exposure and the risk of premature death. When considering the former question, expert elicitation allowed the participants to address a broader range of concentrations than those that were reported in the epidemiological literature. Thus, in theory, the experts could also be asked to extend their analysis to higher concentrations, although it may be more difficult to condition them with the relevant literature and map their thought process. Similarly, as

long there is a basis for drawing their judgments, the experts can give information on the relevant outcomes and relevant populations, using techniques such as analogy.

3.2.4 How suited is the method to the resources?

Policy analysts are often required to develop values under challenging time constraints and limited financial and technical resources. Consequently, the final set of criteria considers the resources associated with these four techniques. We evaluate both the level of resources required to conduct the analysis as well as the accessibility of the results and the ease of use for policy making.

For an analyst, there are substantial advantages to employing readily available data. It can both reduce the amount of time and resources to conduct the analysis and can make it easier to update the assessment to account for new information. All of these methods are designed to make use of the data in the published literature or other publicly available sources. The nature and extent of data required to make particular research synthesis methods possible, however, can differ substantially. Systematic reviews and expert elicitation can both reflect whatever information is available, however incomplete. In fact, structured expert elicitation has been intended primarily to help scientists and policy makers understand what to do when data are lacking or insufficient (e.g. what do experts agree on and where could additional data collection help with issues on which they appear to disagree?). In theory at least, the more data available that are adequate to the task at hand, the more one would expect expert judgment and analytical approaches to converge. If a sufficient number of appropriately similar studies exist, meta-analysis can be relatively easy to specify with risk estimates drawn from existing studies, particularly if these have already been identified through systematic reviews like the ISA. Additionally, while we employed a commercially produced meta-analysis software package, free software abounds (e.g. there are multiple free meta-analysis packages for the open source R statistical software). The IER approach is comparatively more resource intensive and currently, only a few scientists have applied it. While this technique also employs existing data, it draws from a broader literature base and more adjustments are required to unify the $PM_{2.5}$ concentrations across studies and to characterize uncertainty.

The ability to update an analysis as new information becomes available is also a valuable attribute. New studies are constantly being published, and the challenge for scientists and policy analysts is to decide whether or to what extent the new information changes or raises new questions the fundamental scientific inferences that have been drawn from the existing evidence. Systematic reviews, meta-analyses, and the IER analysis – to the extent that the methods and assumptions are explicitly and clearly laid out – are all reasonably updatable. For example, the meta-analysis for this manuscript required little additional effort to include risk estimates from studies published after the release of the 2006 expert elicitation. The expert elicitation is probably the most challenging to update over time. As described above, the $PM_{2.5}$ mortality literature has evolved significantly since the expert elicitation was completed. Several new cohort studies as well as other kinds of evidence were not yet published at the time of the original expert elicitation. Updating the original judgments would require going back to the individuals; alternatively, judgments from a newly chosen set of experts could be elicited. However, in the latter case especially, it would be more difficult to know whether any differences observed in the judgments would be due to the new data or to the new set of experts.

The fundamental question of whether any method is ultimately worth the cost of conducting it or obtaining new information to update it is more challenging to answer and is more appropriately answered through a more careful analysis of the value of information. We could not begin to undertake such an analysis here, although clearly, a number of factors including the costs of compliance with air pollution regulations have driven stakeholders to demand increasing improvements in the accuracy and precision of the estimated health benefits -- that is, in the strength and certainty in the C-R function.

3.3 Evaluating approaches in the context of a regulatory analysis

To illustrate the implication of these different estimates of the C-R function for a policy application, we apply the quantitative results from the different methods to the benefits analysis of the avoided $PM_{2.5}$-related premature deaths avoided due to air quality improvements achieved through implementation of the Mercury and Air Toxics Rule (USEPA, 2011). We also consider how the results are communicated to decision makers.

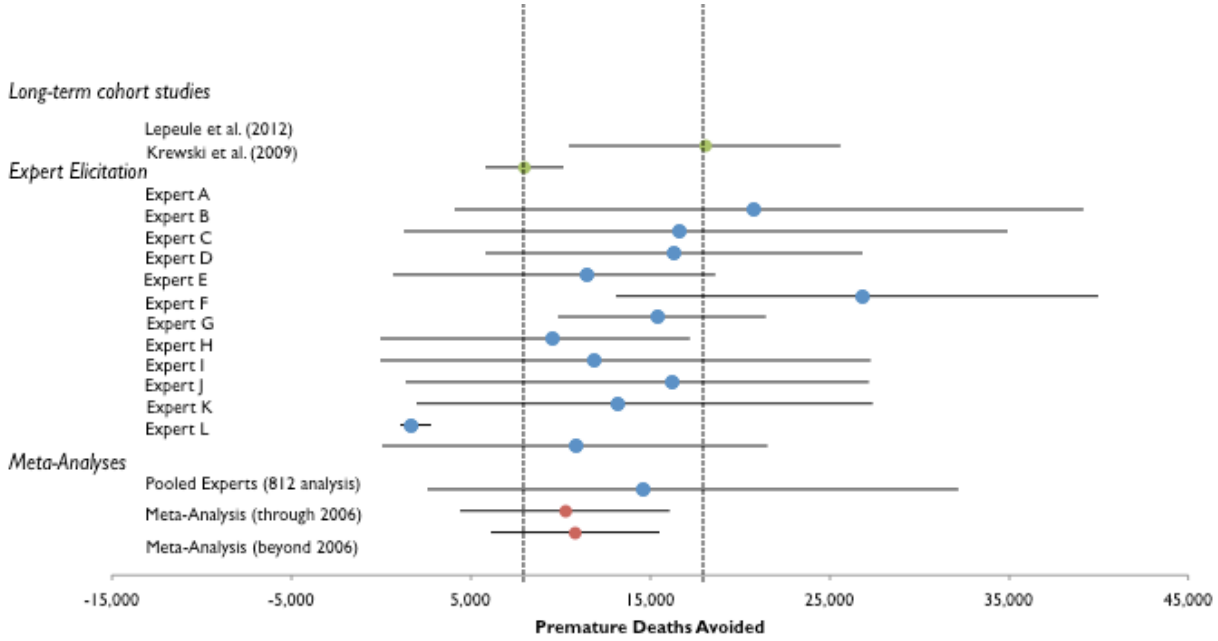
While we are comparing these methods at one point in time, the choice of methods has evolved over time in responses to changes in the amount and strength of evidence.

Following the basic method used by the RIAs conducted by the USEPA, each risk estimate (i.e. the percent change in risk per 10 $\mu g/m^3$) from the various research synthesis methods is translated into a mean value and 95% confidence intervals of premature deaths avoided by the reduction in ambient $PM_{2.5}$ concentrations. We use the air quality model results from the RIA for the Mercury and Air Toxics Rules and combine this with the population and incidence values from the environmental Benefits Mapping and Analysis Program—Community Edition (BenMAP-CE) v0.63 (USEPA, 2013). We conducted this analysis for four groups of risk estimates: 1) the two long-term cohort studies – the Health Effects Institute extended analysis of the ACS cohort (Krewski et al., 2009) and the extended analysis of the H6C study (Lepeule et al., 2012) – that, based on the systematic review of the literature and advice from science advisory board, are the primary estimates used by the USEPA for its RIAs; 2) for each of the 12 judgments elicited in the USEPA Expert elicitation (Roman et al., 2008); 3) from three pooled analyses, a pooled risk estimate of the 12 expert-derived risk values, drawn from the section 812 analysis of the costs and benefits of the Clean Air Act (USEPA, 2012), our meta-analysis of the long-term epidemiological literature that the experts considered in the 2006 elicitation, and our second meta-analysis of the 2006 literature and the additional studies identified in the 2009 ISA; and, 4) from the IER analysis. We show the mean and 95% confidence intervals or credible intervals in the case of the expert elicitation of the estimated avoided deaths in Figure 4. The vertical dotted lines delineate the mean numbers of deaths based on the Krewski et al. (2009) and Lepeule et al. (2012) studies.

These results show a remarkable level of consistency in the estimated number of avoided premature deaths. All but three of the experts' mean risk estimates fell between the ACS (Krewski et al., 2009) and H6C (Lepeule et al., 2012) based values, shown by the vertical dashed lines. This is consistent with the experts who indicated that they found those two studies the most informative to the central tendencies of their judgments. The results from the two meta-analyses yield means closer to that of the ACS analysis reflecting the larger

weight the pooling algorithm gives to studies with larger population and smaller variance. Since all of these studies are drawing on the same literature, one would expect that the mean values would show overall agreement.

Figure 4: Comparison of Premature Deaths Avoided from Different C-R functions for the Mercury and Air Toxics Rule



There is more variability in the bounds of uncertainty as well as what they represent. The 95% confidence intervals for ACS and H6C reflect the population sizes for the cohorts. For the meta-analyses, the confidence intervals suggest a greater precision in the estimates gained from the pooling of information from multiple studies. By contrast, the experts provided 95% credible intervals that reflect the inter-individual variability in each expert's levels of confidence in the causal nature of the relationship and in their own judgments about the existing data. These credible intervals are wider than the statistically based confidence intervals to accommodate these different sources of uncertainty (see Roman et al. 2008 for further discussion of those uncertainties). This is especially apparent for Experts G, H and L who incorporated their judgments about the likelihood that the relationship between exposure to $PM_{2.5}$ and premature mortality was causal directly into their credible intervals. The pooled estimate of the 12 experts' judgments where equal

weight is essentially given to each of the experts. This suggests a similar overall uncertainty of the group.

The juxtaposition of the results using these different approaches to research synthesis illustrates some of the strengths and weaknesses with respect to interpretability and accessibility of the risk estimates to a policy maker. The first set of individual study results, reflect the USEPA's historical approach where the primary emphasis was placed on the large, nationally based ACS study with the H6C study was used as a sensitivity analysis. Reliance on a single estimate may be simpler and easy to communicate. However, using the LePeule et al. (2012) study as a sensitivity analysis, however, is more difficult to interpret since there is no specific guidance on what emphasis the sensitivity analysis should receive or what percentile of the population distribution of risk it might represent. This is of particular concern as the result appears quite different from the larger ACS study. The meta-analysis addresses these concerns by weighting the different studies by population and providing uncertainty estimates based on this pooled sample.

The results from the experts' judgments provide more complexity. On the one hand, presentation of the individual 12 expert-derived C-R functions make it quite clear what the similarities and differences are in the way a range of experts view the full body of evidence and what the. It may also more realistically reflect the nature of scientific opinion before consensus emerges and allow the policy maker to explore what the implications of those differences might be for the policy analysis. On the other hand, this volume of information may be more challenging to process. Specifically, a decision maker may desire greater guidance, in particular certainty, from the science.

The estimate for the IER approach is not available at this time, but it will be presented at the October workshop. It will also be cause-specific not total all-cause mortality.

4. CONCLUSIONS

Four research synthesis techniques have been employed over the past two decades to address what is essentially the same policy question—namely, what is the long-term $PM_{2.5}$ all-cause mortality relationship? The availability of four well-designed and documented research synthesis methods provides a unique opportunity to evaluate and compare these approaches. Overall, the most appropriate method will depend largely on practical constraints, including time, expertise, tolerance for uncertainty and financial resources. However, researchers bound by these constraints, and the decision-makers posing the policy question, will in turn want to evaluate carefully the extent to which each method satisfies the four key sets of criteria we discussed in depth above: validity, transparency, policy relevance and resource efficiency.

We find that the methods are not mutually exclusive. Systematic and critical reviews are a necessary prerequisite for any quantitative method. Other papers in this series will talk more about this process and will debate what exactly systematic review should mean. However, it is clear that an understanding and evaluation of the multidisciplinary evidence is an essential starting point for any research synthesis method. For $PM_{2.5}$, this review has informed both qualitative and quantitative assessments of the likelihood of a causal relationship between exposure and mortality as well as determines what studies the USEPA choses to represent the likely US population risk, either singly or collectively in meta-analyses.

Decision makers can then select from the other three methods for quantitative estimates. We find that each technique can perform well with respect to generating a valid and transparent estimate, but their ranking should be evaluated according to the state of knowledge regarding the $PM_{2.5}$ C-R function at the time the synthesis was performed. This is highlighted in the comparison of our meta-analysis and the expert elicitation. In the lead-up to the 2006 expert elicitation, there was limited or no direct evidence to adequately establish the quantitative likelihood of a causal relationship, the shape of the population C-R function (e.g. threshold effects), or the scientific state of uncertainty in the magnitude of

the mean C-R function (NRC 2002). Here, the expert elicitation was particularly useful for characterizing the basis for and the degree of agreement and disagreement among different scientific experts. Since that time and as reflected in the 2009 ISA, the epidemiological, toxicological and clinical literature has grown and shows a convergence on the causal nature and the magnitude of the $PM_{2.5}$ mortality relationship. Partly for this reason, we believe that the quantitative meta-analysis and the IER approach —which draw upon this growing base of literature—are also capable of providing valid estimates. Our illustrative example of a meta-analysis highlights the analytical appeal of these studies, although only relatively recently have been enough studies to employ this approach (e.g. Hoek et al, 2013). We also caution that differences in comparability among these studies may make the analysis challenging to conduct and interpret; consequently, meta-analysis has not been used for formal policy purposes. The IER approach represents potential substantial improvement by incorporating more data and shows promise in early comparisons; however, it has not been fully validated, and we may need to accept that its biological or mechanistic basis may not be understood for some time.

Ultimately, all of these methods require the considerable judgment of scientists, individually and collectively. Regardless of the amount of data available, the expert elicitation through its systematic approach to characterizing expert opinion can play an important role in enhancing transparency. Finding ways for all these methods to acknowledge, appropriately elicit and examine the implications of that judgment would be a step forward.

5. REFERENCES

ALPERT M. & RAIFFA H. 1969. A progess report on the training of probability assessors. *Judgement under uncertainty: heuristics and biases* 294-305.

ANDERSON H.R., ATKINSON R.W., PEACOCK J.L., SWEETING M.J. & MARTSON L. 2005. Ambient particulate matter and health effects: Publication bias in studies of short-term associations. *Epidemiology* 16: 155-163.

BELL M.L., DOMINICI F. & SAMET J.M. 2005. A meta-analysis of time-series studies of ozone and mortality with comparison to the national morbidity, mortality and air pollution study. *Epidemiology* 16: 436-445.

BORENSTEIN M., HEDGES L., HIGGINS J. & ROTHSTEIN H. 2005. Comprehensive meta-analysis version 2. *Englewood, NJ: Biostat*

BURNETT R.T., POPE III C.A., EZZATI M., OLIVES C., LIM S.S., MEHTA S., SHIN H.H., SINGH G., HUBBELL B.J., BRAUER M., ANDERSON H.R., SMITH K.R., KAN H., LADEN F., PRUESS A., TURNER M.C., THUN M. & COHEN A. 2013. An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. *Under review*

COOKE R.M. 1991. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press, USA.

COOKE R.M., WILSON A.M., TUOMISTO J.T., MORALES O., TAINIO M. & EVANS J.S. 2007. A probabilistic characterization of the relationship between fine particulate matter and mortality: Elicitation of European experts. *Environmental science & technology* 41: 6598-6605.

DOCKERY D.W., POPE C.A., XU X., SPENGLER J.D., WARE J.H., FAY M.E., FERRIS JR B.G. & SPEIZER F.E. 1993. An association between air pollution and mortality in six US cities. *New England journal of medicine* 329: 1753-1759.

HIGGINS J.P.T. & GREEN S. 2011. Cochrane Handbook for Systematic Reviews of Interventions: The Cochrane Collaboration, Version 5.1.10, http://handbook.cochrane.org/.

HOEK, G., KRISHNAN, R.M., BEELEN, R., PETERS, A., OSTRO, B., BRUNEKREEF, B., KAUFMAN, J.D. 2013. Long-term air pollution exposure and cardiovascular respiratory mortality: A review. *Environmental Health* 12(1): 43

INDUSTRIAL ECONOMICS INCORPORATED 2006. Expanded expert judgment assessment of the concentration-response relationship between PM2.5 exposure and mortality. *http://www.epa.gov/ttn/ecas/regdata/Uncertainty/pm_ee_report.pdf*

KREWSKI D., M. JERRETT, R.T. BURNETT, R. MA, E. HUGHES, Y. SHI, M.C. TURNER, C.A. POPE III, G. THURSTON & E.E. CALLE 2009. *Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality*. Health Effects Institute Cambridge, MA, USA.

LADEN F., SCHWARTZ J., SPEIZER F.E. & DOCKERY D.W. 2006. Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities study. *American Journal of Respiratory and Critical Care Medicine* 173: 667.

LEPEULE J., LADEN F., DOCKERY D. & SCHWARTZ J. 2012. Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities Study from 1974 to 2009. *Environmental health perspectives* 120: 965.

LICHTENSTEIN S., B. FISCHHOFF & L.D. PHILLIPS 1977. *Calibration of probabilities: The state of the art*. Springer.

MOHER D., TETZLAFF J., TRICCO A.C., SAMPSON M. & ALTMAN D.G. 2007. Epidemiology and reporting characteristics of systematic reviews. *PLoS Medicine* 4: e78.

MORGAN M.G. & M. HENRION 1992. *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press.

NATIONAL RESEARCH COUNCIL (US) 2002. *Committee on Estimating the Health-Risk-Reduction Benefits of Proposed Air Pollution Regulations, Board on Environmental Studies. Estimating the public health benefits of proposed air pollution regulations*. National Academy Press.

OFFICE OF MANAGEMENT AND BUDGET (OMB) 2013. Draft Report to Congress on the Benefits and Costs of Federal Regulations and Agency Compliance with the Unfunded Mandates Reform Act. *Available on the Internet at*

http://www.whitehouse.gov/sites/default/files/omb/inforeg/2013_cb/draft_2013_cost_benefit_report.pdf

POPE III C.A., BURNETT R.T., TURNER M.C., COHEN A., KREWSKI D., JERRETT M., GAPSTUR S.M. & THUN M.J. 2011. Lung cancer and cardiovascular disease mortality associated with ambient air pollution and cigarette smoke: shape of the exposure–response relationships. *Environmental health perspectives* 119: 1616.

POPE III C.A., BURNETT R.T., THUN M.J., CALLE E.E., KREWSKI D., ITO K. & THURSTON G.D. 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA: the journal of the American Medical Association* 287: 1132-1141.

POPE III C.A. & DOCKERY D.W. 2006. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air & Waste Management Association* 56: 709-742.

POPE C.A., BURNETT R.T., KREWSKI D., JERRETT M., SHI Y., CALLE E.E. & THUN M.J. 2009. Cardiovascular mortality and exposure to airborne fine particulate matter and cigarette smoke shape of the exposure-response relationship. *Circulation* 120: 941-948.

POPE C.A., THUN M.J., NAMBOODIRI M.M., DOCKERY D.W., EVANS J.S., SPEIZER F.E. & HEATH C.W. 1995. Particulate air pollution as a predictor of mortality in a prospective study of US adults. *American journal of respiratory and critical care medicine* 151: 669-674.

ROMAN H.A., WALKER K.D., WALSH T.L., CONNER L., RICHMOND H.M., HUBBELL B.J. & KINNEY P.L. 2008. Expert judgment assessment of the mortality impact of changes in ambient fine particulate matter in the US. *Environmental science & technology* 42: 2268-2274.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (USEPA) 1997. The Benefits and Costs of the Clean Air Act, 1970 to 1990. *Available on the Internet at http://www.epa.gov/cleanairactbenefits/1970-1990/chptr1_7.pdf* 34.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (USEPA) 1999. Final Tier 2 Rule: Air Quality Estimation, Selected Health and Welfare Benefits Methods, and Benefits Analysis Results. EPA 420-R-99-032. Office of Air Quality Planning and Standards, US Environmental Protection Agency, Research Triangle Park, NC.

Available on the Internet at http://www.epa.gov/otaq/regs/ld-hwy/tier-2/frm/tsd/r99032.pdf

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (USEPA) 2004. Final Regulatory Analysis: Control of Emissions from Nonroad Diesel Engines, EPA 420-R-04-007. Office of Transportation and Air Quality, Washington DC. *Available on the Internet at http://www.epa.gov/nonroad-diesel/2004fr/420r04007a.pdf*

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (USEPA) 2005. Regulatory Impact Analysis for the Final Clean Air Interstate Rule. EPA-452/R-05-002. Office of Air and Radiation, Washington DC.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (USEPA) 2009. Integrated science assessment for particulate matter. *US Environmental Protection Agency Washington, DC*

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (USEPA) 2011. Regulatory Impact Analysis for the Final Mercury and Air Toxics Standards. EPA-452/R-11-011. December. *Available on the Internet at http://www.epa.gov/ttn/ecas/regdata/RIAs/matsriafinal.pdf*

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (USEPA) 2012. Regulatory Impact Analysis for the Final Revisions to the National Ambient Air Quality Standards for Particulate Matter. *Available on the Internet at http://www.epa.gov/ttnecas1/regdata/RIAs/finalria.pdf*

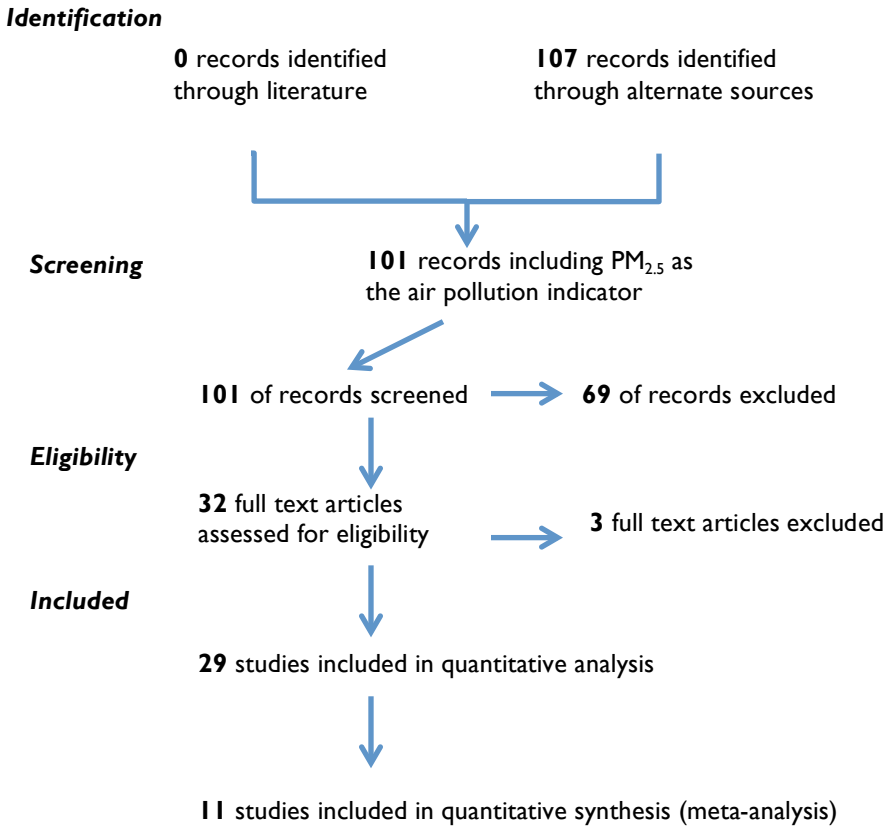
UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (USEPA) 2013. Environmental Benefits Mapping and Analysis Program - Community Edition (Version 0.xx). Research Triangle Park, NC. Available on the Internet at http://www.epa.gov/air/benmap/beta.html

WALKER K.D., CATALANO P., HAMMITT J.K. & EVANS J.S. 2003. Use of expert judgment in exposure assessment: Part 2. Calibration of expert judgments about personal exposures to benzene. *Journal of Exposure Science and Environmental Epidemiology* 13: 1-16.

WALKER K.D., EVANS J.S. & MACINTOSH D. 2001. Use of expert judgment in exposure assessment. Part I. Characterization of personal exposure to benzene. *Journal of exposure analysis and environmental epidemiology* 11: 308.

Supplementary Material for "Characterizing the long-term $PM_{2.5}$ concentration response function: Comparing the strengths and weaknesses of alternate research synthesis approaches" by Neal Fann, Elisabeth Gilmore and Katherine Walker

---



Study	Cohort (population age)	Study period	Hazard Ratio (per 10 ug/m3, 95% confidence interval)	Stage of meta-analysis
Studies considered in 2006 Expert Elicitation				
Dockery et al. (1993)	Six Cities (age >24)	1975-1989	1.13 (1.04—1.23)	Stage 1
Enstrom et al. (2005)	11 California counties (>42)	1973-2002	1.007 0.991—1.023)	Stage 1 & 2
Jerrett et al. (2005)	ACS/Los Angeles (>29)	1982-2000	1.11 (0.99—1.25)	Stage 1
Lipfert et al. (2006)	32 veterans clinics (>48)	1989-1996	1.15 (1.05—1.26)	Stage 1 & 2
McDonnell et al. (2000)	Adventist California (>26)	1997-1992	1.09 (0.98—1.2)	Stage 1 & 2
Pope et al. (2002)	ACS/51 cities (>29)	1979-2000	1.06 (1.02—1.11)	Stage 1
Studies published after 2006 Expert Elicitation				
Eftim et al. (2008)	Medicare/ACS 50 cities	2000-2002	1.08 (1.05—1.18)	Stage 2
Krewski et al. (2009)[A]	ACS/151 cities (>29)	1979-2000	1.06 (1.04—1.08)	Stage 2
Lepeule et al. (2012)	Six Cities (>24)	1974-2009	1.14 (1.07—1.22)	Stage 2
Ostro et al. (2010)[B]	California teachers (>29)	2002-2007	1.1 (0.93—1.28)	Stage 2
Zeger et al. (2008)	Medicare (>64)	2000-2005	0.989 to 1.132[C]	Stage 2
Puett et al. (2009)	Nurses/11 US states (>54)	1992-2002	1.26 (1.02—1.54)	Stage 2
Puett et al. (2011)	Health professionals (40-75)	1986-2010	0.86 (0.7—1.05)	Stage 2

[A] Selected results from model that accounted for 44 individual and 7 community-level variables
[B] Selected results from erratum
[C] Range reflects results from Western U.S. (low end) and Eastern U.S. (high end)