AN EXPERT JUDGMENT ASSESSMENT OF THE CONCENTRATION-RESPONSE RELATIONSHIP BETWEEN $PM_{2.5}$ EXPOSURE AND MORTALITY

Prepared for:

Office of Air Quality Planning and Standards
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

Prepared by:

Industrial Economics, Incorporated
2067 Massachusetts Avenue
Cambridge, MA 02140

Under Subcontract to:

Abt Associates, Incorporated
Hampden Square, Suite 600
4800 Montgomery Lane
Bethesda, MD 20814

April 23, 2004

EXECUTIVE SUMMARY

The effect of changes in ambient fine particulate matter ($PM_{2.5}$) levels on mortality constitutes a key component of the U.S. Environmental Protection Agency's (EPA's) approach for assessing potential health benefits associated with air quality regulations targeting emissions of $PM_{2.5}$ and its precursors. Avoided premature deaths constitute, on a dollars basis, between 85 and 95 percent of the benefits reported in EPA's retrospective and prospective Section 812A benefit-cost analyses of the Clean Air Act (U.S. EPA, 1997 and 1999) and in Regulatory Impact Analyses (RIAs) for rules such as the Heavy Duty Diesel Engine/Fuel Rule (U.S. EPA, 2000). Uncertainties regarding the mortality effects of $PM_{2.5}$ exposure could have a significant impact on the range of plausible benefit values associated with air pollution regulations and on the interpretation of the results of benefit analyses.

A recent National Research Council (NRC) report (NRC, 2002), *Estimating the Public Health Benefits of Proposed Air Pollution Regulations*, presented the results of an NRC review of EPA's benefits assessment methods for evaluating proposed regulations of air pollutants. The NRC committee approved of EPA's benefits analysis framework but provided a number of recommendations for improving the implementation of that framework, including recommendations for improving uncertainty analysis for benefits. The committee recommended that EPA take steps to improve its characterization of uncertainties associated with key components of its health benefits analyses, such as PM-related mortality effects. The NRC report recommended that probability distributions for key sources of uncertainty be developed using available data or, where scientific data are limited or conflicting, using probabilistic distributions obtained through formal elicitation of expert judgments.

In response to the NRC recommendations, EPA is exploring how it might incorporate expert judgment in policy analysis. As a first step in this direction, IEc worked with EPA and OMB scientists to design a pilot expert elicitation to characterize the uncertainty in the ambient $PM_{2.5}$/mortality relationship. This pilot was designed to provide EPA with an opportunity to improve its understanding of the design and application of expert elicitation methods to economic benefits analysis. For instance, the pilot was designed to provide feedback on the efficacy of the protocol developed and the analytic challenges, as well as to provide insight regarding potential implications of the results on the degree of uncertainty surrounding the C-R function for $PM_{2.5}$ mortality. The scope of the quantitative questions was limited in that we focused the elicitation on the C-R function of PM mass; this initial elicitation was not intended to thoroughly characterize the uncertainty surrounding individual elements of the $PM_{2.5}$/mortality relationship, such as the relative toxicity of specific PM components (e.g., diesel particulates).

Over the next year, EPA may plan a second phase of this project. That phase may involve a more comprehensive elicitation that could specifically evaluate the contribution of individual factors to the magnitude and uncertainty of the $PM_{2.5}$ mortality relationship.

The pilot elicitation conducted by IEc for EPA consisted of personal interviews with five experts, who were selected from the membership lists of two PM-related National Research Council (NRC) committees via a peer nomination process. The size of the pilot expert panel was

dictated by time and resource constraints, and the decision to restrict the initial expert pool to the NRC committees was made to help expedite the expert selection process. The rosters of both NRC committees included recognized experts in pertinent fields such as epidemiology and toxicology who had already undergone extensive review of their qualifications by the NRC, producing a reasonable initial list of experts likely to meet our expert selection criteria. The peer nomination process served to minimize the influence of the analysts on expert selection.

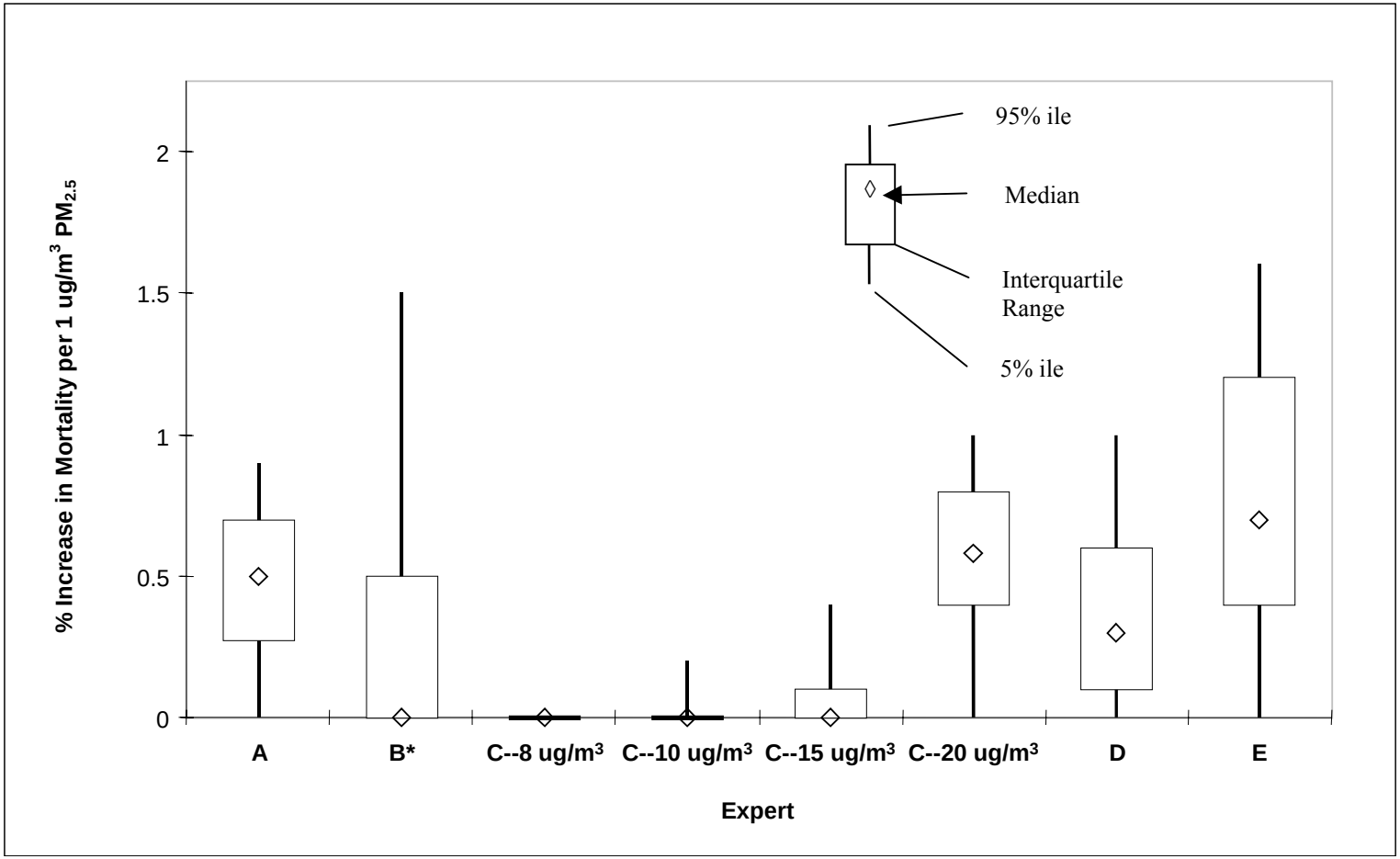
Experts were provided a briefing book of reference materials and a copy of the elicitation protocol prior to the interviews. Each interview lasted approximately 6-8 hours and covered both qualitative questions about various aspects of the $PM_{2.5}$/mortality concentration-response (C-R) relationship (e.g., nature of the mechanism, likelihood of causality) and quantitative questions asking experts to estimate the percent changes in mortality associated with specific changes in long-term and short-term exposure to $PM_{2.5}$. Specifically, they were asked to estimate: 1) the percent change in annual non-accidental mortality associated with a 1 $\mu g/m^3$ change in annual average $PM_{2.5}$ (long-term exposure to baseline concentrations ranging from 8 to 20 $\mu g/m^3$); and 2) the percent change in daily non-accidental mortality associated with a 10 $\mu g/m^3$ change in daily 24-hour average $PM_{2.5}$ (short-term exposure to baseline concentrations ranging from background to 60 $\mu g/m^3$). For each type of exposure, each expert provided minimum, maximum, and median estimates, plus 5^{th}, 25^{th}, 75^{th}, and 95^{th} percentile values for the distribution describing his uncertainty in the mortality effect of the specified change in $PM_{2.5}$. IEc both reported the judgments of individual experts and combined all five judgments into a single distribution, based on equal weighting of expert responses.

Figures ES-1 and ES-2 display the responses of the experts to the quantitative elicitation questions for the mortality effects of changes in long-term and short-term $PM_{2.5}$ exposures, respectively. The distributions provided by each expert, identified by the letters A through E, are depicted as boxplots with the diamond symbol showing the median (50th percentile), the box defining the interquartile range (bounded by the 25th and 75th percentiles), and the whiskers defining each expert's 90 percent confidence interval (bounded by the 5th and 95th percentiles of the distribution).

As illustrated by both these figures, the experts exhibited considerable variation in both the median values they reported and in the spread of uncertainty about the median. In response to the question concerning the effects of changes in long-term exposures to $PM_{2.5}$, the median value ranged from values at or near zero to a 0.7 percent increase in annual non-accidental mortality per 1 $\mu g/m^3$ increase in annual average $PM_{2.5}$ concentration. The variation in the experts' responses regarding the effects of long-term exposures largely reflects differences in their views about the degree of uncertainty inherent in key epidemiological results from long-term cohort studies, the likelihood of a causal relationship, and the shape of the concentration-response (C-R) function.

Figure ES-1

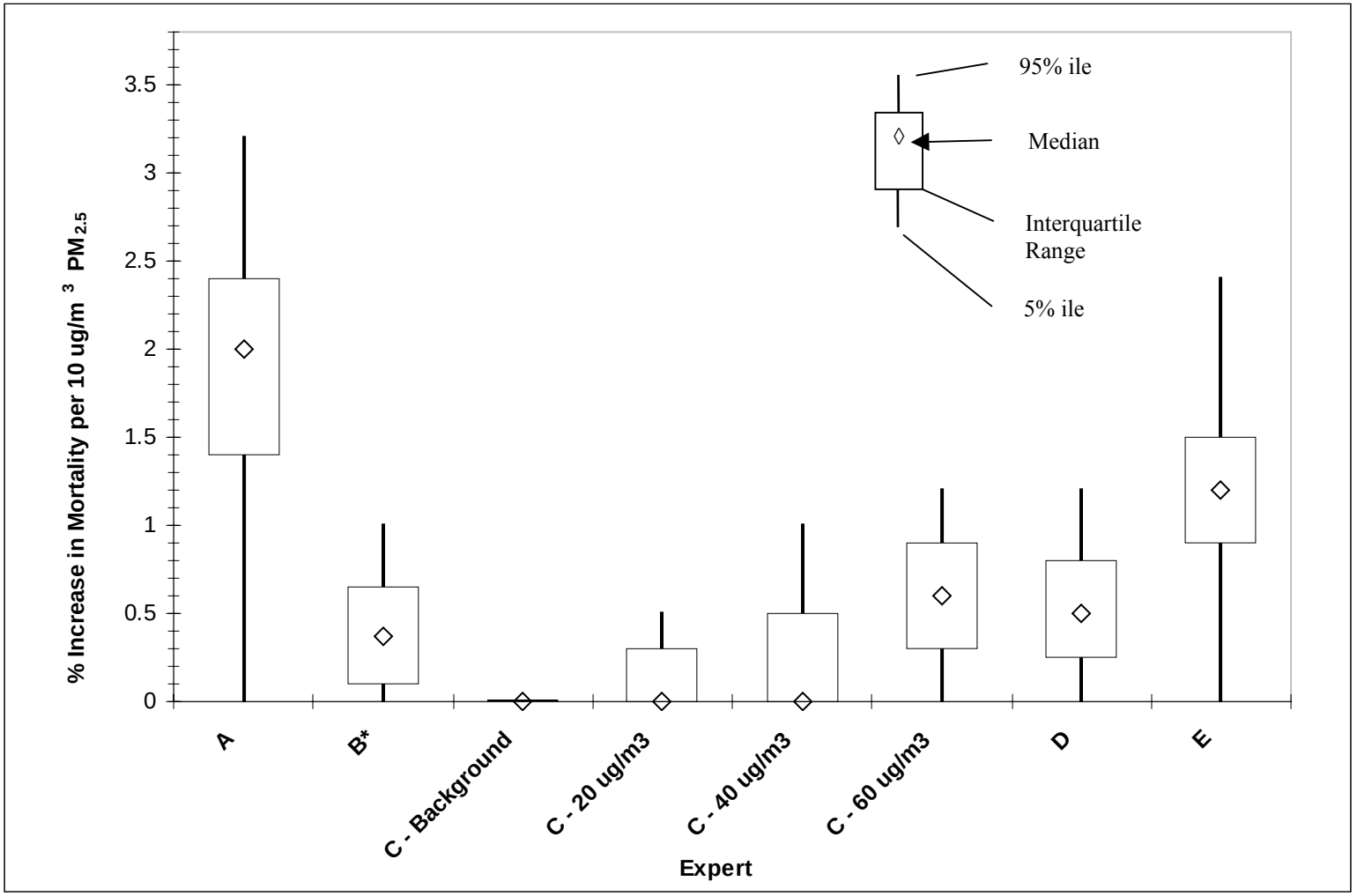
Comparison of Experts' Judgments about the Percent Increase in Annual Non-Accidental Mortality Associated with a 1 μg/m^3 Increase in Annual Average Exposures to $PM_{2.5}$ (U.S. Baseline 8 to 20 μg/m^3)



*Expert B specified this distribution for the PM/mortality coefficient above an uncertain threshold which he characterized as ranging between 4 and 15 with a modal value of 12 ug/m^3

Figure ES-2

Comparison of Experts' Judgments about the Percent Increase in Daily Non-Accidental Mortality Associated with a One-day 10 μg/m^3 Increase in 24-hour Average $PM_{2.5}$ Concentration (U.S. Baseline: Background to 60 μg/m3)



* Expert B specified this distribution for the PM/mortality coefficient above an uncertain threshold which he characterized as ranging between 4 and 25 with a modal value of 15 ug/m^3.

The experts' non-zero responses for the percent change in annual mortality associated with long-term exposures were mostly influenced by two cohort-based epidemiological studies: the Krewski et al., (2000) reanalysis of the original ACS cohort study, and the later Pope et al. (2002) update of the ACS study, with additional years of follow-up.

Three of the five experts gave distributions more heavily weighted towards zero. Those experts also gave the lowest probability of a causal effect of long-term exposure to $PM_{2.5}$ in the preliminary questions. All of the experts placed at least a 5 percent probability on the possibility that there is no causal relationship between fine PM exposure and mortality. Doubts about causality arose both from questions about the strength of the mechanistic data and from uncertainties in the epidemiological data.

Finally, three experts (A, D, and E) assumed that the function relating mortality with annual mean $PM_{2.5}$ concentrations would be log-linear over the specified range. The other two experts provided more complex responses. Expert B assumed a population threshold in his model, below which there would be no effect of increased $PM_{2.5}$ exposure and above which the relationship would be log-linear. Because he felt location of the threshold was uncertain, he specified a separate uncertainty distribution for the threshold concentration, in addition to his distribution for the slope of the C-R function above the threshold. Expert C expected there to be a declining effect on mortality with decreasing levels of $PM_{2.5}$, as well as some practical concentration threshold below which we would not observe any increase in mortality. He reflected these beliefs by developing a non-linear model within the specified $PM_{2.5}$ range, specified using separate distributions for the slope of the curve at four separate $PM_{2.5}$ concentrations.

As seen in Figure ES-2, the responses to the question asking experts to describe the mortality effects of changes in short-term $PM_{2.5}$ exposure showed more variation than the results for the effects of long-term exposure. The median estimates ranged from zero percent (no change) to a 2 percent increase in non-accidental mortality per 10 $\mu g/m^3$ increase in daily 24-hour average $PM_{2.5}$ concentration. The same types of factors that strongly affected each individual expert's uncertainty distribution for the effects of long-term PM exposures influenced their central tendency and spread for the mortality effects of short-term exposures (key epidemiological results from time-series studies, the strength of the causal relationship, and the nature of the C-R function).

On the issue of short-term mortality effects, the experts could be separated into two groups: those who relied primarily on results from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS; Dominici et al., 2003a), and those who relied primarily on other studies. Three experts (B, D, and E) relied primarily on NMMAPS data but differed in their decisions about adjusting the results to account for distributed lag effects. Of the two remaining experts, one (Expert C) relied primarily on the Canadian Eight City study (Burnett et al. 2000; Burnett and Goldberg, 2003) while the other (Expert A) used all of these studies as well as results from Air Pollution and Health: A European Approach 2 (APHEA2); Samoli et al., (2003) and Schwartz (2003).

The experts' distributions for short-term effects of $PM_{2.5}$ exposure on mortality reflected their views on the likelihood of a causal relationship between short-term exposures to $PM_{2.5}$ and

increased mortality. Although most of the experts assigned a greater probability to the likelihood of increased mortality resulting from short-term exposures than they did for long-term exposures, all of the experts again placed at least a five percent probability on the true value of the effect being zero. The experts specified C-R functions for short-term exposure effects that mirrored those for long-term exposure effects. A, D, and E assumed that the C-R function was log-linear across the entire range specified in the question. Experts B and C again specified different functional forms for the C-R relationship than the other three experts, with Expert B specifying an uncertain threshold and a log-linear relationship above the threshold, and C specifying distributions at four $PM_{2.5}$ concentrations to represent a non-linear C-R function.

Based on an *a priori* design decision made by the EPA and OMB scientists, used an equal weighting approach to develop combined distributions describing the premature mortality effects of long-term and short-term exposures. (The schedule of this analysis precluded efforts to explore alternative means of weighting expert's responses.) Calculation of these combined results necessitated developing distributions for Experts B and C that could be averaged with the results provided by the other experts. These distributions were based on the expected values of Expert B's and C's responses over the range of PM concentrations specified in the elicitation questions, and are sensitive to the assumptions used to calculate those expected values. As a result, the combined expert judgment distributions presented in this study are illustrative of the panel's collective uncertainty, but are not a direct replacement for the judgments of the individual experts and thus should not be used as direct input to policy analyses. The preferred approach for benefits analysis would be to use each individual expert's uncertainty distribution for the $PM_{2.5}$/mortality C-R coefficient to generate separate benefits estimates and then pool the results using equal weights.

The quantitative results of this pilot study reflect clearly the influence of the primary epidemiological studies the experts used in forming their judgments. However, a comparison of the combined distribution with the confidence intervals reported in those studies suggests that experts did incorporate additional sources of uncertainty besides statistical error in their predictions of the mortality effects of long-term and short-term $PM_{2.5}$ exposure. For example, the estimated variance of the combined expert distribution for effects of long-term exposure is nearly three times that of the pooled estimated of variance based an inverse variance weighting of the most recent analyses of the Six-Cities and ACS data.

The results of the pilot study suggest that a carefully constructed, comprehensive expert judgment study of the relationship between PM exposures and mortality can be helpful for characterizing the uncertainty in this component of benefits analysis for PM-related air pollution regulations. The experts in this study included a range of expertise and diverse opinions on issues such as causality, thresholds, C-R functions, and the importance of confounders. In addition, sensitivity analyses conducted on the combined results for both types of exposures suggest that no one individual's judgments dominated the combined distribution. The protocol questions covered most of the major issues the experts found relevant to developing their quantitative responses and was flexible enough to allow for consideration of additional issues raised by the experts, such as the effect of a distributed lag on estimates of the mortality effect of short-term exposures. The protocol was also flexible enough to allow experts to specify alternative forms of the C-R function for the mortality effects of $PM_{2.5}$ exposures.

The elicitation process also yielded several useful insights about applying an expert judgment approach in this context and identified some key areas on which to focus in subsequent efforts. In particular, a future elicitation study would benefit from a more structured and disaggregated approach to eliciting judgments that more clearly links the experts' opinions on key data and sources of uncertainty to their final probability distributions. This study employed an aggregate approach, asking directly about the final distribution. At the same time, the experts were asked to incorporate into their estimates any uncertainty associated with numerous complex issues (e.g. mechanisms, confounding, PM components, co-pollutants, exposure errors and misclassification). This constituted a challenging task for the experts. Also, future protocols should clearly specify all key assumptions underlying the questions, in order to minimize confusion and ensure that experts have a common understanding of the information being elicited.

In addition, future elicitation projects should focus on identifying a panel size likely to be representative of the breadth of respected scientific opinion on this issue. Although the decision analysis field tends to use relatively small sample sizes (typically 5-10 experts), and we selected the panel to reflect a diversity of opinions, it is possible that the expert panel interviewed for this pilot may not be fully representative of the distribution of relevant expert opinions on the question of interest. A larger sample size may provide better representation of the range of opinions on this topic both across and within different disciplines.

Though resource constraints precluded IEc from conducting a workshop prior to conducting the interviews for this pilot elicitation exercise, we believe that holding a workshop with the experts prior to the individual elicitations would be an important part of any future elicitation exercise. For example, development of a more structured approach could be facilitated through a workshop at which the participating experts could attempt to come to agreement on the structure or set of structures that might be used. In addition, a workshop could provide the experts with a more thorough introduction to expert judgment elicitation. It could be used to discuss and critique the vast literature the experts are expected to reference and to suggest important elements of the briefing book that will be provided to experts, such as supplemental data summaries or analyses addressing particular sources of uncertainty or bias. Furthermore, the workshop would provide an opportunity for EPA to clearly communicate to the experts the role and potential importance of the elicitation results for future regulation, information that would help increase the level of expert engagement in the process.

Finally, EPA should evaluate the feasibility of developing and validating a set of questions to help assess the accuracy and precision of the experts' judgments about the $PM_{2.5}$/mortality relationship. We recognize that identifying a set of independent questions that will be suitable for testing knowledge across all pertinent technical disciplines constitutes a significant challenge. However, without additional research into how to compare the quality of experts, questions will remain about how much confidence to place in the individual judgments obtained.

TABLE OF CONTENTS

APPENDICES

TABLES

FIGURES

1. INTRODUCTION

The effect of changes in ambient fine particulate matter ($PM_{2.5}$) levels on mortality constitutes a key component of the U.S. Environmental Protection Agency's (EPA's) approach for assessing potential health benefits associated with air quality regulations targeting emissions of $PM_{2.5}$ and its precursors. Avoided premature deaths constitute, on a dollars basis, between 85 and 95 percent of the benefits reported in EPA's retrospective and prospective Section 812A benefit-cost analyses of the Clean Air Act (U.S. EPA, 1997 and 1999) and in Regulatory Impact Analyses (RIAs) for rules such as the Heavy Duty Diesel Engine/Fuel Rule (U.S. EPA, 2000). Uncertainties regarding the mortality effects of $PM_{2.5}$ exposure could have a significant impact on the range of plausible benefit values associated with air pollution regulations and on the interpretation of the results of benefit analyses.

A recent National Research Council (NRC) report (NRC, 2002), *Estimating the Public Health Benefits of Proposed Air Pollution Regulations*, presented the results of an NRC review of EPA's benefits assessment methods for evaluating proposed regulations of air pollutants. The NRC committee approved of EPA's benefits analysis framework but provided a number of recommendations for improving the implementation of that framework, including recommendations for improving uncertainty analysis for benefits. The committee recommended that EPA take steps to improve its characterization of uncertainties associated with key components of its health benefits analyses, such as PM-related mortality effects. The NRC report recommended that probability distributions for key sources of uncertainty be developed using available data or, where scientific data are limited or conflicting, using probabilistic distributions obtained through formal elicitation of expert judgments.

Purpose and Scope

In response to the NRC recommendations, EPA is exploring how it might incorporate expert judgment in policy analysis. As a first step in this direction, IEc worked with EPA and OMB scientists to design a pilot expert elicitation to characterize the uncertainty in the ambient $PM_{2.5}$/mortality relationship. This pilot was designed to provide EPA with an opportunity to improve its understanding of the design and application of expert elicitation methods to economic benefits analysis. For instance, the pilot was designed to provide feedback on the efficacy of the protocol developed and the analytic challenges, as well as to provide insight regarding potential implications of the results on the degree of uncertainty surrounding the C-R function for $PM_{2.5}$ mortality.

The scope of the quantitative questions was limited in that we focused the elicitation on the C-R function of PM mass; this initial elicitation was not intended to fully characterize the uncertainty surrounding individual aspects of the $PM_{2.5}$/mortality relationship, such as the role of specific PM components (e.g., diesel particulates). Over the next year, EPA may plan a second phase of this project. That phase may involve a comprehensive elicitation and may explore the factors contributing to the magnitude and uncertainty of the $PM_{2.5}$ mortality relationship in a more explicitly disaggregated format.

The pilot elicitation consisted of a series of carefully structured questions, both quantitative and qualitative, about the nature of the $PM_{2.5}$/mortality relationship. The primary objective of the elicitation was to obtain experts' quantitative, probabilistic judgments about the average expected decrease in mortality rates associated with decreases in $PM_{2.5}$ levels in the United States. These judgments were expressed in terms of median estimates and associated percentile values of an uncertainty distribution. IEc developed an elicitation protocol including both quantitative and qualitative questions, in consultation with a Project Team of EPA and OMB staff scientists (hereafter "Project Team"). The quantitative questions in the protocol asked experts to provide judgments about both changes in mortality due to long-term exposure (a 1 $\mu g/m^3$ decrease in annual average $PM_{2.5}$ levels) and changes in mortality due to short-term exposure (a 10 $\mu g/m^3$ reduction in daily average $PM_{2.5}$ levels). The quantitative questions were designed to yield results appropriate for EPA's quantitative cost-benefit analyses. The questions requiring qualitative responses were designed to get experts thinking about the key issues, to allow them to establish a conceptual basis for supporting their quantitative judgments, and to provide EPA with information that would be useful for designing a more comprehensive and disaggregated elicitation assessment.

Organization of the Document

The remainder of this document is organized into four sections. The first describes IEc's analytical approach to conducting the expert judgment assessment, including the selection of experts, design of an elicitation protocol, testing of the protocol, and approach to combining expert judgments. The second presents the results of the assessment, summarizing expert responses to both quantitative and qualitative questions. The last two sections present IEc's conclusions regarding the findings of this pilot-scale assessment and recommendations for future EPA expert judgment efforts.

2. ANALYTICAL APPROACH

While researchers have been studying the process of eliciting and interpreting expert judgments for several decades, no single accepted, standardized method has emerged for this type of assessment. Expert judgment studies do, however, have certain common elements, such as the criteria used in the selection of experts and the development of an elicitation protocol. IEc describes in this section its approach to each part of its expert judgment elicitation process: expert selection, development of a briefing book, development of an elicitation protocol, the procedure for conducting elicitation interviews, and the aggregation of experts' probability distributions.

Expert Selection

Expert selection is critically important to a well-conducted expert elicitation exercise. The expert selection process should be designed to ensure that the final panel of experts has the appropriate expertise to address the questions posed to them and represents, as a group, a reasonably balanced range of respected scientific opinion on the issues being addressed. Previous studies (e.g., Hawkins and Graham, 1990) have identified several additional criteria on

which an expert selection approach should be evaluated. Ideally, the process should be explicit and reproducible; be reasonably cost-effective and straightforward to execute; and minimize, to the extent possible, the level of control by the risk assessor or manager. In selecting an expert panel for this assessment, IEc followed a process intended to meet as many of these objectives as possible within the time and resource constraints imposed by the project. Ultimately, we gave greatest emphasis to ensuring the recruitment of appropriate expertise and achieving a reasonable range of opinions in a small group of experts (five).

IEc and the Project Team recognize that the number of experts is a potential limitation of this study. A panel size of five was selected in consultation with the team, considering both the pilot nature of this elicitation exercise, and the need to expedite completion of the project in time to inform ongoing EPA regulatory analysis. The size of expert panels involved in complex expert judgment elicitations has historically been relatively small (i.e. typically 5-10). In addition, formal quantitative methods are not typically used to define the number of experts. Expertise in the panels has typically been spread among several disciplines. Morgan et al. (1984) used nine atmospheric chemists to estimate the mass balance of sulfur during long range transport from coal-fired power plants, and seven health scientists to estimate subsequent regional health impacts. The health scientists consisted of two inhalation toxicologists, two epidemiologists and three scientists experienced in the analysis of historical mortality data. Wallsten and Whitfield (1984) employed five experts to give judgments about lead-induced hemoglobin decrements and six to estimate uncertainty in IQ decrements associated with blood lead levels. Whitfield et al. (1991) used six experts to assess the risk of chronic lung injury attributable to ozone exposure. An analysis of uncertainty about the risks due to chloroform employed nine experts, but of the nine, the six non-epidemiologists followed one protocol, while the three epidemiologists chose an alternative approach (Evans et al., 1994). Finally, seven exposure assessment experts participated in a study to predict the ambient, indoor and personal air concentrations of benzene in the U.S. (Walker et al., 2001, 2003). Although the decision analysis field tends to use relatively small sample sizes (typically 5-10 experts), some members of the Project Team are not comfortable with obtaining a combined distribution from such small numbers in the absence of more extensive evaluation of the degree to which the expert panel is likely to be statistically representative of the overall population of relevant experts on the question of interest.

IEc staff identified a pool of 34 candidate experts and then used a peer nomination process to facilitate expert selection from among the members of this pool (and minimize the level of control exercised by IEc over the selections). The process of expert selection consisted of the following steps:

1. **Identifying an Expert Pool.** To expedite the selection process, the Project Team overseeing this analysis proposed that the selection of experts be limited to a pool derived from the list of members of the following two National Research Council (NRC) Committees: (1) the Committee on Estimating the Health-Risk Reduction Benefits of Proposed Air Pollution Regulations, and (2) the Committee on Research Priorities for Airborne Particulate Matter. The rosters of both committees include recognized experts in fields pertinent to the topic of this elicitation exercise, such as epidemiology and toxicology. Because these experts have already undergone extensive review of their qualifications by

the NRC, membership of these committees provided a reasonable first cut at experts likely to meet the criteria for expert selection.

IEc recognizes that restricting the pool to the NRC committees omits from consideration a number of well-regarded experts who would be appropriate for this elicitation. However, the current project was intended as an initial, pilot-scale exploration of the use of expert judgment to characterize the uncertainty in the $PM_{2.5}$/mortality relationship. Should EPA continue its efforts in this direction, we anticipate that a subsequent study would apply a refined elicitation process and a broader pool of experts. At this point, the pool of NRC committee members is sufficiently large to obtain a group of five well-qualified experts representing a range of views on this topic. We recognize, however, that by limiting both the pool and the panel size, we may not have obtained a sample fully representative of scientific opinions about the $PM_{2.5}$/mortality relationship.

2. **Narrowing the Expert Pool.** IEc reduced the pool of experts from the original NRC membership lists based on consideration of the following selection criteria provided by the Team:

    - Experts should possess the necessary scientific training for evaluation and integration of epidemiological, toxicological, and physiological evidence addressing the relationship between exposure to PM and mortality.

    - Selection of experts should consider the collective breadth and depth of the experts' professional experience in critically evaluating the scientific literature related to estimating the relationship between ambient $PM_{2.5}$ exposures and total non-accidental premature mortality. Expertise and experience should be demonstrated through a history of publications and presentations relevant to evaluating the relationship between ambient $PM_{2.5}$ and premature mortality, as well as participation in expert committees (e.g., CASAC, WHO guidelines), workshops, and document reviews relevant to the topic.

    - The overall set of experts selected should be a balanced group that reflects the range of respected scientific opinions concerning the evidence for premature mortality being associated with elevated ambient PM concentrations.

    - Experts should be U.S.-based and be available and willing to participate in late 2003.

    - Experts should preferably have no financial conflicts of interest.

Note that the above selection criteria consider not only the qualifications of the individual experts but the composition of the expert panel as a whole. The reduced pool consisted of 19 experts.

3. **Identifying Experts to Provide Peer Nominations.** To maintain an unbiased expert selection process, IEc decided to obtain peer nominations that could be used to rank qualified NRC committee members. IEc separately identified a group of experts in the field (not limited to the NRC committees) and asked them to objectively nominate individuals to participate in this assessment from our pool of 19 experts. IEc selected potential nominators using unpublished results of a literature search and publication count performed by researchers at the Harvard School of Public Health. The Harvard effort involved a Medline and Current Contents search of publications published between 1966 and 2001 concerning the mortality effects of particulate matter combined with a publication count and ranking of experts by first authorships and all authorships during that period (Wilson, 2003, personal communication). IEc selected as nominators the top 11 authors ranked by first authorships.[1]

4. **Obtaining Peer Nominations.** IEc contacted the top 11 authors, ranked by first authorships, and requested they each nominate five individuals from the narrowed expert pool from Step 2.[2] Each nominator was given a copy of the selection criteria used in Step 2 and asked to consider those criteria in making their nominations. Of the eleven nominators contacted, eight provided nominations to IEc.[3]

5. **Ranking Experts.** The experts in the narrowed NRC pool were ranked based on the number of peer nominations they received. IEc then invited the five most highly nominated experts in the pool to participate in the assessment.

6. **Replacing Experts.** If an invited expert was unwilling or unable to participate in the assessment, IEc sought to replace that expert with the next most highly nominated candidate of similar background and perspective, based on available biographical information.[4] One expert who declined to participate provided an additional set of nominations; these nominations were used to help re-rank remaining experts for purposes of identifying a replacement.

---

[1] The 11 nominators contacted included (in order of publication count): Dr. Joel Schwartz, Dr. C. Arden Pope, Dr. Bart Ostro, Dr. Frederick Lipfert, Dr. Richard Burnett, Dr. Suresh Moolgavkar, Dr. Jonathan Samet, Dr. Annette Peters, Dr. Klea Katsouyanni, Dr. Morton Lippmann, and Dr. David V. Bates.

[2] Two of the nominators were also included in the expert pool for this study. These experts were allowed to self-nominate; a review of the ranking of the experts following peer nomination indicated that self-nominations did not significantly affect the expert rankings.

[3] Drs. Schwartz, Ostro, Lipfert, Burnett, Moolgavkar, Samet, Lippmann, and Bates provided nominations.

[4] The following experts were invited to participate, but declined: Dr. Joe Mauderly, Dr. Frank Speizer, and Dr. Daniel Krewski.

The expert selection process IEc followed was generally consistent with the practices outlined at the beginning of this section. It was reasonably explicit and reproducible and was cost-effective to execute, though costs were substantially reduced by reliance on a previously conducted literature search and nomination count. In addition, use of the peer nomination process helped minimize the influence of IEc staff in selecting experts; where such influence was exerted, it served to help preserve a reasonable balance of opinion when developing a reduced pool of experts (Step 2) and when selecting replacement experts for this small panel (Step 6).

The experts ultimately selected were among the top eight experts in the pool, based on the rank ordering following the peer nomination process, and each expert in the final group had received at least three peer nominations. The resulting panel included two experts specializing in epidemiology, one specializing in biostatistics and epidemiology, and two specializing in respiratory toxicology. Of these experts, three are currently employed by academic institutions, one works for a government agency, and one is currently a consultant and President Emeritus of the Chemical Industry Institute of Toxicology. Specifically, the expert panel for the elicitation included the following participants:

> **Dr. Roger McClellan,** President Emeritus of the Chemical Industry Institute of Toxicology (CIIT);
>
> **Dr. Bart Ostro**, Chief of the Air Pollution Epidemiology Unit of the California Office of Environmental Health Hazard Assessment (OEHHA);
>
> **Dr. Jonathan Samet**, Professor and Chair of the Department of Epidemiology, Johns Hopkins School of Public Health;
>
> **Dr. Mark Utell**, Professor of Medicine and Environmental Medicine, Director of Occupational Medicine Program and Associate Chair of Environmental Medicine, University of Rochester School of Medicine; and
>
> **Dr. Scott Zeger**, Professor of Biostatistics and Epidemiology, Chair of the Department of Biostatistics, Johns Hopkins School of Public Health.

Each expert was provided a stipend to cover one half day of preparation and an eight hour interview with the elicitation team.

Briefing Book

Approximately two weeks in advance of each interview, IEc provided each expert with a binder of briefing materials on the topic of PM exposure and mortality. This briefing book was intended as a reference that experts could use when preparing for the elicitation and when responding to questions during the full day interview. The briefing book contained comprehensive recent summaries of the PM/mortality literature, such as EPA's fourth external review draft of the PM Criteria Document (hereafter, "draft PM CD"; U.S. EPA, 2003a), as well as more recently published articles not included in that summary. IEc encouraged each expert to review the briefing book prior to his interview but did not limit the information that he could consider when answering questions to that contained in the binder.

The most comprehensive summary of the PM mortality literature included in the briefing book is the draft PM CD (U.S. EPA, 2003a). Chapter 8 of that document summarizes epidemiological studies that report associations for total non-accidental, total all-cause, and cause-specific mortality and various PM measures. Studies from both North America and other locations around the world accepted for publication as of April 2002 are summarized and evaluated. The draft PM CD also evaluates the human clinical, and animal toxicological evidence with regard to the physiological changes and adverse health effects of PM. The briefing book also includes other recent assessments and reports that address the role of ambient PM in contributing to premature mortality, including EPA's *Review of the National Ambient Air Quality Standards for Particulate Matter, OAQPS Staff Paper – First Draft* (2003c); the World Health Organization (WHO) Working Group report, *Health Aspects of Air Pollution with Particulate Matter, Ozone, and Nitrogen Dioxide* (2003); and the Committee on Medical Effects of Air Pollutants (COMEAP) *Statement On Long-Term Effects Of Particles On Mortality* (2001).

IEc provided experts with electronic and/or paper copies of the draft PM CD, as well as the other reports cited above. Hard copies of the summary tables in Chapters 8 and 9 of the draft PM CD were provided for easy reference during the interviews. In addition, IEc included copies of papers concerning PM-mortality associations that were published after the cutoff date for the draft PM CD, as well as a paper by Kunzli et al. (2001) useful for conceptualizing the distinction between the mortality effects associated with short-term and long-term PM exposure. IEc included a copy of the 2001 EPA *Trends* report (2002), to assist experts in understanding current levels and composition of PM in the U.S., and provided an excerpt from the NRC report, *Estimating the Public Health Benefits of Proposed Air Pollution Regulations* (NRC, 2002), to provide experts with context for the purpose of this expert judgment assessment. The complete Table of Contents for the briefing book is provided as Appendix A to this report.

IEc's use of the briefing book in this assessment differs in some respects from how it has been used by other researchers conducting expert judgment studies (e.g., Morgan et al., 1984). Instead of providing a compendium of relevant documents, other studies have prepared briefing books that summarize key technical concepts and provide additional analyses related to the elicitation subject matter, as well as extensive materials to train experts in how to give unbiased and well-calibrated judgments. These books may be used as part of a pre-interview workshop with the experts. Given the limited time available to complete this initial phase of the assessment, IEc was unable to hold a workshop and found it necessary to simplify the nature of the briefing book in the analysis – providing a set of useful, comprehensive summary reference materials for the experts to consult in preparation for (and during) their interview. As part of the elicitation protocol, IEc did include a short primer on potential sources of bias in expert judgments that was discussed with each expert at the beginning of his interview.

Protocol Development

IEc developed an elicitation protocol to standardize the expert interview process. The protocol evolved through a series of collaborative discussions with the Project Team.[5] The final version of the protocol consisted of the following five parts:

1. **Introduction.** The first section described the objectives of the elicitation project and introduced the expert to the process by which the interviewers would elicit judgments. This introduction involved examining a sample question and response, reviewing criteria for a good, well-calibrated expert[6], discussing potential pitfalls in giving judgments (i.e., use of common heuristics and biases that can lead to poorly calibrated judgments), and conducting a practice elicitation exercise.

2. **Preview of Key Questions.** The second section provided an opportunity for the elicitation team to preview for each expert the key quantitative questions concerning the coefficient of the $PM_{2.5}$/mortality C-R function. This served to emphasize the ultimate goal of the elicitation and to allow the expert to ask any clarifying questions he might have after reviewing the protocol.

3. **Preliminary Questions: Factors to Consider.** The third section consisted of preliminary, largely qualitative questions about factors to consider when characterizing the relationship between $PM_{2.5}$ exposure and premature mortality. These preliminary questions covered the following categories:

    - evidence for the impact of short-term and long-term PM exposure on the risk of premature mortality;
    - physiological mechanisms leading to mortality;
    - causes of death;
    - form of the concentration-response (C-R) function;
    - potential thresholds in the C-R function;
    - lag/cessation period, or more simply, the time course of effects;
    - relative effect of PM components;
    - relative effect of PM sources;

[5] IEc would also like to acknowledge the early guidance and input from Dr. Roger Cooke of Delft University in the Netherlands, researchers from the Harvard School of Public Health, and the EPA Science Advisory Board Health Effects Subcommittee for the Section 812 Benefit-Cost Analysis of the Clean Air Act.

[6] In the expert judgment context, calibration refers to the ability of an expert to "capture" true values within the intervals of his probabilistic judgment with the appropriate relative frequencies. In other words, the expert accurately characterizes what he knows and does not know. This is an important concept for understanding the goals of the elicitation process. (See the section, "Evaluating Experts' Judgments," for a more thorough discussion).

- effects of exposure misclassification;
- effects of confounding; and
- effect modification.

These questions were intended to serve three purposes: (1) to get experts to think carefully about key aspects of the nature of the $PM_{2.5}$/mortality C-R relationship; (2) to allow them to establish a conceptual basis to support their quantitative judgments; and (3) to provide information to the Project Team about issues that may deserve more detailed investigation in a subsequent phase of the expert judgment assessment.

4. **Elicitation of Quantitative Judgments.** The fourth section presented the two key quantitative questions to be elicited:

   Long-term Exposure Effects: What is your estimate of the true, but unknown percent reduction in total annual, non-accidental mortality (excluding any short-term effects) in the adult U.S. population resulting from a long-term 1 $\mu g/m^3$ reduction in annual average $PM_{2.5}$ (ranging from about 8 to 20 $\mu g/m^3$) across the U.S. (e.g., the population-weighted mean effect)?

   Short-term Exposure Effects: What is your estimate of the true, but unknown percent reduction in total daily, non-accidental mortality (excluding any long-term effects) in the adult U.S. population resulting from a one-day 10 $\mu g/m^3$ reduction in daily average $PM_{2.5}$ (ranging from background up to 60 $\mu g/m^3$) across the U.S. (e.g., the population-weighted mean effect)?

For each question, the expert was asked to specify his 5th, 25th, 50th, 75th, and 95th percentile values to characterize the uncertainty in his estimates. The concentration ranges specified in each question represent the range of baseline annual and daily average ambient $PM_{2.5}$ concentrations currently found throughout the U.S.[7]

5. **Follow-up Questions.** The last section included questions that asked for each expert's judgments about the potential for differential toxicity of specific subcategories of fine particulate matter. Each question asked the expert how he might revise his responses to the key quantitative questions in Section 4 if he were told the PM mixture were higher in a particular subcategory (e.g., sulfates) than originally assumed in Section 4. Like the preliminary questions, the purpose of these questions was to provide the elicitation team with information useful for planning a subsequent, more detailed phase of the elicitation assessment.

[7] These ranges were developed through review of EPA monitor data and consultation with the EPA Project Team.

A copy of the full elicitation protocol for this assessment is included as Appendix B to this report.

Aggregated Versus Disaggregated Approach

The expert judgment literature discusses two broad approaches to elicitation of judgments; an aggregated and a disaggregated approach. As the term implies, an aggregated approach asks the expert to estimate the quantity of interest directly; for example, the numbers of newspapers sold in the U.S. in a particular year. In a disaggregated approach, the expert (or group of experts) would be asked to construct a model for estimating the quantity of interest and would be asked directly about the inputs to that model (e.g. population in each state, percentage of the population that reads newspapers, etc.) The theory behind the disaggregated approach is that it is easier for experts to answer questions about the intermediate quantities than about the total quantity. Relatively few studies have been done to compare the two approaches, however. Morgan and Henrion (1990) note that studies that have attempted to document the superiority of the disaggregated approach have shown mixed results.

The Project Team carefully considered the relative advantages and disadvantages of the two approaches. A major advantage of the disaggregated approach is a more structured and transparent characterization of the key inputs and sources of uncertainty in the final quantity of interest. However, the method does require additional time and resources to develop a model structure (or in some cases, multiple models) and set of inputs on which the experts can agree prior to the individual elicitations.

The limited time frame available to complete this assessment drove the decision to undertake an aggregate approach to elicit the C-R coefficient for the $PM_{2.5}$/mortality relationship.[8] Nonetheless, a major goal of the preliminary and follow-up questions in the protocol was to identify critical issues that could be addressed through the development of a more disaggregated approach.

Pilot Testing

Pilot testing is critical to developing a well-functioning protocol. It enables the elicitation team to test the clarity of the questions using expert volunteers and enables the team to practice and refine the process for administering the protocol. IEc conducted three pilot tests of the protocol on experts from academia and government:

[8] IEc and the Project Team initially considered using a highly aggregated approach that would have asked experts to characterize a single overall $PM_{2.5}$/mortality effect due to both short- and long-term exposures. However after advice from the SAB-HES, we opted to disaggregate effects due to long- and short-term exposures. The Project Team felt that separate questions to address effects of long- and short-term exposures, though still at a high level of aggregation, would prove to be easier for experts to address than a question that "rolled up" all the effects into a single estimate. This level of disaggregation also enabled the elicitation team to explore with experts possible overlap in reported mortality effects detected using long-term and short-term epidemiological studies.

- Dr. Kazuhiko Ito, NYU Medical Center and Dr. Lester Grant, Director, National Center for Environmental Assessment, RTP Office, USEPA, July 1, 2003;

- Dr. Jonathan Levy, Harvard School of Public Health, September 15, 2003; and

- Dr. John Vandenberg, Acting Associate Director for Health of EPA's National Center for Environmental Assessment, USEPA, October 3, 2003.

The first "pilot test" was not a formal run-through of the protocol, but more a consultation to review the key questions in an early draft of the protocol for appropriateness and clarity. The later two pilots were conducted as formal elicitations designed to gauge the time required to complete the questions and to obtain feedback on the clarity and format of questions in the protocol. None of the pilot subjects was paid for his participation.

Following the formal pilot tests, the protocol underwent final revisions. The primary changes made in response to the pilot tests involved clarifications to the wording of the quantitative questions concerning short- and long-term effects, clarifications to the assumptions underlying the quantitative questions, and revisions to the questions concerning the relationship between short- and long-term exposures on mortality.

Elicitation Process

IEc chose to elicit the judgments of each expert individually during a personal interview. This choice reflects a preference based on a review of the expert judgment literature, and the goals of the project. Numerous approaches for obtaining subjective judgments from groups of experts with differing opinions have been proposed and demonstrated in the expert judgment literature (Morgan and Henrion, 1990; Cooke, 1991). Approaches vary widely in the degree of interaction between experts during or after the process and in whether the process is intended to achieve group consensus. Unfortunately, little research has been done that rigorously examines the relative ability of the different methods to achieve well-calibrated results (Morgan and Henrion, 1990). Most of the research comparing methodologies has involved almanac-type questions (e.g., "What is the height of Mount Everest?") and has shown little difference in the quality of the results obtained by the various approaches The research does suggest that interactions between experts can increase rather than decrease the problem of overconfidence (and thus, poor calibration) (Morgan and Henrion, 1990). Whether these findings can be extrapolated to studies like this one that feature complex questions posed to relevant experts requires further research.

Individual elicitations offer several advantages for this project. First, because the issue of PM-associated mortality is a polarizing one, the potential for achieving consensus was likely to be very limited. Second, because the experts for this pilot phase of the project were selected to reflect a range of respected scientific opinions on this issue, individual elicitations enabled us to examine the variability in the experts' responses. The use of individual elicitations also avoids

the dominance of the group opinion by any one individual and "attempts to preserve the unique perspective of each expert" (Wolff et al., 1990).

Several authors (Morgan et al., 1984; Evans et al., 1994; Walker et al., 2001) have convened workshops with experts prior to individual elicitations. The workshop may have many goals including introduction of the experts to subjective judgment elicitation, explanation of the goals of the project to which the elicitation is an input, critique of the scientific literature relevant to the questions being posed, and, where appropriate, discussion of the appropriate structure for decomposing the responses to a complex issue. Although IEc and EPA would have preferred to hold a workshop prior to the elicitations, the project schedule precluded it.

Approximately two weeks prior to each interview, IEc sent each expert a copy of the protocol and the briefing book. The material was provided in advance to allow each expert to familiarize himself with the questions to be asked and the resources available in the briefing book.

The elicitations were conducted during October and November 2003. Four of the five interviews were conducted at the expert's institution; one was conducted at IEc's offices.[9] The elicitation team consisted of two interviewers, one of whom is experienced in expert judgment elicitation, Dr. Katherine Walker of IEc, and the other who provided expertise in PM exposure assessment, Dr. Patrick Kinney of Columbia University. Dr. Kinney was selected for his expertise in the subject matter of the elicitation and for his ability to objectively evaluate the responses of the various experts.

Most of the elicitations were conducted over the course of a single 8-hour day. Due to scheduling constraints, two of the elicitations were conducted over slightly shorter timeframes; however, one of these was supplemented with a follow-up teleconference to clarify responses and resolve issues remaining at the end of the in-person interview. Typically, covering the introductory material and calibration exercise took about an hour, while the remaining time was split approximately evenly between answering the preliminary and follow-up questions, and answering the quantitative questions.

For each question in the protocol, experts were asked to think systematically about the relevant evidence and to consider any sources of uncertainty, error or bias that might influence their interpretation of that evidence. The elicitation team requested that experts cite specific studies or other evidence supporting their judgment on a particular issue, and also prompted experts to consider specific evidence that would support an opposing or alternative position. This gave experts the opportunity to more fully evaluate the robustness of their conclusions, allowing them to make adjustments to their judgments in light of the full range of evidence or explain why they found the alternative evidence unpersuasive.

For the quantitative questions, each expert was first asked to specify his assumptions about the overall shape of the $PM_{2.5}$/mortality C-R function for the range of $PM_{2.5}$ concentration

[9] One expert's schedule could not accommodate an interview at his office during the timeframe of this study. The expert was scheduled to travel to Massachusetts during this time period, however, and he agreed to conduct the interview at IEc's offices in Cambridge instead.

changes specified (e.g., whether he assumed a linear or log linear relationship, a threshold at some level, etc.) Experts were then asked directly for their estimates of the individual percentiles describing the coefficient (i.e., slope) of their C-R function (the "C-R coefficient"). Typically, discussion began with the theoretical basis for bounding the estimates, although most experts preferred to begin with the median. As part of the process of eliciting quantitative values, we asked each expert to identify the most important factors that might contribute to bias in published, peer-reviewed estimates of the percent increase in mortality associated with PM exposures and to quantify the effect of such biases on the overall uncertainty in those estimates. Each expert was then instructed to consider whether his quantitative responses needed to be adjusted to account for those potential biases.

The elicitation team took extensive notes during the interviews. In some cases, experts may have written or sketched responses to certain questions. As time permitted, the expert's responses to the quantitative questions were plotted as cumulative distributions and or as boxplots and compared to the literature on which the individual expert had relied. Following each interview, IEc provided each expert with a summary of his quantitative results for review, adjustment and/or confirmation of his responses. Expert responses were not shared with the rest of the expert group.

To maintain confidentiality, each expert was assigned a randomized letter (between A and E) prior to his interview for purposes of note-taking and a second randomized letter (again between A and E) with which his judgments would be associated in this report. We provided confidentiality to allow experts the freedom to express candid, independent opinions even if they should differ from those he has expressed publicly or from those of his employer.

Approaches to Presenting Results

When faced with differing judgments across experts, analysts must give careful thought to whether and how to combine these judgments into a single value or distribution. Combination of expert judgments is not strictly necessary; some investigators (e.g., Hawkins and Graham, 1990; Winkler and Wallsten, 1995; and Morgan et al., 1984) have preferred to keep expert opinions separate in order to preserve the diversity of opinion on the issues of interest. In such situations, the range of values expressed by the experts can help decision-makers by serving as inputs to sensitivity analyses of analytical models and thereby bounding possible outcomes. Individual judgments can also illustrate dichotomies of opinion arising from different disciplinary perspectives or from the rational selection of alternative theoretical models or data sets (Morgan and Henrion, 1990). Nonetheless, analysts are often interested in developing a single distribution of values that reflects a synthesis of the judgments elicited from a group of experts. For this analysis, IEc has presented both the individual quantitative distributions of the C-R coefficient elicited from the five experts interviewed and estimates of the combined results based on an equal weighting of the individual judgments.

An extensive literature exists concerning methods for combining expert judgments. These methods can be broadly classified as either mathematical or behavioral (Clemen and Winkler, 1999). Mathematical approaches range from simple averaging of responses to much more complex models incorporating information about the quality of expert responses, potential

dependence among expert judgments, or (in the case of Bayesian methods) prior probability distributions about the variable of interest. Behavioral approaches require the interaction of experts in an effort to encourage them to achieve consensus, either through face-to-face meetings or through the exchange of information about judgments among experts. As noted earlier, IEc believes both methodological and practical issues argue against a behavioral approach. Therefore, we used a mathematical combination process to derive a single distribution.

One advantage of mathematical combination over behavioral approaches is the ability to be completely transparent about how weights have been assigned to the judgments of specific experts and about what assumptions have been made concerning the degree of correlation between experts. Several approaches can be used to assign weights to individual experts. Weights can be assigned based on the analyst's opinion of the relative expertise of each expert; on a quantitative assessment of the calibration and informativeness (i.e., precision) of each expert based on their responses to a set of calibration questions (as described in Cooke, 1991); or on weights assigned by each expert, either to him or herself or to the other experts on the panel (see Evans et al., 1994 for an example of this approach). Ideally, such a weighting system would address problems of uneven calibration and informativeness across experts, as well as potential motivational biases (Cooke, 1991).[10] In practice, appropriate weights can be difficult to determine, though Cooke and others have conducted considerable research on this issue.

In consultation with the Project Team prior to the beginning of the elicitation exercise, IEc chose to combine the experts' judgments using equal weights, essentially calculating the arithmetic mean of the expert responses. The reasons for choosing equal weights were both practical and methodological. Development of defensible differential weights was not possible given the expedited schedule for this project. Although we did conduct a sample elicitation exercise with each expert, the purpose of the exercise was to train the experts in providing quantitative responses, not to develop calibration scores that would be used to weight experts. In addition to the practical considerations, some empirical evidence suggests that the simple combination rules, like equal weighting, perform equally well when compared to more complex methods in terms of calibration scores for the combined results (Clemen and Winkler, 1999).

The equal-weight combination method we used involves averaging responses across experts for each percentile and for the minimum and maximum values elicited. We note that this method does not address the potential for dependence among expert responses that may result, for example, from reliance on the same data sources. While some amount of dependence is likely, it is difficult to quantitatively assess the extent of that dependence. Treatment of the experts' responses as completely independent likely leads to a combined distribution that reflects less uncertainty in the PM coefficient than would exist if dependencies were taken into account.

While the equal-weight combination method is straightforward in principle, applying it in this context was complicated by the fact that the elicitation protocol gave the experts freedom to specify different forms for the C-R function. If all the experts had chosen the same function, the combination of results would have been a simple exercise. For example, if each expert had

[10] "Motivational bias" refers to the willful distortion of an expert's true judgments. The origins of this bias can vary, but could include, for example, a reluctance to contradict views expressed by one's employer or a deliberate attempt to skew the outcome of the study for political gain.

specified a log-linear C-R function with a constant, but uncertain, C-R coefficient (i.e., slope) over the PM range specified in the protocol, the combination of their distributions for the C-R coefficient would require a simple averaging across experts at each elicited percentile. In this study, three experts specified log-linear functions with constant C-R coefficients over the specified range of $PM_{2.5}$ concentrations, but two of the experts believed the C-R coefficient was likely to vary over the range of $PM_{2.5}$ specified for both long- and short-term exposures. These more complex C-R functions necessitated some additional steps in the calculation of the combined results.

As discussed in detail in the Results section, one expert (B) specified a distribution for the C-R coefficient for $PM_{2.5}$ concentrations above a threshold and assigned the coefficient a value of zero for all PM concentrations below the threshold. He then specified a probability distribution to describe the uncertainty about the threshold value. Expert C specified separate distributions for the C-R coefficient at four discrete points within the concentration ranges defined in the protocol, to represent a continuous C-R function whose slope varied with the $PM_{2.5}$ concentration.[11] Both experts assumed the same functional forms in responding to questions about long-term and short-term exposures.

To derive a single distribution across all experts for a particular range of exposures (e.g. 8-20 $\mu g/m^3$ annual average $PM_{2.5}$), we first needed to estimate an "effective" distribution of uncertainty about the C-R coefficient for both Experts B and C across that range. The "effective" distribution would be a reduced form of Expert B's and C's functions that would allow averaging with the other three experts' distributions. We derived the "effective" distributions for B and C by using Monte Carlo simulation (Crystal Ball® software) to estimate the expected value of each percentile elicited across the full $PM_{2.5}$ range specified.

For Expert B, we conducted Monte Carlo sampling using two distributions: his uncertainty distribution for the threshold, and an assumed distribution of baseline $PM_{2.5}$ concentrations for the $PM_{2.5}$ range specified in the elicitation protocol. For each iteration, we selected a value from each of these two distributions and compared them. If the selected baseline concentration was less than or equal to the selected threshold value, each of the percentiles of Expert B's uncertainty distribution was assigned a zero value (no mortality effect); if the concentration was greater than the threshold, we assigned each percentile the "above-the-threshold" value specified by Expert B in his interview.[12] We repeated this process for thousands of iterations and then took the average value for each of the percentiles to obtain Expert B's "effective" distribution of uncertainty about the C-R coefficient across each range of exposures.

---

[11] Expert C indicated that the coefficient value between these points was best modeled as a continuous function, rather than a step function.

[12] An example for mortality effects from long-term exposures helps illustrate this approach. Expert B estimated that he was 75 percent sure (i.e., his 75th percentile) that the percent increase in mortality would be less than or equal to 0.5 percent per 1 $\mu g/m^3$ change in $PM_{2.5}$ concentration if the baseline concentration were above the threshold, but zero percent if it were below the threshold. If on a given iteration, the program selects a baseline concentration of 12 $\mu g/m^3$ and a threshold level of 10 $\mu g/m^3$, we assign his 75th percentile the value of 0.5. If the threshold level selected were 15 $\mu g/m^3$, the 75th percentile would be assigned a value of zero.

Expert C, who provided uncertainty distributions at several discrete points, required a different approach. We first randomly sampled from the assumed distribution of baseline PM concentrations. We then linearly interpolated between Expert C's responses at the two points nearest to the sampled PM concentration, to estimate his uncertainty distribution for the C-R coefficient at the sampled concentration. For example, Expert C provided slope values at $PM_{2.5}$ concentrations of 8, 10, 15 and 20 for mortality effects of long-term exposure. If, on a given iteration we selected a $PM_{2.5}$ concentration of 12 $\mu g/m^3$, we would generate a slope at each percentile of his uncertainty distribution by interpolating between Expert C's responses at 10 and 15 $\mu g/m^3$. We repeated this process for thousands of iterations and then took the average value for each of the percentiles to obtain the "effective" distribution of the average slope of Expert C's C-R function.

The estimates of Expert B and C's "effective" distributions, and thus the combined expert distribution, are all sensitive to the probability density function chosen to describe the U.S. baseline $PM_{2.5}$ concentrations in the simulations. This sensitivity arises because both Experts B and C assume that the effect of an increase in $PM_{2.5}$ concentration on mortality depends on the initial $PM_{2.5}$ concentration.

We have illustrated this sensitivity to PM concentrations in two ways. First, using long-term exposures as an example, we compared combined expert distributions derived using two different distributions: 1) a uniform distribution defined by the concentration ranges specified in the elicitation protocol (e.g. 8-20 $\mu g/m^3$ for long-term exposures); and 2) a normal distribution describing population-weighted annual average $PM_{2.5}$ concentration data generated from EPA's Environmental Benefits and Mapping Analysis Program (BenMAP), the model EPA currently uses for health benefits analyses of air quality regulations affecting PM and other criteria pollutants.[13] Second, for both long- and short-term exposures, we calculated combined expert distributions at four different $PM_{2.5}$ baseline concentrations. Using the methods described above, we first calculated Expert B's and C's distributions at the four concentration points and then averaged them with the distributions of the other three experts (which remain constant over the concentration range) using equal weights.

We also evaluated the sensitivity of the combined expert distribution to the removal of individual experts from the panel. This enabled us to assess the relative influence of individual experts on the combined results. To conduct the sensitivity analysis, we recalculated the combined expert judgment distribution, systematically removing one expert at a time from the average.

Evaluating Experts' Judgments

To understand how experts' performance is assessed, both in the broader expert judgment field and for this project, it is important to understand that each expert's probabilistic judgment is a reflection of his or her own state of knowledge. (How well or with what degree of certainty

[13] To facilitate Monte Carlo sampling, we evaluated the fit of the BENMAP data to several distributional forms, ultimately selecting a normal distribution, truncated at zero, with a mean of 11.04 $\mu g/m^3$ and a standard deviation of 2.32 $\mu g/m^3$.

does he think he can predict the quantity of interest?) One measure of his success is therefore related to how well he knows the limits of his knowledge. In the expert judgment field, this measure is known as calibration. An expert is well-calibrated if, for example, when asked to give his 90 percent confidence intervals for 100 predictions, his intervals contain the true value 90 percent of the time. A second important measure of the quality of an expert's judgment is "informativeness," an attribute reflecting the breadth of his confidence intervals. Two experts, one giving very broad intervals and the other very narrow, can both be well calibrated, but the latter is more informative. Finally, good judgments should be unbiased. That is, even if an expert is well-calibrated using the definition above, his median should fall close to the expected "true" median for the predicted value.

Of particular concern for analysis and decision-making is that research has suggested that experts tend to be overconfident, and therefore poorly calibrated (see Morgan and Henrion, 1990 for a review). In essence, they think they know more about the subject than they do, and therefore they express greater certainty in their predictions than is warranted. Their confidence intervals tend to be overly narrow, causing them to "miss' the true value entirely, or they may provide biased estimates. In the calibration example discussed above, an overconfident expert's 90 percent confidence intervals would contain the true value less than 90 percent of the time. Such inaccuracy in a key input to an analysis could have significant effects on the outcome.

At the same time, we do know that experts can perform well making predictions in their own areas of expertise. Studies have also shown that experts who receive regular feedback on their judgments (e.g., weather forecasters, see Murphy and Winkler, (1992); and physicians, see Winkler and Poses, (1993)). Walker et al. (2003) demonstrated that exposure experts asked to predict benzene concentrations in EPA Region V were relatively well-calibrated.

Techniques do exist to objectively evaluate experts' performance. The "gold standard" for judging their performance requires that the "truth," for example the true $PM_{2.5}$ C-R function, become known. For studies like this one, the gold standard is clearly beyond reach. Other investigators have used additional sets of questions, for which the truth can subsequently be known, to assess the calibration of experts (Cooke, 1991). This approach requires careful selection of a set of questions that are likely to be valid predictors of calibration on the questions that are the focus of the assessment. Neither of these calibration alternatives was available for this pilot assessment.

In the absence of these calibration measures, we developed the protocol and followed elicitation procedures designed to help experts avoid some of the common biases and errors of judgment (also referred to as heuristics) that can lead to poor calibration. Both during and subsequent to the interviews, we have evaluated the results of this pilot considering whether: 1) the judgments were statistically coherent; 2) the judgments were reasonably consistent with the rationales given by the experts; and 3) whether the process did help experts to avoid some of the common pitfalls associated with giving subjective judgments. We therefore discuss in substantial detail the experts' evaluation and use of the scientific literature relevant to answering the questions.

3. RESULTS

This section of the report presents the experts' responses to the questions posed in the elicitation protocol. We divide the results into two sections. We begin by summarizing expert discussions in response to the preliminary questions, which covered topics addressing the key evidence for or against a $PM_{2.5}$/mortality relationship as well as specific characteristics of such a relationship, such as the shape of the C-R function. Following the preliminary results, we summarize the experts' quantitative estimates of the percent change in mortality for a given change in: 1) annual average $PM_{2.5}$ concentrations (long-term exposure); and 2) daily 24-hour average $PM_{2.5}$ concentrations (short-term exposure).

The responses to the follow-up questions in the protocol are also presented as part of the discussion on the relative importance of PM sources and components in the Preliminary Questions section. We group these responses together because of the overlap between the preliminary and follow-up questions on this topic, and because the follow-up questions, which were often addressed with limited time at the end of the interview, yielded little significant additional information.

IEc's discussion of the experts' views about individual questions in the protocol may appear uneven; that is, some experts' views on particular topics are discussed in greater detail than others. This is a function of several factors. The experts represented different areas of expertise and therefore were particularly knowledgeable about different aspects of the PM/mortality issue. Many of the preliminary questions, in particular, invited each expert to discuss the evidence most relevant to him; the experts did not always focus on the same studies or evidence so direct comparisons across experts were not always possible. Similarly, the quantitative questions gave the experts considerable latitude to construct their own approaches to generate estimates and their approaches did differ. Nonetheless, we strove to summarize accurately the important commonalities and differences among the experts' opinions. Our goal was to help identify the rationales, key sources of data, and major uncertainties behind the experts' quantitative estimates of uncertainty. Ultimately, this information may be used to help design a quantitative elicitation protocol that allows for a more structured and explicit incorporation of expert's views on the factors that drive both the magnitude and uncertainty in the PM mortality.

Given the focus on drawing broader lessons from the elicitation, our goal is not to facilitate a critique of individual experts. Thus we frequently, but not consistently, attribute opinions to specific individual experts. At times, more general statements regarding the number of experts holding a particular view are sufficient to describe the information obtained from the interviews. Summaries of individual experts' responses to particular questions may be found in Appendix C.

Responses to Preliminary Questions

As discussed in the Analytical Methods section, the experts were initially asked a series of preliminary questions designed to get them to begin thinking about evidence and important characteristics of the $PM_{2.5}$/mortality relationship. This section presents experts' responses those

preliminary questions. We do not present the results in the order the questions were posed in the protocol. Instead, we present results first for the responses that we believe were most influential in shaping the experts' quantitative assessment of the C-R relationship between $PM_{2.5}$ and mortality (e.g., key epidemiological studies, the likelihood of a causal relationship, and the shape of the C-R function). The topics covered in this section include:

- Key epidemiological literature;
- Potential causes of death and biological mechanisms;
- Likelihood of a causal relationship between $PM_{2.5}$ exposures and mortality;
- Shape of the $PM_{2.5}$/mortality C-R Function and the likelihood of a threshold level;
- The fraction of $PM_{2.5}$-related deaths due to long- and short-term mortality effects of PM exposure;
- Exposure misclassification;
- Confounding;
- Cessation lag; and
- Relative importance of different PM sources and components.

Key Epidemiological Evidence

Discussions of key epidemiological evidence supporting or refuting a relationship between $PM_{2.5}$ and mortality permeated the interviews. Experts first discussed the strengths and weaknesses of the epidemiological evidence regarding the relationship between $PM_{2.5}$ exposure and non-accidental mortality during the preliminary questions. They then revisited this topic during the quantitative questions while deciding the relative emphasis they would place on each study in developing their quantitative estimates. Experts were not required to provide quantitative weights, but were asked to indicate when or how they used specific studies in developing quantitative estimates. The discussion below focuses on the experts' assessment of the strengths and weaknesses of the evidence regarding the relationship between $PM_{2.5}$ exposure and mortality.

Long-term Exposure Studies

All of the experts cited the following long-term, cohort-based studies as major evidence in support of a positive relationship between ambient annual average $PM_{2.5}$ concentrations and mortality:

- the Dockery et al. (1993) "Six-Cities" study;
- the Pope et al. (1995) "American Cancer Society (ACS)" study;
- the Pope et al. (2002) ACS follow-up study; and
- the reanalyses of the Six-Cities and original ACS studies by Krewski et al. (2000) for the Health Effects Institute (HEI).

The experts were uniform in their statements that the original Six-Cities and ACS studies were well-conducted and that their results proved robust upon extensive reanalysis by Krewski et al. (2000). One key strength of the Six-Cities study often noted was that it had been designed specifically for the purpose of evaluating the relationships between air pollution and health. As a result, the Six-Cities study included the recruitment of a representative sample of subjects, use of a questionnaire specifically designed for studying the effects of air pollution, reasonable control for possible confounders and effect modifiers, and control over the location of air pollution monitors. The availability of follow-up measurements of potential confounding variables such as smoking was cited as another strength of the study. Frequently cited weaknesses of the Six-Cities study include the small sample size, limited number of cities, and concerns about the representativeness of the six cities of the U.S. as a whole (since important regions of the U.S., the Southwest, Midwest and California, are not represented). Several experts expressed uncertainty about the adequacy of control for various identified (smoking) and not yet identified confounders ('lifestyle', 'stressors', cultural factors not captured in socioeconomic variables — see Quantitative Results section for more detail).

All of the experts praised the ACS cohort study for its large sample size (nearly 300,000 people in the 1995 study), the large number of cities (50 in the 1995 study), and broad geographic scope, although one expert remained more skeptical of its findings. Expert A pointed out that the air pollution characteristics in the cities also encompass a wide distribution of particle composition and chemistry, allowing for the additional sensitivity analyses conducted by Krewski et al. (2000) in their reanalysis. Further more, although the ACS questionnaire was not developed for the purpose of studying the effects of air pollution, he noted that it nonetheless provides a richer source of data on possible confounders and effect modifiers than the Six-Cities study. The weaknesses of the ACS study mentioned by most experts include the method of recruitment for the study, which favored higher income, more education, and a greater proportion of whites than is representative of the general U.S. population. Also, Expert A noted that the ACS exposure assessment was more problematic than the Six-Cities; the study had to rely on whatever monitors were available to the study which raises issues of quality control and representativeness of the exposures for the study population. Expert C questioned whether the control for smoking was adequate, citing concerns that smoking status was ascertained only upon enrollment to the study but that smoking patterns change over time.

Few experts cited any other cohort studies as primary evidence for a $PM_{2.5}$-mortality relationship, although they pointed to some short-term exposures studies as providing insight into how short-term exposures might contribute to a long-term impact. For example, Experts A and D both mentioned the Hoek et al.(2000) traffic-related study in the Netherlands as

supporting evidence. Expert A felt that the Southern California Children's Health Study finding of lung function changes in response to air quality was also overall evidence of the potential mechanism through which PM could ultimately affect increased risk of mortality.

Most of the experts did not feel that there have been any studies conducted that present strong evidence to refute the hypothesis that long-term exposures to $PM_{2.5}$ are related to increased mortality. Each of the experts was asked about the Lipfert et al. (2000) (Veterans' Cohort Study) and the Abbey et al., (1999) (Adventist Health Study of Smog or AHSMOG) study of Seventh Day Adventists in southern California, two studies that could be considered possible evidence suggesting no positive relationship. All of the experts raised questions about the strength of evidence provided by the Veteran's Cohort Study. Three of the five experts expressed the view that the AHSMOG study did not rise to the level of strong evidence, one did not discuss the study and the fifth expert ranked it more highly than the ACS study. The experts noted that the study populations of these two studies are not representative of the US.[14] Most of these experts found the analysis and results of the Veterans' Cohort study to be unclear based on currently available reports, and several expressed concern that it had not undergone a high level of peer review. Two experts noted that the AHSMOG study had advantages of a relatively unconfounded population (non-smokers and non-drinkers) and a good residence history for assigning exposures. Nonetheless, the exposure measures were not ideal; Expert A noted that, for some years of the study they used PM_{10} levels estimated from Total Suspended Particle (TSP) measurement. Expert C suggested the measurements do not represent geographically diverse areas or a large differential in air quality levels. Expert A noted that the McDonald et al. (2000) study, a variant of AHSMOG that studied a subset of people located within a given range of a local airport and estimated fine particle levels from airport visibility data (after correcting for humidity), has found a mortality effect in a subset of people living in high concentration areas.

In their quantitative assessments, most of the experts ultimately placed relatively less weight on the mortality estimates from the Six-Cities data and relied more heavily on estimates from the Krewski et al., (2000) reanalysis of the ACS cohort and, because of the benefits of its additional years of followup, the later Pope et al. (2002) study. Two experts noted the Pope et al. (2002) finding of a decrease in relative risk with additional years of follow-up as consistent with a concern that mortality effects observed at the time the studies were conducted (and exposures measured) could be associated with earlier, higher exposures. (See discussion in Exposure Misclassification/Exposure Error section).

Short-term Exposure Studies

The experts also discussed evidence that they would take into consideration when estimating the magnitude of the short-term mortality effects of exposure to ambient $PM_{2.5}$ concentrations. Experts cited the body of evidence from numerous individual city time-series studies and the findings from specific multi-city studies --- the Dominici et al. (2003a) reanalysis

[14] The Veterans' Cohort assessed male Veterans with high blood pressure, originally recruited to assess the efficiency of anti-hypertension drugs, and the AHSMOG study involves only Seventh Day Adventists, a relatively small pool likely to be representative of a healthier population.

of the National Morbidity, Mortality, and Air Pollution Study (NMMAPS); the Burnett and Goldberg, (2003) analyses of eight Canadian cities; Schwartz et al. (1996); and the Schwartz (2003) analysis of $PM_{2.5}$ exposure in the Six-Cities cohort; the Stieb, Judek and Burnett (2002) meta-analysis; the Hoek et al. (2001) traffic-related study in the Netherlands; and the European study by Samoli et al. (2003), known as Air Pollution and Health: A European Approach 2 (APHEA 2). The discussions of key studies centered around three major issues:

- **Single-city vs. Multi-city Studies.** When evaluating the body of data from time-series studies, most experts tended to rely more heavily on data from multi-city analyses. Three of the experts argued strongly that multi-city studies using a consistent analytical methodology across cities were the most reliable. Three experts noted that the use of a consistent analytical approach determined *a priori* was a particular strength in that it would minimize the potential for analyst bias in choosing model form, lag structure, and options. Of the studies discussed, they believed the large, geographically diverse NMMAPS study in the U.S. to be the strongest. The other two experts, A and C regarded the consistent analytical model as a possible weakness. Experts A and C argued that the benefit of single city studies was that they could control for weather and seasonality in ways that were better tailored to conditions in the geographical area. Even among the experts who favored the approach, two believed that the NMMAPS study may have been overly conservative in its control for seasonality, thus potentially underestimating the "true" time-series effect.[15]

- **Relevance of PM_{10} Studies.** Another factor in the experts' consideration of the available evidence was their view of the relevance of PM_{10} exposure data for estimating the mortality effects of $PM_{2.5}$. All but Expert C were readily willing to convert results from studies using PM_{10} to obtain a mortality coefficient for $PM_{2.5}$, using a factor derived from national data on $PM_{2.5}/PM_{10}$ ratios (U.S. EPA, 2000). Such a conversion essentially assumes that all of the PM_{10} effect is attributable to that portion of PM_{10} that is $PM_{2.5}$. Expert C raised questions about the relative roles of the various PM fractions in causing mortality and was only willing to rely on studies using $PM_{2.5}$ as an exposure measure.

- **Distributed Lag Effects.** Finally, although not an explicit part of the question about the effects of short-term exposures, the experts' views on distributed lag effects were a critical part of discussion of the likely magnitude of the effects. The Schwartz (2000b), Zanobetti and Schwartz (2003) and the APHEA 2 study often were the focus of discussion about the appropriateness and quantitative impact of using a distributed lag model.

[15] This concern did not necessarily lead the experts to adjust their quantitative estimates, however.

Major Causes of and Mechanisms for $PM_{2.5}$-related Mortality

Early in the interview, the experts were asked to identify what they believed to be the major causes of death associated with long-term and short-term exposures to $PM_{2.5}$. For each of these causes of death, they were asked to discuss the potential causal mechanisms linking premature mortality to a) long-term exposures and b) short-term exposures. In most interviews, however, experts saw the mechanisms as part of a continuum of exposure and therefore discussed the impact of long- and short-term exposures together. Where they did draw distinctions between the mechanisms leading to mortality related to long-term and to short exposures, we have noted them in our discussion below.

Major Causes of $PM_{2.5}$-related Deaths

All of the experts identified cardiovascular disease and pulmonary disease (non-cancer), in order of importance, as the most significant causes of death related to long-term exposures to $PM_{2.5}$ However, as will be discussed in later sections, they disagreed on the strength of the causal association. The experts did not agree that cancer is an important cause of death related to $PM_{2.5}$; three of the five experts believed it plausible that cancer deaths could be related to $PM_{2.5}$ exposure. Two experts were unconvinced, with one expert calling the data supporting PM relationships to cancer deaths "highly uncertain". Several experts noted that the order of importance of the three causes of death reflected the baseline rates of these causes of death in the general population.

All of the experts cited cardiovascular events (e.g., heart attack) as the primary cause of death related to short-term exposures to $PM_{2.5,}$ although one expert also cited pneumonia or other infections as possible causes in individuals with compromised lung function. A common theme in these discussions among several of the experts was the difficulty in determining the ultimate cause of death in these acute cases. For example, a death can result from heart failure in someone suffering from chronic obstructive lung disease just as a respiratory infection can precipitate the death of someone with underlying heart failure. The experts noted that death certificate data frequently obscure such distinctions.

Potential Mechanisms

The experts varied in their knowledge of and familiarity with the evidence for the possible mechanisms by which long and short-term exposures might lead to various causes of death. Nonetheless, they all laid out conceptually similar frameworks. Expert A's discussion, which was one of the more detailed we had on this subject, captures well the elements of several experts views.

Expert A defined three general categories of mechanisms:

- circulatory and cardiac events (related to inflammatory and atherosclerotic changes);
- pulmonary and systemic inflammation; and

- disturbances of the cardiac-autonomic nervous system.

He cited a growing body of evidence for plausible mechanisms by which cardiovascular and pulmonary disease might develop from fine particulate exposures. He discussed both toxicological and epidemiological studies relating PM exposures to increases in C-reactive protein, increases in fibrinogen, and increases in coagulation and plasma viscosity (Ghio et al. 2000; Peters A. et al., 1997; Peters A., et al. 2000a; Peters A., et al., 2000b; Peters A, et al., 2001a; Peters A, et al., 2001; Seaton et al., 1999). These factors are indicators of injury and inflammation and can be predictors of subsequent heart disease and mortality. Although the studies have observed these effects in most cases following short-term exposures, Expert A felt that they are indicative of a mechanism that could also be a part of a longer-term process.

He described a conceptual model for the pulmonary and systemic inflammation mechanism involving the deposition of smaller particles, in particular, to the deep lung. These particles can cause direct injury as well as inflammatory responses that can amplify that injury and initiate another chain of adverse effects. For example, increased respiratory infections, hyper-responsiveness, and other markers of lung injury could precede chronic obstructive pulmonary disease (COPD). Several experts in addition to Expert A cited the Utah studies that exposed cell lines to concentrated air pollution particles both before and after the closure of the local steel mill as illustrative of the increased inflammatory responses due to PM exposure (e.g. Dye et al., 2001).

Expert A also described a third type of mechanism that involves impact on the nervous system, in particular, the cardiac-autonomic nervous system. He noted that several studies (Gold, et al., 2000; Pope, et al., 1999; and Liao et al., 1999) have shown associations between PM exposures and heart rate variability and/or cardiac arrhythmias. The evidence from "defibrillator studies" showing associations between increased numbers of arrhythmias with increased particle concentrations is particularly strong since there is no reliance on recall by patients and the doctors downloading the defibrillator data are blind to the particulate concentrations (Peters et al., 2000a; Peters et al., 2001a).

Expert B also described the possible mechanism for PM-related cardiovascular disease as operating through the increased risk of atherosclerosis, resulting from chronic inflammation of the arteries. Expert B discussed studies (epidemiological and laboratory) that showed increases in biomarkers of inflammation, c-reactive proteins, fibrinogen, conduction disturbances, and heart rate variability following exposure to fine particles. He found the Peters et al. (2000a,b; 2001a,b) work showing relationships between particulate exposure and cardiac arrhythmias and other irregularities intriguing as a possible mechanism for $PM_{2.5}$ to trigger cardiac events. He also thought studies showing decreases in oxygen diffusing capacity in healthy human volunteers exposed to ultrafine particles to be suggestive evidence of a mechanism for influencing cardiac events in susceptible individuals. In general, he believed fine particles to be a more likely explanation for the cardiovascular effects than coarse particles.

Expert C laid out a general conceptual framework for mechanisms of cardio-respiratory disease related to deposition of particles in the respiratory system, cytotoxicity, and "a cascade of events that take place both locally and … beyond." Much of his discussion, however, centered on concerns about disentangling the effects of $PM_{2.5}$ from those of other particulate fractions (i.e.,

$PM_{10-2.5}$) and the role of higher historical exposures in the etiology of underlying levels of frailty and rates of death observed in recent epidemiological studies.

Experts D and E gave less detailed responses, but described conceptually similar mechanisms for the impact of $PM_{2.5}$ on cardiovascular and pulmonary disease as Expert A and B. However, Expert D felt the plausible arguments for mechanisms existed mostly by analogy to the effects of smoking or higher levels of exposure to PM. Like several other experts, he thought that short-term exposure effects on mortality probably occur via mechanisms that affect individuals who are already in a state of frailty. Expert E generally felt that the mechanistic models were not well established and remained a source of uncertainty.

Likelihood of a Causal Relationship between Long- and Short-term $PM_{2.5}$ Exposures and Mortality

Following their discussion of the mechanistic and epidemiological data available to characterize the relationship between long-term and short-term $PM_{2.5}$ exposures and premature mortality, the experts were next asked to sum up their views by assessing the likelihood of a causal relationship. The question requested that the experts select from a range of qualitative categories and then specify a quantitative probability of causality for each type of effect. While the qualitative categories were specified in the protocol, each expert was free to choose any quantitative probability value that they thought best reflected the category they chose.

Table 1 summarizes the experts' judgments about the likelihood of causal relationships between both long and short-term exposures to $PM_{2.5}$ and premature mortality. Both the categorical and numerical values are presented. The summary shows that, in general, the experts displayed higher confidence in the causal relationship between short-term exposures and mortality than they did for a causal relationship for long-term effects. Experts who expressed increased confidence for short-term effects cited the large base of time series studies supporting such effects and the strength of the data sets on which those studies were based. The experts assigned at least a five percent probability that no causal relationship between $PM_{2.5}$ exposure and premature mortality exists for either long-term or short-term exposures.

Table 1
Likelihood of a Causal Relationship between $PM_{2.5}$ and Non-Accidental Premature Mortality

	Effects of Long-term $PM_{2.5}$ Exposure		Effects of Short-term $PM_{2.5}$ Exposure	
Expert	Likelihood (categorical)	Likelihood range[a] (best estimate) %	Likelihood (categorical)	Likelihood range[a] (best estimate) %
A	Highly likely	85-90 (88)	Highly likely	90-95 (93)
B	Somewhat unlikely	40-50	Somewhat likely	65-80
C	Somewhat unlikely	50[b]	Somewhat unlikely to somewhat likely	50 [b]
D	Somewhat likely	50	Somewhat to highly likely	80-90
E	Likely	80-98 (95)	Likely	80-98 (95)

a. Represents minimum to maximum for categorical likelihood specified. Each expert specified his own quantitative probability estimates to match his qualitative categorical description.
b. Expert C wanted to answer this question separately for different parts of the range. The 50 percent value represents his "average" for the range, with little or no probability of a causal relationship at the low end of the range and a high probability at the upper end of the range.

The experts' responses to the questions about causality highlighted some issues in the design of the causality question. Some of the experts suggested that they would have chosen to answer the questions differently had they been asked specifically about: a) the likelihood of causality in different parts of the $PM_{2.5}$ concentration ranges; and b) the likelihood of causality for different types of mortality (cardiovascular deaths versus cancer). Table 1 also shows that the qualitative probability categories did not always have the same quantitative meaning for all subjects. This is a common problem with use of qualitative categories.

Shape of the C-R Functions / Thresholds

The experts were asked to provide their judgments about the true shape of the C-R function relating mortality with both short-term and long-term $PM_{2.5}$ concentrations and also to discuss the potential for the existence of a threshold concentration below which no mortality effects would be expected. For both short- and long-term effects, three of five experts ultimately chose a log-linear C-R function with no threshold, one chose a log-linear function that incorporated an uncertain threshold, and one chose a non-linear function that also included a threshold. [16]

When considering the relationship for the effects of long-term exposure, Experts A, D, and E assumed that mortality would be log-linearly related to concentration, although they recognized that this might be an oversimplification of reality. Expert D stated that, "the relative risk model… is probably not right." Expert E noted that the relationship is probably monotonic but that it is not currently possible to distinguish between a log-linear function and alternative mathematical descriptions of the curve. All three believed that existing data did not support the identification of any population thresholds and that the available cohort data were consistent with a log-linear interpretation.

Expert B assumed some population threshold could exist below which there would be no effect of increased $PM_{2.5}$ exposure and above which the relationship would be linear. Although he acknowledged that little evidence exists to support identification of a population threshold, he argued that one was likely on biological, mechanistic grounds. He did characterize his estimate of a possible threshold as uncertain, ranging between background (about 4 $\mu g/m^3$) and 15 $\mu g/m^3$, with a modal value of 12 $\mu g/m^3$.

Expert C believed that the log-linear relationships found in existing studies are not well-grounded in biological theory, but rather are dictated by the statistical methods used in those studies. Expert C believed that the increased relative risks for mortality observed in the cohort studies were likely to be the result of exposures at the higher end of the exposure range (or of earlier high historical exposures) and that he expected there to be a declining effect on mortality

[16] The experts all discussed a theoretical model in which the C-R relationship should exhibit a threshold at the individual level. That is, an individual might have a threshold for a particular cause of death. However, individual thresholds may vary both across individuals and by cause of death for a given individual. Most agreed that variation in thresholds among individuals or population subgroups limited the ability of epidemiological studies to detect a population-level threshold. Though some experts' quantitative estimates employed thresholds, none of the experts could cite a study or studies that provided strong evidence of a threshold for either short- or long-term $PM_{2.5}$ mortality effects.

with decreasing levels of $PM_{2.5}$. He also argued that there was likely to exist some practical concentration threshold below which we would not observe any increase in mortality.

The responses considering short-term effects were similar to those for long-term effects. Experts A, D, and E assumed a log-linear relationship with no threshold, expert B assumed a log-linear function with a threshold, and Expert C assumed a non-linear function with declining mortality effects at lower $PM_{2.5}$ levels and a practical threshold. Expert A said he was more confident positing a log-linear relationship for short-term mortality because of larger base of time-series studies of $PM_{2.5}$ and mortality. Expert D cited the NMMAPS study as being consistent with a log-linear relationship for short-term effects. Expert B again defended his choice of threshold on biological and mechanistic grounds and defined an uncertain threshold, with a distribution ranging between background levels (about 4 $\mu g/m^3$) to 25 $\mu g/m^3$, with a modal value of 15 $\mu g/m^3$. Expert C argued that linear or log-linear function is not likely to be consistent with the underlying biology of effects and believed that a one-day increase in $PM_{2.5}$ would likely have a decreasing impact on mortality with a decreasing baseline daily $PM_{2.5}$ concentration.

Fraction of $PM_{2.5}$-related Deaths Due to Short and Long-term Exposures

The experts were asked two questions designed to help them consider the differences between the findings of the cohort and time-series studies. The first question asked what percent of the mortality effect reported in the cohort studies was likely attributable to short-term PM exposures.[17] The second question asked what percent of mortality due to short-term exposure to PM is not encompassed within the effect reported in the cohort studies.[18] The questions required the experts to discuss the possible conceptual framework for describing the relationship between the different types of mortality effects. They were asked to draw on their understanding of the mechanisms for PM-related mortality with which we began the interview and to consider several alternative models, including one posited in a paper by Kunzli et al. (2001; see Figure 1)[19] The purpose of the question was ultimately to help the experts think about a 'net' effect of long-term exposure (minus any short-term effects captured by cohort studies) in developing quantitative estimates of the effects of long-term exposures on non-accidental mortality.

[17] If for example, one believed the cohort effect to be a function of deaths due to both short-term (Cst) and long-term exposures (Clt), the percent of the cohort effect due to short-term exposures alone would be given by Cst/Ctotal *100.

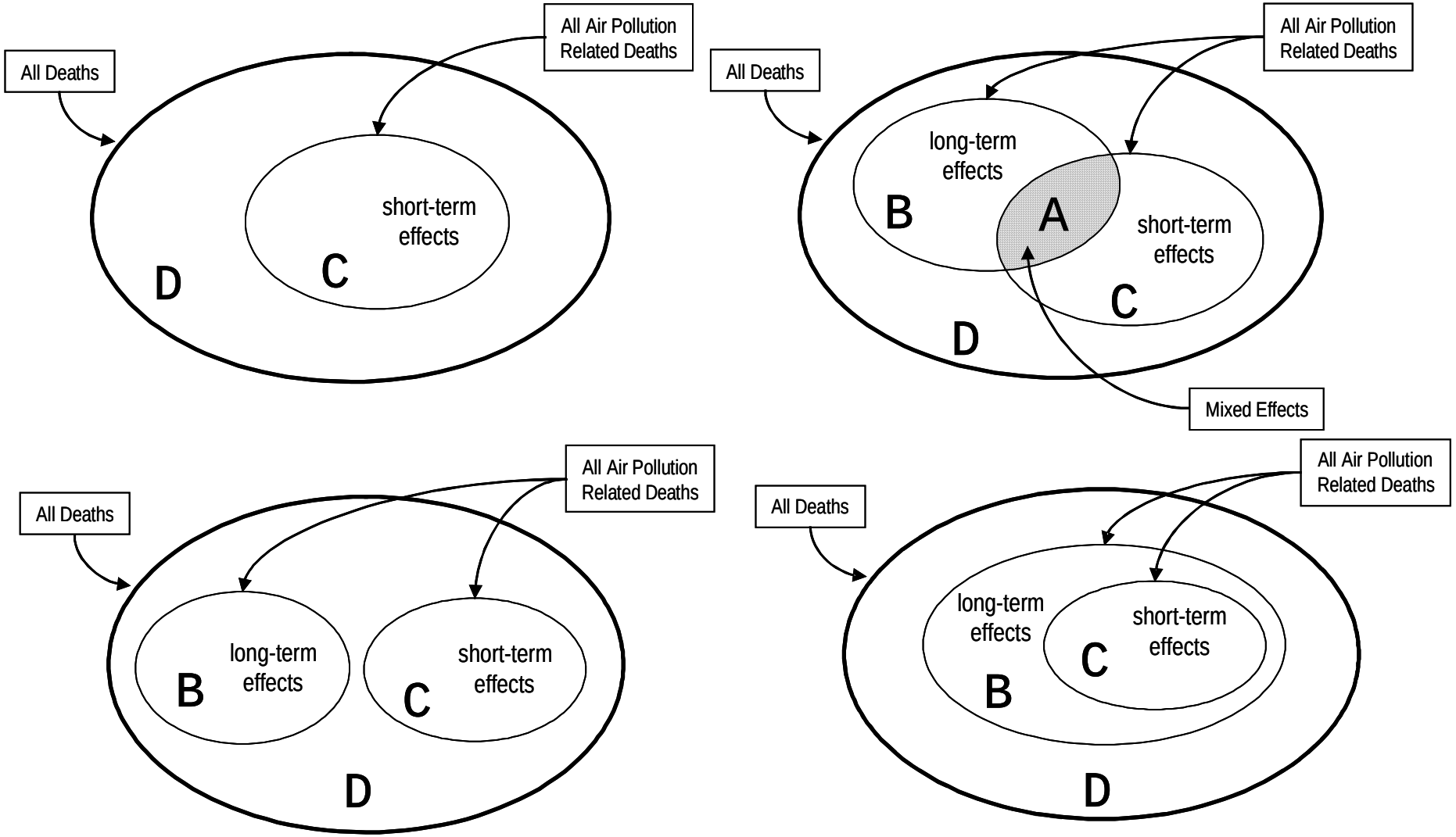
[18] Assuming that the time-series effect includes some mortality captured by the cohort studies (TSc) and some additional amount that is not (TSadd), the percent additional mortality is estimated by TSadd/TStotal *100.

[19] The Kunzli et al (2001) version is in the upper right quadrant of the figure.

Figure 1

ALTERNATIVE MODELS OF DEATHS ATTRIBUTABLE TO AIR POLLUTION

(Adapted from Kunzli et al. (2001) Note that sizes of circles have no quantitative meaning.)



All found the structure described by Kunzli et al. (2001) a useful starting point for discussions though most added modifications to his basic conceptual framework. All the experts still found it difficult to develop quantitative responses to these questions. The main difficulties arose as experts tried to distinguish between the types of deaths that were likely to occur as a result of long- and short-term exposures, how they affected days or years of life lost, and how that affected whether they would be captured by the cohort and time series studies designs. Expert B observed that the question was not clear about whether his views on causality should be a part of his response, an implicit issue underlying responses given by others. It generally appears that they did incorporate their views on causality into their responses. For example, Expert B, who was skeptical about the causal relationship between long-term $PM_{2.5}$ exposures and mortality, estimated that 50 to 85 percent of the mortality effects were likely due to short-term exposures.

The experts' responses to these questions, summarized in Table 2, varied widely even when experts gave somewhat similar rationales. Ultimately, the experts did not feel confident about using their responses to this question when providing their quantitative judgments about the potential increase in mortality associated with long-term exposures. Summaries of each individual expert's responses to this question can be found in Appendix C.

Table 2
Summary of Expert's Opinions about the Percent of Effects Reported from Cohort Studies That Represent Short-term Mortality Effects and the Percent of Short-term Effects not Captured by Cohort Studies

Expert	Percent of cohort mortality effects that are short-term[a]	Additional mortality effect from short-term exposures not captured in cohort studies[b] (percent)
A	10-15	30
B	50-85	~10
C	Up to 50	50%=0 50%= some small number (5-10%)
D	<10	>90
E	10-20%	20%

a. $Cohort_{Total\ Effect} = C_{ST} + C_{LT}$
$\%C_{ST} = C_{ST}$ /cohort * 100

b. $Time\ Series_{Total} = TS_{cohort} + TS_{additional}$
$\%TS_{additional} = TS_{additional} / TS_{Total} * 100$

Exposure Misclassification/Exposure Error

Concerns about exposure misclassification and/or exposure error, in one form or another, permeated the interviews. The following paragraphs briefly summarize the experts' views. Experts did not always distinguish clearly between exposure measurement for cohort (long-term) and time-series (short-term) studies in their remarks.

Expert responses exhibited some consistency in the concerns they cited. All experts expressed some concern about how well the exposures used in the cohort epidemiological studies reflect the relevant exposure history of the subjects. Two experts specifically mentioned the possibility that early life exposures may contribute to increased mortality risk (one included pre-natal exposures). Some pointed out that ambient measurements taken at enrollment or during follow-up in a cohort study may underestimate the level and range of relevant exposures that occurred in prior years, when ambient levels of $PM_{2.5}$ were likely to have been higher. Four of the five experts cited concern about the representativeness of monitor locations relative to study subjects' personal exposures; two of these specifically mentioned the issue of population migration (commuting or moving to another city). (Two experts noted that this issue is less of a problem for time-series results than for cohort studies.) One expert mentioned uncertainty related to use of $PM_{2.5}$ as a surrogate for other particulate components.

The experts expressed a range of views as to the ultimate quantitative impact of exposure misclassification and errors on estimates of risk. Expert B recalled data indicating that random, non-differential exposure misclassification was likely to result in underestimates of the risk, a view shared by others, although it seemed mostly to increase his uncertainty about the validity of the risk estimates. On the exposure misclassification issue, expert D's view was that it was "very doubtful, both in the short- and long-term that exposure error could be leading us to upwardly biased estimates." Expert E agreed that exposure misclassification, where the true individual exposure is measured with some error, probably leads to an underestimate of the PM effect. However, he and others noted that there were likely to also be exposure biases in the cohort studies (as noted above) that could lead to upward biases in risk estimates. Two experts saw evidence for this phenomenon in Pope et al. (2002)'s finding of a lower mortality effect using exposure data from 1980 than using exposure data from 2000 to assess risks of $PM_{2.5}$ in the extended ACS follow-up analysis.

Confounding/Effect Modification

In these questions, the experts were asked to identify the most important sources of confounding and effect modification for both the existing cohort studies and time-series studies. In this context, a "confounder" would be a factor that contributes to mortality risk and is also associated with PM exposure. Smoking is a classic example of a potential confounder.[20] An effect modifier is a factor whose value influences the association between exposure and effect. For the PM / mortality effect, age would be considered an effect modifier. It is worth noting that some factors can be both a confounder and an effect modifier for the same association (Hennekens and Buring, 1987). For example, if smokers were found to have a higher relative risk of mortality due to PM exposure than non-smokers, even after controlling for potential confounding due to smoking, then smoking would serve as both a confounder and an effect modifier.

[20] For example, if people in more polluted cities tended to smoke more than people in less polluted cities, smoking would be a potential confounder, because smoking also increases one's mortality risk. If smoking status was not included in a PM/mortality risk model, the mortality effect of smoking would erroneously be attributed to PM exposure, leading to an overestimate of the PM/mortality effect.

The experts' responses are summarized in the text below and in Table 3 for long-term studies. In general, we reached these questions near the close of the qualitative discussions and they did not always receive the same level of attention as earlier questions. In particular, the experts did not discuss in much detail whether particular confounders or effect modifiers were likely to contribute to their characterization of uncertainty in the $PM_{2.5}$/mortality coefficient.

Table 3
Variables Mentioned by Experts as Potentially Significant Confounders or Effect Modifiers of the $PM_{2.5}$/Mortality Relationship (Long-Term Studies Only)

Variable	Number of Experts Who Cited As:		Comments
	Confounder	Effect Modifier	
Smoking	5	1	
Diet	4	0	
SES Variables / Education	3	1	One expert cited SES as a proxy effect modifier for proximity to traffic.
Health Status / Pre-existing disease	3	0	e.g., diabetes, obesity
Co-pollutants	3	0	2 experts specifically cited SO_2
Occupation	1	0	

SES Variables = Socioeconomic Status Variables

For cohort studies, the experts cited a number of commonly-discussed confounders: smoking, socioeconomic status (SES) variables, diet, pre-existing disease / health status, and co-pollutants. On the issue of co-pollutants, two experts specifically cited SO_2, based on the ACS study; one noted that SO_2 only resulted in a 5-10 percent decrease in effects estimate for PM. One expert expressed frustration with the tendency of cohort and time series study critics to raise vague and poorly thought-out arguments about confounding. He suggested that there exist systematic ways to consider such issues that would be more productive and useful.

The experts cited few effect modifiers for cohort studies. One specifically cited SES variables as a proxy for proximity to traffic, a variable difficult to measure in practice.

The experts' discussions of confounding and effect modification in the time-series studies generally identified many fewer variables. Because the time-series studies follow the same group of people over time, the cross-city or region confounders are of lesser importance. One expert cited SES variables as possible effect modifiers; another cited seasonal variation as an effect modifier, noting that there is a more sizeable effect in the spring, summer and fall than in the winter, and in the Northeast than in the rest of the country. The latter expert also cited temperature and the changing flora of infectious diseases as potential confounders of time-series studies.

Cessation Lag

Each expert was asked to discuss his views on the potential length of the cessation lag associated with fine PM-related mortality effects, where cessation lag is defined as the length of

the time period between a reduction in ambient $PM_{2.5}$ and achievement of a new, lower steady-state level of mortality risk.

In general, the experts believed that insufficient data exist to provide specific, well-substantiated estimates of the cessation lag for PM-related mortality. Two experts used the term "guesswork" to describe specific quantitative lag estimates; one referred to such estimates as "pure speculation." Nonetheless, two of the five experts (A and E) were willing to explore how one might generate rough quantitative lag estimates. Both felt that some of the deaths avoided would be rather immediate, specifically the uncertain fraction of deaths that are captured by cohort studies but are in fact due to short-term fluctuations in daily PM. For long-term mortality effects, both these experts drew a distinction between risks of cardiovascular deaths, which they felt would be reduced within five years based on the smoking literature , and other causes of death such as lung cancer, which they hypothesized would be reduced over a longer timeframe. Expert A suggested 15 to 20 percent of effects might be short-term mortality that would yield relatively immediate benefits, another 40 to 50 percent of deaths avoided could be realized within 5 years, and the remainder would be realized over a period of up to 25 or 30 years (e.g., lung cancer deaths). Expert E suggested that 20 percent of the total deaths avoided might be realized immediately, another 30 percent might be realized within two years, and the remainder might be realized over a period of 20 years.

Four of five experts cited the smoking literature on cessation lags as a reasonable starting point from which to begin thinking about lags for PM-related effects, though most felt that that literature alone was insufficient to resolve the issue for PM-related mortality.

Contribution of PM Components and Sources to the C-R Relationship

The experts' responses to both preliminary and follow-up questions regarding the relative contribution of individual $PM_{2.5}$ components to the observed premature mortality associated with fine PM was generally consistent. With few exceptions, all experts concluded that there was insufficient evidence to allow them to conclude that one PM component might contribute more to PM toxicity than another. In response to the follow-up questions, Expert A responded that he would have increased his C-R coefficient estimate if the $PM_{2.5}$ was much higher in black carbon (soot).[21] Expert B, in response to the follow-up questions, said he might lower his quantitative estimate by some unspecified amount if nitrate concentrations were greater than expected and said he might increase his response if the ultrafine particle concentrations were greater than expected.

In response to a question asking about the relative contributions of $PM_{2.5}$ from different source types, four experts mentioned some evidence suggesting that PM from motor vehicles might be more potent than that from other sources. Two experts (A and D) specifically mentioned the Laden et al. (2000) study and the reanalysis by Schwartz (2003) that reported that transportation sources are more toxic. However, Expert D described the findings of that study as "exploratory," and an insufficient basis for policy decisions. Expert B thought motor vehicle emissions were much more important than other sources (i.e., consistency in finding

[21] He said he might increase his response 1.5 to 2 times if the $PM_{2.5}$ consisted entirely of black carbon.

PM/mortality effects in different cities and country probably attests to the presence of motor vehicles); in order to conclude that utilities were an important source, he would want to see clearer evidence that PM mortality effects in the eastern U.S. were worse than in the west.[22] Expert C thought the existing studies showed that crustal materials are probably not potent. He also mentioned a recent paper by Schlesinger and Cassee (2003) that implies that sulfates are not a significant contributor to premature mortality.

Responses to Key Quantitative Questions

The elicitation protocol asked experts to provide quantitative estimates of the percent change in non-accidental mortality associated with a specified change in long-term and short-term exposure to $PM_{2.5}$. Regarding the potential effects of long-term exposures, experts were asked for the percent change in annual, non-accidental mortality associated with a permanent 1 $\mu g/m^3$ increase in the annual average concentration of $PM_{2.5}$.[23] Regarding the potential effects of short-term exposures, experts were asked for the percent change in daily non-accidental mortality associated with a single day increase of 10 $\mu g/m^3$ in the daily 24-hour average concentration of $PM_{2.5}$. In both cases, each expert was asked to express his uncertainty about the predicted percent change in mortality per unit increase in $PM_{2.5}$ in the form a cumulative distribution. Specifically, the elicitation team asked each expert to estimate a median (50^{th} percentile) value for the percent change in mortality as well as a 5^{th}, 25^{th}, 75^{th} and 95^{th} percentile of an uncertainty distribution. Experts were also asked to provide a minimum and a maximum value to bound this distribution. The questions were predicated on a number of important assumptions that are discussed in Analytical Methods section and in the elicitation protocol in Appendix B.

The summary of responses to the quantitative questions is organized in two broad sections, the first addressing the results for the effects of long-term exposures and the second summarizing the results for the effects of short-term exposures. For each type of exposure, we first present the experts' individual responses, both their quantitative estimates and their rationales for their characterization of uncertainty. We then present the results of combining the experts' distributions using the equal-weighting approach described in the Analytical Methods section, and then compare the results of this elicitation with the results from selected epidemiological studies. We also present the results of analyses done to test the sensitivity of the combined results to assumptions about the underlying $PM_{2.5}$ distributions and to the judgments of individual experts. The results of the analyses are all presented in the form of figures; tables of the quantitative values can be found in Appendix D.

[22] Expert A noted that the NMMAPS study, by showing that the Northeast has a steeper dose-response curve, might suggest that power plant emissions are important, though he also acknowledged that this difference could be due to higher motor vehicle emissions along that corridor.

[23] Note that the question in the protocol was framed in terms of *decreases* in $PM_{2.5}$ and in percent mortality, However, the experts preferred to frame their responses in terms of *increases* in $PM_{2.5}$ and mortality, reflecting the way results are often reported in the epidemiology literature on which they relied.

Effects of Long-term $PM_{2.5}$ Exposure on Mortality

The first set of quantitative estimates IEc elicited from the experts was the expected percent increase in total annual non-accidental mortality that could be associated with a permanent 1 $\mu g/m^3$ increase in the annual average concentration of $PM_{2.5}$. The experts' responses to this question differed in an important way from the original intent of the protocol. The protocol asked that the experts estimate changes in mortality associated only with long-term exposures, excluding any effects of short-term exposures. Our intent was that the experts could use their responses to preliminary questions in the protocol intended to estimate what proportion of mortality effects captured in the cohort studies was likely due to long-term exposures (see Question F in Appendix B). As discussed earlier, these questions were problematic and, in practice, none of the experts developed their estimates this way. Their estimates are instead reflective of an overall "cohort"-based effect (i.e., an effect derived from a long-term cohort study that may include mortality effects due to both long- and short-term exposure). This observation is true even for individual experts who, in the preliminary discussions, believed that the cohort studies were capturing some percentage of short-term effects.[24]

To provide a clear picture of the experts' responses regarding the effects of long-term exposures, we first discuss the elicited values, then describe briefly the experts' approaches to developing their distributions, and finally discuss the key sources of uncertainty discussed by the experts.

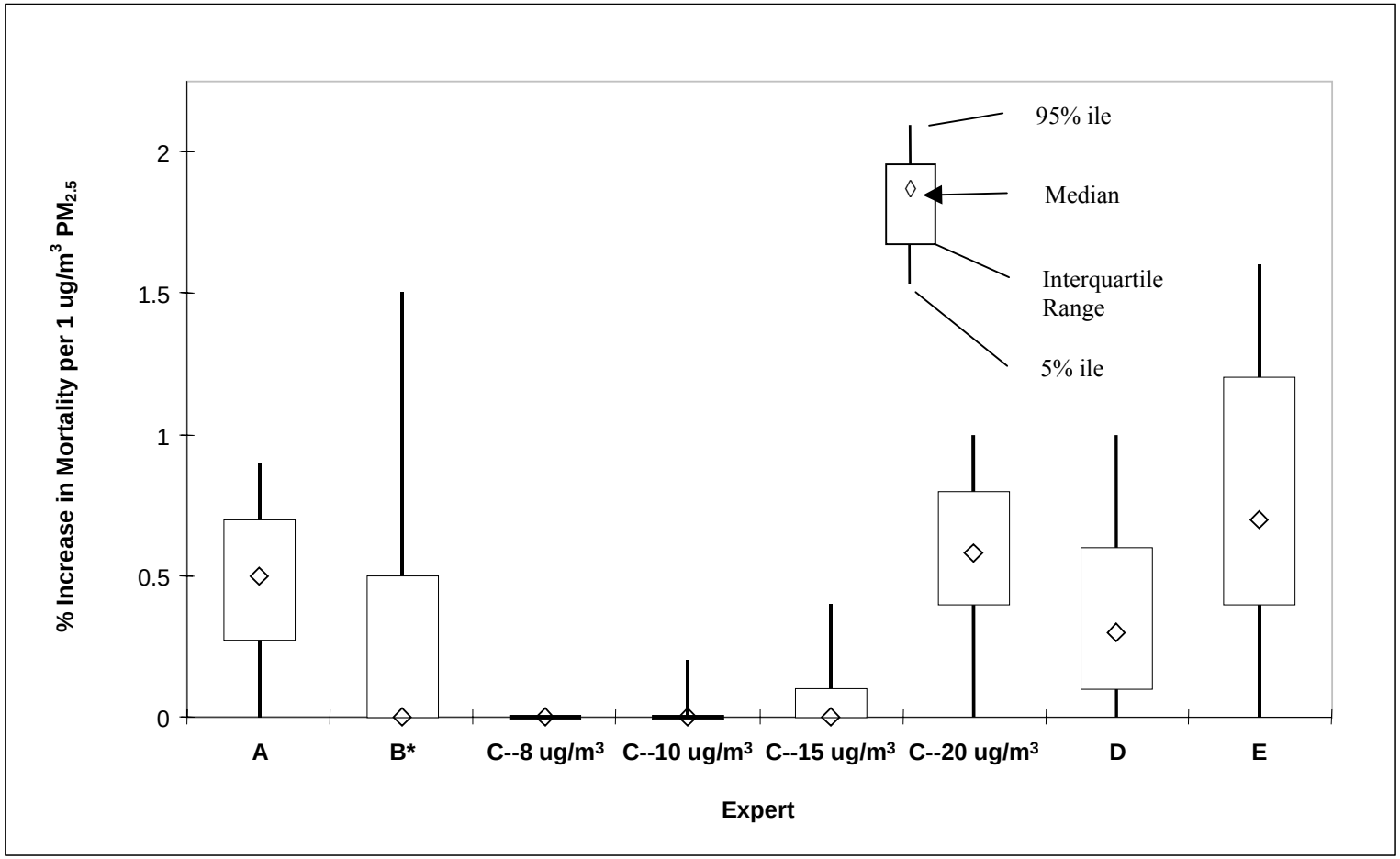
Elicited Values - Long-Term Exposure Effects

Figure 2 displays the results of the quantitative elicitation for each of the experts. Their distributions are depicted as boxplots with the diamond symbol showing the median, the box defining the interquartile range, and the whiskers defining the 90 percent confidence interval.

[24] We did obtain quantitative estimates of those percentages (see Table 1, summarized in the qualitative results section) and could develop non-overlapping estimates of long- and short-term mortality effects. However, we have not done so in reporting results back to individual experts, or for this report. We found that using the judgments the experts provided (i.e., C-R coefficients based on effects reported in cohort studies) facilitated evaluating them in the context of the existing epidemiological evidence.

Figure 2

Comparison of Experts' Judgments about the Percent Increase in Annual Non-Accidental Mortality Associated with a 1 μg/m^3 Increase in Annual Average Exposures to $PM_{2.5}$ (U.S. Baseline 8 to 20 μg/m^3)



*Expert B specified this distribution for the PM/mortality coefficient above an uncertain threshold which he characterized as ranging between 4 and 15 with a modal value of 12 ug/m^3.

As illustrated in Figure 2, considerable variation exists in both the median values and the spread of uncertainty provided by the experts. The median value of the percent increase in annual non-accidental mortality per unit increase in annual $PM_{2.5}$ concentration (within a range of $PM_{2.5}$ concentrations from 8 to 20 $\mu g/m^3$) ranged from values at or near zero to a value of 0.7 percent. The variation in the responses largely reflects differences of opinion among the experts concerning key epidemiological results from long-term cohort studies, the likelihood of a causal relationship, and the shape of the C-R function. The previous section of this report provided detailed descriptions of the experts' judgments about these factors, but we present a few brief observations relative to their responses below.

Key Cohort Studies. The experts' non-zero responses for the percent change in annual mortality were mostly influenced by the Krewski et al., (2000) reanalysis of the original ACS cohort study and by the later Pope et al. (2002) update of the ACS study, with additional years of follow-up. None of the experts ultimately placed substantial weight on the mortality estimates from the Six-Cities data in composing their quantitative responses, despite citing numerous strengths of that analysis. Concerns about sample size and representativeness of the six cities for the entire U.S. appeared to be the major reasons for de-emphasizing those results.

Causality for Long-Term Effects. Three of the five experts gave distributions more heavily weighted towards zero. Those experts were also the ones who gave the lowest probability of a causal effect of long-term exposure to $PM_{2.5}$ in the preliminary questions. All of the experts placed at least a 5 percent probability on the possibility that there is no causal relationship between fine PM exposure and mortality; as a result, all experts gave a fifth percentile value for the C-R coefficient of zero.[25] For most of the experts, this was based primarily on residual concerns about the strength of the mechanistic link between the exposures and mortality.

Shape of the C-R Function for Long-Term Effects. The other key determinant of each expert's responses for long-term effects was his assumption about the nature of the C-R function across the range of baseline annual average $PM_{2.5}$ concentrations assumed in the study (8 to 20 $\mu g/m^3$). Three experts (A, D, and E) assumed that the function relating mortality with PM concentrations would be log-linear with constant slope over the specified range. They therefore gave a single estimate of the distribution of the slope describing that log-linear function. The other two experts provided more complex responses.

Expert B assumed a population threshold in his model, below which there would be no effect of increased $PM_{2.5}$ exposure and above which the relationship would be log-linear. He characterized his estimate of a possible threshold as uncertain, ranging between 4 $\mu g/m^3$ and 15 $\mu g/m^3$, with a modal value of 12 $\mu g/m^3$. He then described a distribution for the slope for the

[25] The experts' distributions are not all completely consistent with their response to the preliminary question about the probability of a causal relationship. For example, one of the experts gave a 50 percent probability of causality (i.e., a 50 percent chance that the C-R coefficient is zero) in response to the preliminary question on causality. When providing his quantitative assessment of the C-R coefficient, however, he gave a non-zero value for his 25th percentile of the distribution and did not state that his distribution was conditional on the existence of a causal relationship. The elicitation team encouraged experts to try to resolve such inconsistencies, though some experts opted not to make changes to their judgments.

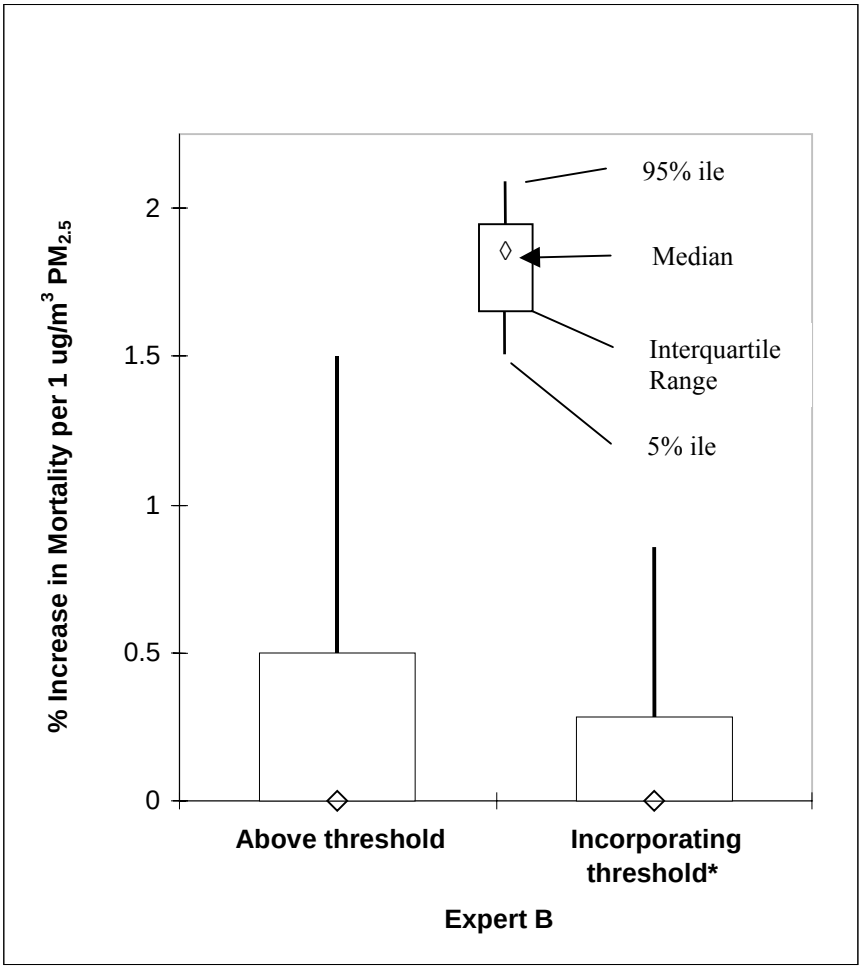
log-linear function that might exist above the threshold; this distribution is depicted in Figure 2. Figure 3 compares Expert B's distribution for the C-R coefficient in the log-linear range above a threshold to his "effective" distribution for the full concentration range, obtained by probabilistically combining his uncertain threshold and his uncertainty about the C-R coefficient above the threshold as described in the Analytical Methods section. The effect of incorporating the uncertain threshold is essentially to shift his entire distribution downward.

Expert C believed that the increased relative risks for mortality observed in the cohort studies were likely to be the result of exposures at the higher end of the exposure range, and he expected there to be a declining effect on mortality with decreasing levels of $PM_{2.5}$.[26] He also argued that some practical concentration threshold was likely to exist below which we would not observe any increase in mortality. He reflected these beliefs by developing a non-linear model within the range from 8 to 20 $\mu g/m^3$; he described the model by providing distributions for the slope of the curve at four discrete concentrations within the range. His four distributions (including a set of all zero values at 8 $\mu g/m^3$) are shown in Figure 2.

[26] In general, he critiqued the predication of the quantitative questions on a range of 8-20 $\mu g/m^3$ believing that the effects observed in epidemiological studies are likely to be the result of higher historical exposures.

Figure 3

Expert B's Distributions for the Percent Increase in Annual Non-Accidental Mortality Associated with a 1 μg/m^3 Increase in Long-term Exposures to PM2.5: Comparison of His Distribution Above a Threshold to His Expected Distribution* for the Range 8-20 μg/m^3



* Expert B specified the threshold as uncertain between 4 and 15 μg/m^3 with a modal value at 12 μg/m3. He assumed the percent increase in mortality to increase linearly with concentration above the threshold. His effective distribution was simulated using Monte Carlo techniques assuming an underlying distribution of population-weighted annual average $PM_{2.5}$ concentrations for the U.S. generated from the BenMAP model (see Analytical Approach section for details).

Approaches to Developing Distributions for Long-Term Exposure Effects

When developing their estimates of uncertainty for the C-R coefficient, four of the five experts generally took an "informal" approach to weighing the evidence from the studies, considering the uncertainty represented by the statistical error and limitations from the studies. That is, most did not begin with some prior estimate of a mean or median and standard error and then work outward to estimate particular percentiles of their distributions. Instead, most experts approached the elicitation of each percentile separately.

Expert A's overall approach was to consider carefully the different epidemiological studies we had discussed, their published central estimates and confidence intervals and uncertainties arising from their design or implementation. The central tendency of his estimate is dominated by the Pope et al., 2002 study because of the strengths of the ACS studies generally and because of the additional years of follow-up included in this later study. The decrease in the relative risk compared to the first published study was consistent with his view that these newer results may be more reflective of current air pollution levels. The 25th and 75th percentiles were selected to reflect a greater amount of uncertainty than that reflected by the statistical error in the Pope study.

Expert A's minimum and 5th percentiles reflect the small probability he assigns to the likelihood that the $PM_{2.5}$/mortality relationships observed to date are not causal. He derived his maximum value both from considering the confidence intervals from the ACS study and also considering what might be a plausible upper bound on the percent of total mortality attributable to $PM_{2.5}$. Expert A estimated that, at a 1 percent increase per $\mu g/m^3$ of $PM_{2.5}$ (roughly the 90 percent upper confidence limit on the Pope, 2002 study), we would have to attribute roughly 25 percent of all mortality to $PM_{2.5}$ across the range 5 to 30 $\mu g/m^3$. He believed these values seemed to strain credulity; he set the 95th percentile at a slightly lower value, 0.9 percent per $\mu g/m^3$. The higher central estimates from the Six-Cities study seemed more implausible to him.

Expert B believed it likely that a population threshold for the long-term mortality effects could exist. Above a threshold, he assumed that the CR relationship would be likely to be log-linear with increased exposure, absent any data to the contrary. The maximum of Expert B's distribution was informed by the upper 95th percentiles of the Six-Cities and ACS studies. He could not imagine a scenario in which the percent increase could be higher than about 2, the upper 95th percentile of the Six-Cities study, noting again that this study had wider confidence intervals because of its smaller size. He felt the Pope et al. studies provided a more plausible basis for his upper 95th percentile, adjusted upward to allow for uncertainties associated with possible exposure misclassification and differences in $PM_{2.5}$ composition between cities. His median was driven by the likelihood he placed on a causal relationship (50 percent probability at zero).

Expert C believed that a 1 $\mu g/m^3$ increase in $PM_{2.5}$ would likely have a decreasing impact on mortality with a decreasing baseline annual average $PM_{2.5}$ concentration. He argued that the linear and log-linear models primarily reflect limitations in the statistical methods available to investigators and that they are not likely to be consistent with the underlying biology. Consequently, he specified a non-linear C-R function, specifying distributions for the

instantaneous slope of the curve at four discrete points in the $PM_{2.5}$ range for this assessment as shown in Figure 2. As in the case of Expert B, the effect of these assumptions is to shift his 'effective' distribution for the full range of annual average $PM_{2.5}$ concentrations downward (not shown).

Expert D's final uncertainty distribution is bounded at the high end by his judgment that air pollution is likely to be a small contributor to chronic heart and lung disease; he felt these diseases have multiple etiologies others of which have been shown to be more important. Current smokers, who receive on the order of milligrams of toxic particles daily, experience about a doubling of their risk of cardiovascular disease relative to non-smokers. An increase of two percent per 1 $\mu g/m^3$ increase in $PM_{2.5}$, comparable essentially to the 95 percent upper confidence interval for the Six-Cities study, seemed uncomfortably high to him. It serves as his maximum value. The minimum and 5^{th} percentiles of his distribution were set at zero based on his residual concerns about the causal relationship between air pollution and mortality.[27] Expert D argued that one percent was as high as he could plausibly set the 95^{th} percentile. The remaining percentiles (25^{th}, 50^{th}, and 75^{th}) were set subjectively, with little specific discussion of each data point, because he felt there was such limited data on which to rely in making these judgments.

Expert E approached estimation of his subjective confidence intervals in two ways, initially starting with direct estimation of quantiles using a informal subjective weighting of the primary studies he wished to rely on, the Krewski et al. (2000) re-analyses of the Six-Cities and ACS studies. Subsequently, he chose to develop his estimates of the percentiles using a more structured approach. His first objective was to evaluate what confidence intervals should be expected on the basis of statistical error alone assuming the two studies were independent estimates of the "true" long-term mortality effect. He calculated an inverse-standard variance weighted average of the median and a weighted estimate of the variance. He then reconsidered his more subjectively estimated percentiles, considering the impact of various sources of uncertainty (see next section).

Sources of Uncertainty - Long-Term Exposure Effects

The experts were asked at several points during the interview to discuss the key sources of potential bias and uncertainty in current evidence on which they relied for their judgments. In the context of the quantitative discussion they were asked to list the top 5 issues. They were encouraged to think about how these issues would affect the uncertainty surrounding their best estimate of the potential impact on total mortality of a small change in long-term exposure to $PM_{2.5}$. The tables summarizing the factors identified by each expert may be found in Appendix E.

[27] Expert C was invited to consider whether his 25^{th} and 50^{th} percentiles were consistent with his 50:50 probability of a causal relationship. He did not want to change either estimate.

Many of the same factors appeared in the experts' lists. However, the experts often differed on whether a particular factor was a source of potential bias or uncertainty. We list below some of the common concerns raised as either sources of bias or uncertainty:

- **Residual confounding by smoking.** The same types of mortality associated with $PM_{2.5}$ exposures (i.e., cardiovascular, COPD) are also associated with smoking. The effect of smoking on mortality is so large relative to that associated with $PM_{2.5}$ that it is particularly important to rule out confounding by smoking.

- **Residual confounding by "life style" or other personal factors or "stressors."** These represent a collection of factors – diet, deterioration in social/community support structures, drug use, etc. – that increase frailty or risk of death and that are correlated with air pollution levels.

- **Exposure errors/misclassification.** Experts were concerned about the use of more current exposures as surrogates for historical (likely higher) exposures. Another concern was misclassification of exposure arising from differences between concentrations measured where people live and where they work or where they live at the time of the study versus where they grew up. They also noted changes over time in regional patterns of housing construction and/or air conditioning use (affecting indoor/outdoor ratios of pollutants), and the differential impact of averting behavior (e.g., people staying indoors) during high pollution days.

- **The role of co-pollutants as confounders or effect modifiers.** The SO_2 finding from the HEI reanalysis (Krewski et al., 2000) was cited by two experts.

- **Impact of the relative toxicity of PM components.** While acknowledging this as an uncertainty, experts felt that the limited literature on this issue precluded identifying specific differences among subcategories of PM at this time.

- **Representativeness of the cohort populations with respect to the general U.S. population.** This included concern about selection bias in the ACS cohort (see Responses to Preliminary Questions Section.)

- **Investigator/publication biases.** Expert B and E discussed the concern that, despite the strength of the analytical skills of the key investigators in the field, there may exist bias towards publishing studies featuring positive results.

Despite the many qualitative discussions about sources of uncertainty, the use of an aggregate approach to eliciting experts judgments ultimately made it difficult to evaluate systematically how specific sources of bias or uncertainty influenced individual experts'

results—unless an expert explicitly adjusted his estimates by a particular factor. As discussed earlier, one expert essentially did check his subjective estimates to convince himself that they allowed sufficiently for the uncertainties he expressed. He compared the variance in his subjective distribution to the variance expected from statistical error alone, based on an inverse-weighted variance combination of the results from the Krewski et al., (2000) reanalysis of the Six-Cities data and ACS data. He estimated that his variance was generous --- nearly three-fold greater than the statistical variance based on a pooling of the two studies. He felt that this increase in uncertainty reflected: 1) acknowledgment of potential heterogeneity in the true relative risk across the populations in the two studies and relative to the U.S. population to which theses estimates will be applied; 2) general imperfections in epidemiology as a science (i.e., not a randomized trial); and 3) measured and unmeasured confounders that may not have been adequately controlled for that would cause epidemiologic studies to over or underestimate the true effect.

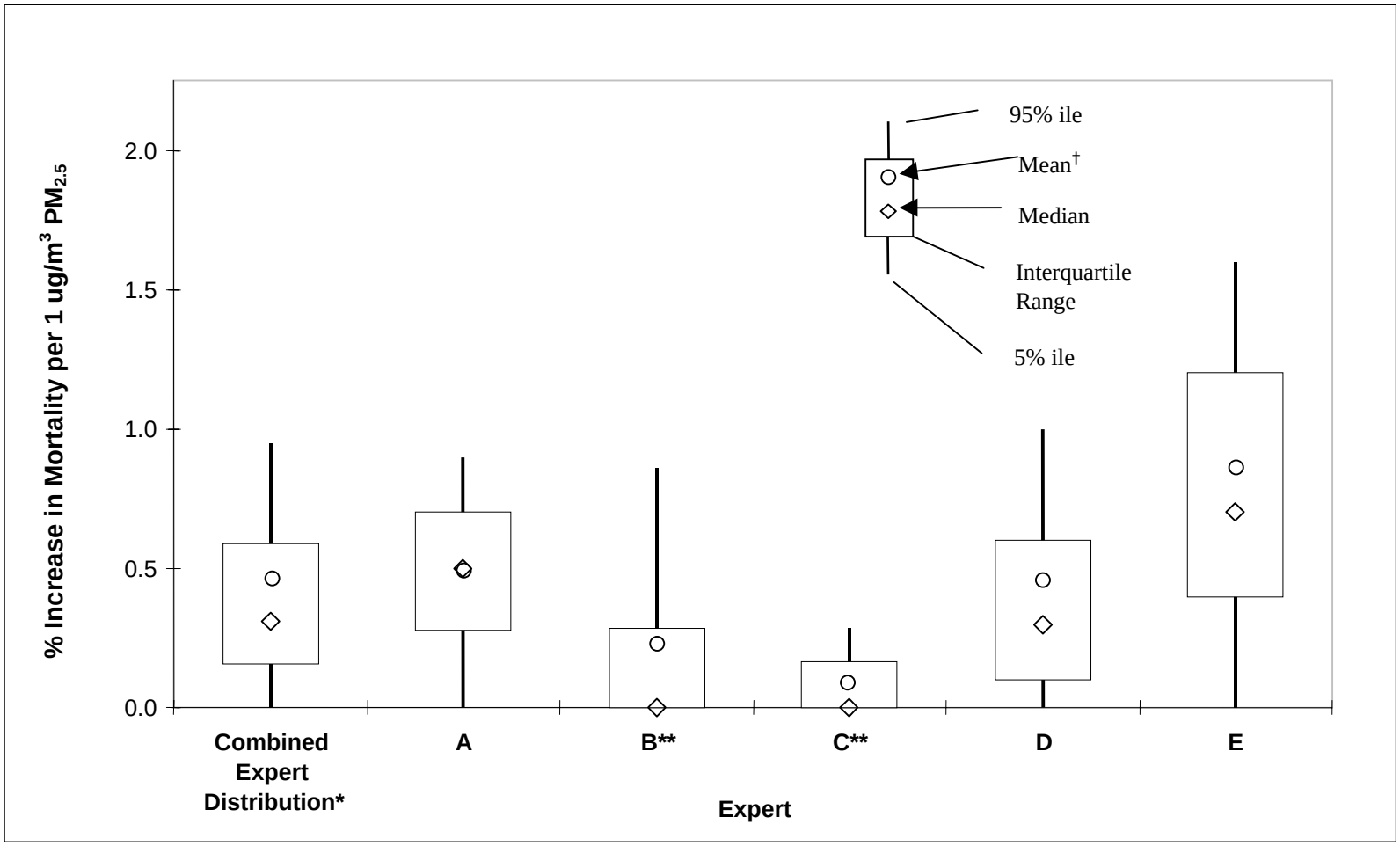
Effects of Long-term Exposure - Combined Results of Experts

As described in the methodology section, we combined the experts' judgments using equal weights. We then evaluated the distribution in several ways, comparing it to the judgments of individual experts, testing its sensitivity to the form of baseline $PM_{2.5}$ distribution used to calculate the "effective" distributions for Experts B and C, and evaluating the influence of individual experts. In order to provide perspective on the amount of uncertainty reflected in the combined judgments of the experts on our panel, we compared the combined distribution to the mean and standard errors reported in the primary cohort studies on which the experts relied for their judgments.

Figure 4 compares the combined expert distribution to the individual distributions given by the experts. Both Expert B's and C's distributions are their "effective" distributions for the range based on the assumption of an underlying population-weighted distribution of annual mean $PM_{2.5}$ concentrations generated from EPA's BenMAP model (as described in the Analytical Methods section). The combined distribution reflects a balance of the quite varied individual distributions, having a median of 0.3, mean of 0.4 and a 90 percent confidence interval (CI) of 0 to 0.93 percent increase in mortality per 1 μg/m3 increase in annual average $PM_{2.5}$.

Figure 4

Experts' Judgments about the Percent Increase in Annual Non-Accidental Mortality Associated with a 1 µg/m^3 Increase in Annual Average Exposures to $PM_{2.5}$ (U.S. Baseline 8 to 20 µg/m^3): Comparison to Combined Expert Distribution



* The experts' judgments were combined assigning equal weight to each expert.

** The distributions for Expert B and expert C are their expected distributions for the $PM_{2.5}$ range specified. Prior to combining their results with other experts, Monte Carlo sampling was used to calculate the expected values for Expert B's and C's distributions for the 8-20 ug/m3 range assuming an underlying distribution of population weighted annual average PM2.5 concentrations generated from the U.S. EPA's BenMAP model (see Analytical Methods section for details).

†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

The figure also includes estimated means for each distribution, calculated using the elicited distributions as input to a Monte Carlo simulation (the means were not elicited from the experts). The means are typically greater than the medians given by the experts suggesting some skewing of the subjective distributions toward higher concentrations. This likely reflects bounding of the lower half of the distribution by zero and for some experts, greater uncertainty about the potential upper values of the distribution.

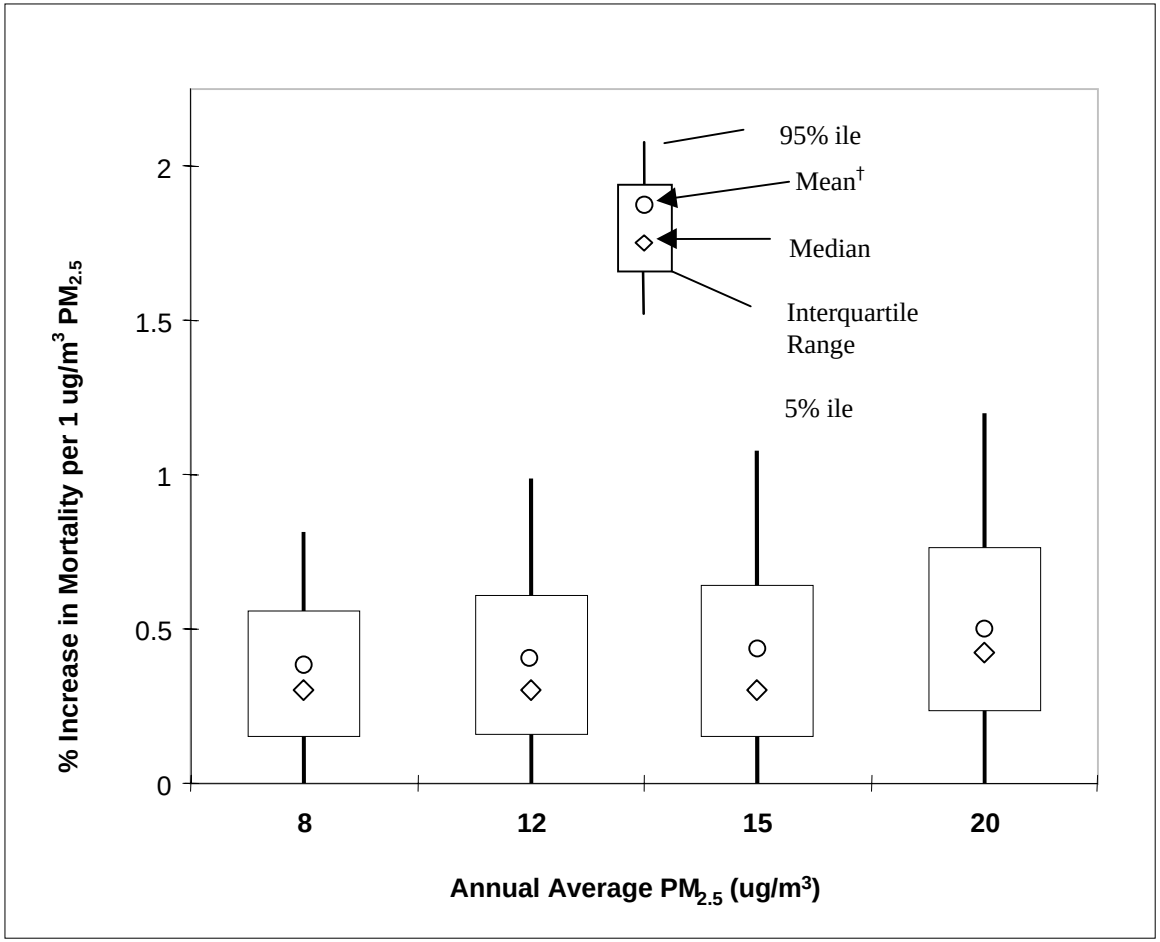
Collapsing the experts' distributions to one distribution for the full range obscures the influence of Expert B and C, whose C-R coefficients are conditional on the baseline concentration. Figure 5 more clearly illustrates the concentration dependence of the combined distribution by plotting the distribution at four baseline annual average $PM_{2.5}$ concentrations: 8, 12, 15, and 20 $\mu g/m^3$. From the high end of the range, where B and, and particularly C, placed higher probabilities on increases in mortality related to $PM_{2.5}$, the distributions shift progressively, though not dramatically, downward. This downward shift reflects incorporation of higher probabilities placed by B and C on zero percent increases in mortality at lower concentrations.

The concentration dependence of the combined distribution is an important limitation on use of the overall combined distributions presented in this report. Because B's and C's distributions are conditional on concentration, the form of the underlying distribution used to generate their "effective" distributions matter. The probability density function essentially provides the set of weights by which the concentration-dependent distributions of Expert B and C are multiplied to obtain an overall expected value for the concentration range.

Figure 6 illustrates the impact of changing the assumption about the underlying $PM_{2.5}$ distribution when generating the effective distributions for B and C. The left hand boxplot is the combined expert distribution developed using a uniform distribution (8 to 20 $\mu g/m^3$). The right hand boxplot was developed using a normal distribution (mean = 11.04, standard deviation = 2.32) based on output from the BenMAP model as described in the Analytical Methods section. Although not dramatically different, the combined expert distribution based on the uniform distribution is shifted higher at every percentile except the 5^{th} percentile when the uniform distribution is used. More of the probability density in the uniform distribution falls at concentrations more likely to be above Expert B's threshold than in the normal distribution, a distribution likely to be more representative of the true distribution of concentrations in the U.S. For all of the subsequent figures involving long-term $PM_{2.5}$ exposures, we have used the normal distribution based the BenMAP data.

Figure 5

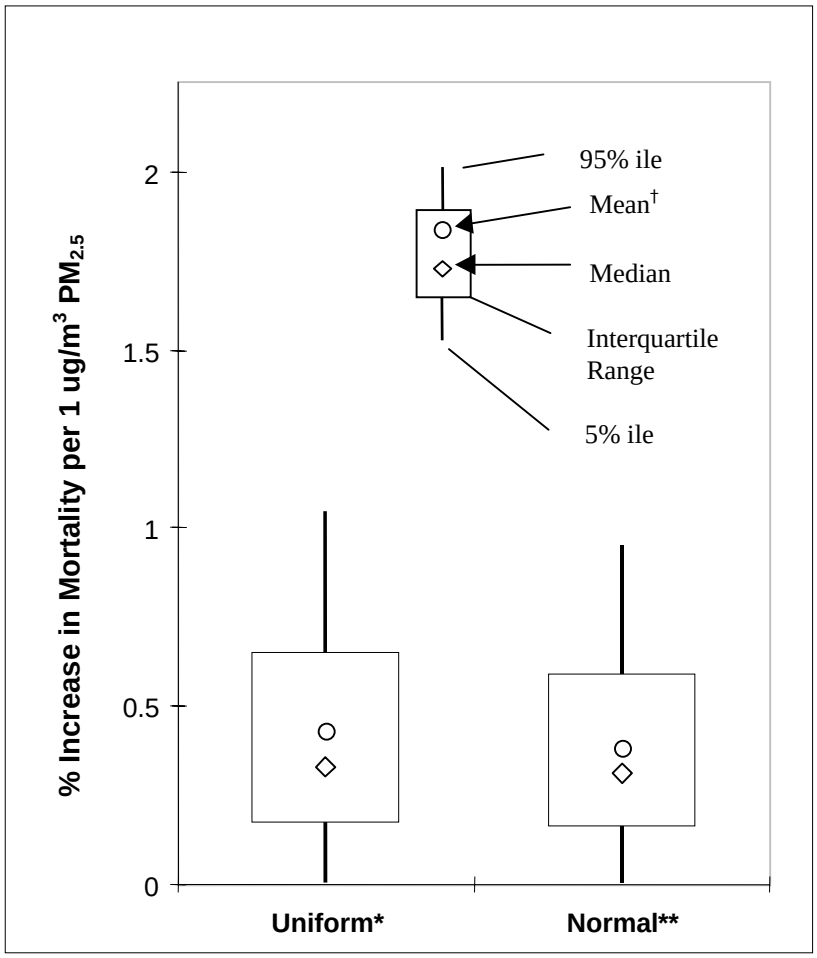
Combined Expert Judgment Distributions for Percent Increase in Annual Non-Accidental Mortality Associated with a 1 µg/m^3 Increase at Specific Baseline Annual Average $PM_{2.5}$ Concentrations*



* The experts' judgments were combined assigning equal weight to each expert.
†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

Figure 6

Comparison of Combined Expert Distributions Using Alternative Underlying $PM_{2.5}$ Distributions for Incorporating Expert B's and Expert C's Judgments



* Uniform $PM_{2.5}$ distribution, range 8 to 20 μg/m^3 as specified in the elicitation questions.
** Distribution of population-weighted annual average $PM_{2.5}$ values generated from the USEPA's BenMAP model. Data best fit a normal distribution, truncated at zero (see Analytical Methods section for details).
†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

The concentration dependency limits the direct use of the combined distributions presented for policy analysis. The equal weighting assumption in the combined distribution only holds if the underlying distribution used to adjust Expert B and C's distribution is the same as that used in the ultimate analysis. Otherwise, depending on the form of the distribution, implicitly more or less weight may be given to Experts B and C, distorting the analysis. The preferred approach is to apply the individual experts' distributions in conjunction with the specific $PM_{2.5}$ distribution for a particular application and then pool the results for each expert.

IEc also conducted a sensitivity analysis to evaluate the influence of individual experts' distributions on the combined distribution. Figure 7 compares the combined expert distribution with all five experts to a sequence of combined distributions with each of the five experts removed in turn. The expert omitted is indicated on the x-axis of the figure.

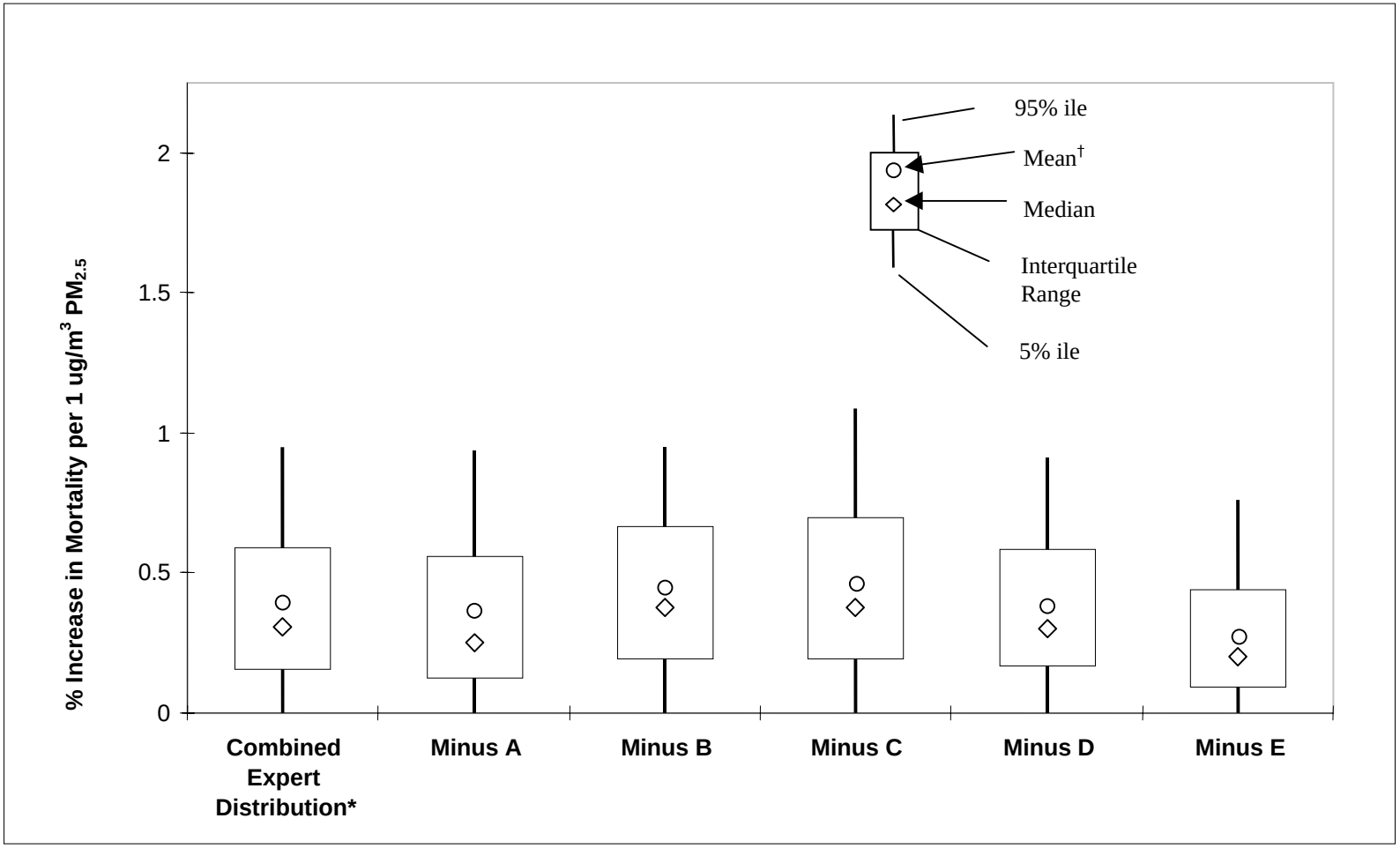
Not surprisingly, Experts C and E appear to have the greatest influence on the combined distribution. Without Expert C, the distribution shifts upward by about 20 percent on average at each percentile. Without Expert E, the combined distribution shifts downward by about 30 percent. The detailed results of the sensitivity analysis, including the percent change at every percentile for each expert, may be found in tables in Appendix D.

To provide some perspective on the combined results, we compared the combined expert distribution to the mortality coefficients reported by the key studies discussed by the experts while developing their subjective judgments (Figure 8). Some of the values from these studies have also been used in various EPA analyses. The Pope et al. (1995) result was used in the first Section 812A prospective analysis (U.S. EPA, 1999). The next is the estimate from Krewski et al.(2000) reanalysis of the ACS data using mean $PM_{2.5}$ for 63 cities, which was used in the RIA for the Heavy Duty Engine/Diesel Fuel Rule (U.S. EPA, 2000). The third distribution is the adjusted relative risk averaged over the 1979-83 and 1999-2000 time periods from the Pope et al. (2002) study. The last distribution from the Krewski et al. (2000) reanalysis of the Six-Cities study.

The distributions presented in the figure for each long-term epidemiologic study show the mean, rather than the median (although most were normally distributed), and the same percentiles as for the combined expert distribution. This comparison provides some indication of the influence of particular studies as well as provides some perspective on the nature and degree of uncertainty expressed in the experts' estimates. Recall that one of the major goals of this study was to try to develop a more complete characterization of uncertainty about the $PM_{2.5}$/mortality relationship than is likely to be represented by statistical error of the mortality coefficient in a particular study. Since the "true" value of the C-R coefficient is unknown, we have no way of knowing how well the experts actually performed.

Figure 7

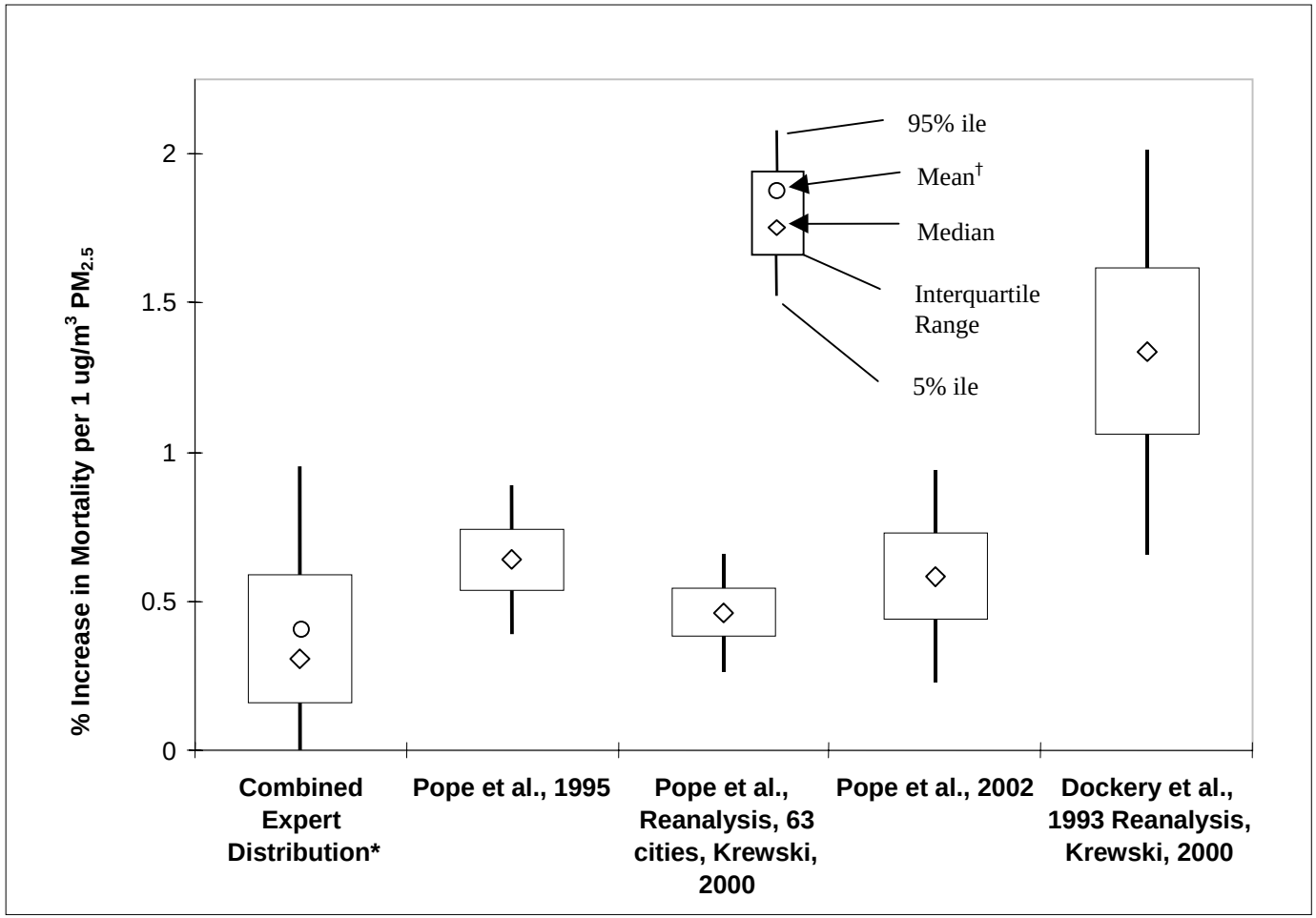
Sensitivity Analysis of Individual Experts Influence on Combined Expert Distribution for Percent Increase in Non-Accidental Mortality Associated with a 1 µg/m^3 Increase in Annual Average $PM_{2.5}$



*The experts' judgments were combined assuming equal weight to each expert and an underlying distribution of population weighted annual average $PM_{2.5}$ concentrations generated from the USEPA's BENMAP model (see Analytical Methods section for details).
†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

Figure 8

Comparison of Combined Expert Judgment Distribution to Results from Selected Studies: Percent Increase in Annual Non-Accidental Mortality Associated with a 1 µg/m^3 Increase in Annual Average $PM_{2.5}$



*The experts' judgments were combined assuming equal weight to each expert and an underlying distribution of population weighted annual average $PM_{2.5}$ generated from the BENMAP model (see Analytical Methods section for details).
†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

The mean of the combined distribution is slightly lower than the means of the various ACS studies, which we believe reflects the greater weight most experts placed on the analyses of the ACS cohort in general as well as the greater weight given to a zero result by Experts B and C. The combined distribution is broader than any of the distributions from the various ACS studies, particularly at the low end of the distribution reflecting the inclusion of zero at the 5th percentile. The 95th percentile of the experts' combined distribution is greater than those of the ACS analyses, but not by a large factor.

Calculation of a standard deviation (SD) for each distribution does give some perspective on the percent increase in uncertainty relative to the standard errors reported in the various studies shown in Figure 8, although experts' distributions are not generally symmetrical. The standard deviation of the combined distribution was roughly 50 percent greater than that from the Pope et al., 2002 study and twice that of the original Pope et al., 1995 study. Or, in terms of variance, the variance of the combined distribution is roughly twice that of Pope, 2002 and about four times that of the original Pope et al. (1995) study.

Effects of Short-term $PM_{2.5}$ Exposure on Mortality

The second set of quantitative estimates IEc elicited from each expert concerned the percent change in total non-accidental mortality associated with a one-day, 10μg/m^3 change in 24-hour average $PM_{2.5}$ that might result from a hypothetical regulatory action. The decrease was assumed to occur relative to a baseline daily average concentration typical of the U.S., defined as somewhere between background levels and 60 μg/m^3.

The "total" mortality described in the question refers to the cumulative change in non-accidental daily mortality that occurs in the day or days following the one-day change in $PM_{2.5}$ concentration. Thus, the question required the experts to aggregate the lagged effects of PM-related mortality, if they felt it appropriate, and express the total effect as a percent change in daily mortality. The question also asked the experts to discuss their views of the appropriate lag period and to make sure that their quantitative estimates reflected those views.

As in the case of the long-term questions, the experts framed their responses to the question in terms of an increase in mortality associated with an increase in daily $PM_{2.5}$, which is consistent with the reporting of values in the epidemiological literature.[28]

To provide a clear picture of the experts' responses, we first discuss the elicited values, then describe briefly the experts' approaches to developing their distributions, and then discuss the key sources of uncertainty discussed by the experts.

[28] The original formulation of the question implicitly assumed that the absolute value of the mortality response to a 10 μg/m^3 decrease of $PM_{2.5}$ is the same as the absolute value of the mortality response to an increase of $PM_{2.5}$ of equivalent magnitude, an assumption about which some of the experts expressed concern. This assumption should be carefully reviewed in any future expert elicitation studies.

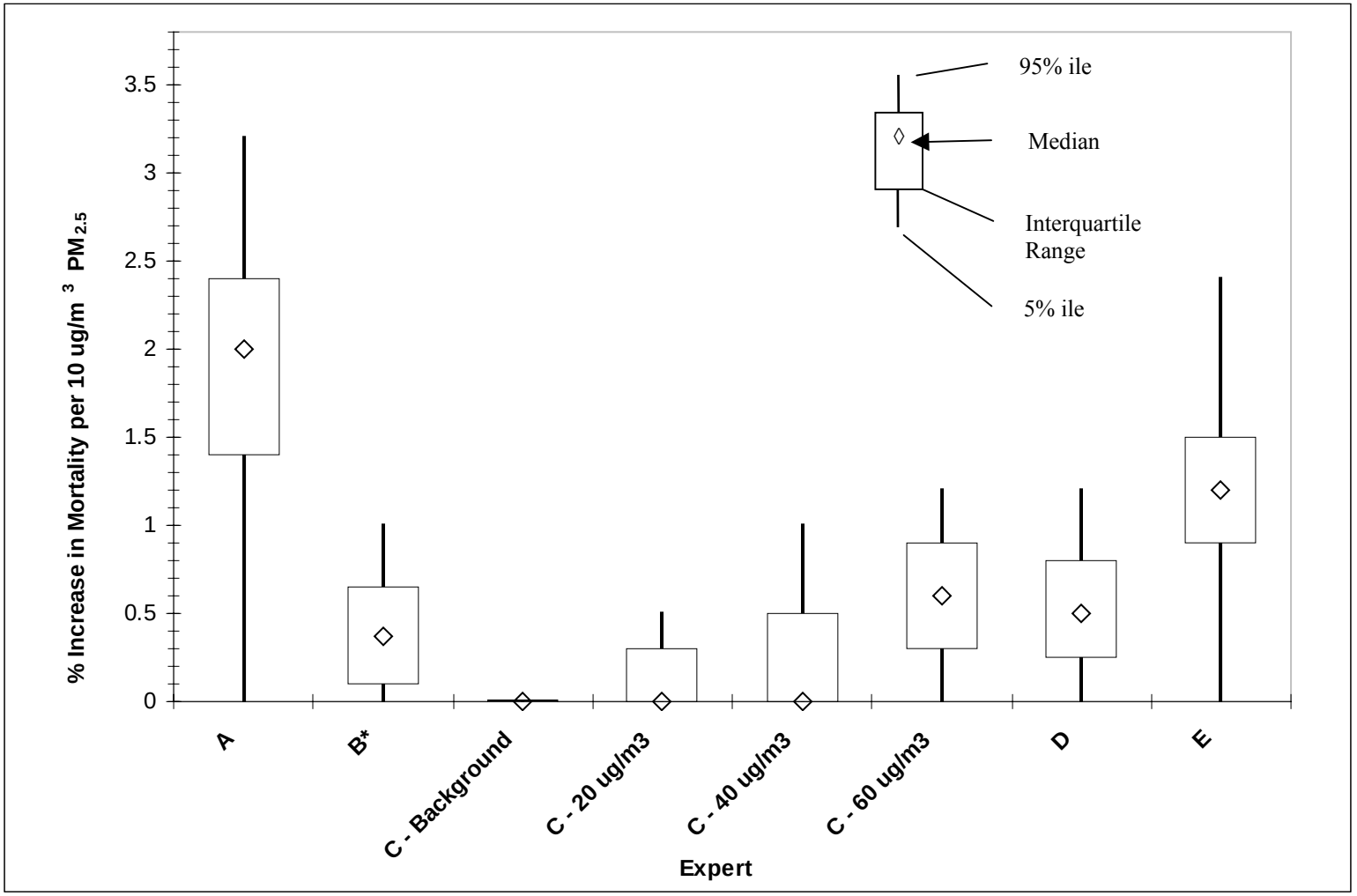
Elicited Values for Individual Experts - Short-Term Exposure Effects

Figure 9 displays the individual distributions given by each of the experts. Their distributions are depicted as boxplots with the diamond symbol showing the median, the box defining the interquartile range, and the whiskers defining the 90 percent confidence interval. As in the case of the effects of long-term exposure, Experts B and C specified different functional forms for the C-R relationship than the other three experts. Expert B again defined an uncertain population threshold and specified an uncertainty distribution for the C-R coefficient at concentrations above this threshold. The distribution in this figure characterizes his distribution for the C-R coefficient above the threshold. Expert C again provided four separate uncertainty distributions for the C-R coefficient at four specific $PM_{2.5}$ concentrations, as representative of points along a continuous, non-linear C-R function. Figure 9 presents the four distributions specified by Expert C. The data for all the distributions in Figure 9 are presented in tabular form in Appendix D.

As Figure 9 illustrates, considerable variation exists in the median value and the spread of uncertainty provided by the experts for the effect of short-term $PM_{2.5}$ exposure on mortality. Their median estimates ranged from 0 percent (no change) to a two percent increase in non-accidental mortality per 10 $\mu g/m^3$ increase in 24-hour average $PM_{2.5}$ concentration. The same types of factors that strongly affected each individual expert's uncertainty distribution for the effects of long-term PM exposures influenced their central tendency and spread for the mortality effects of short-term exposures. These included each expert's assumptions about the key epidemiological results from time-series studies, the strength of the causal relationship, and the nature of the C-R function.

Figure 9

Comparison of Experts' Judgments about the Percent Increase in Daily Non-Accidental Mortality Associated with a One-day 10 μg/m^3 Increase in 24-hour Average $PM_{2.5}$ Concentration (U.S. Baseline: Background to 60 μg/m3)



* Expert B specified this distribution for the PM/mortality coefficient above an uncertain threshold which he characterized as ranging between 4 and 25 with a modal value of 15 ug/m^3.

Key Time-Series Studies. Ultimately, the experts could be separated into two groups: those who relied primarily on results from NMMAPS (Dominici et al., 2003a), and those who relied primarily on other studies. Experts B, D, and E all based their judgments to varying degrees on NMMAPS, with the responses of experts B and D relying most directly on that study.[29] All three argued that the NMMAPS study's use of a consistent methodology and set of assumptions across a large number of cities that were geographically representative of the U.S. provided the most defensible estimate.[30] All three were also willing to adjust the NMMAPS PM_{10}-based estimate to derive an effect associated with $PM_{2.5}$. All the experts who relied on NMMAPS assumed that the $PM_{2.5}$ fraction was likely to be the most important determinant of PM-related mortality. All adjusted their base estimates from NMMAPS based on the $PM_{2.5}/PM_{10}$ ratio in the U.S. (approximately 0.6 to 0.7) (US EPA, 2002).[31] For a $PM_{2.5}/PM_{10}$ ratio of 0.6 for example, the PM_{10} mortality coefficient was multiplied by a factor of 1.7.

Experts A and C relied much less on NMMAPS for their estimates. Expert A more explicitly relied on a broad group of studies to inform his quantitative estimates, including:

> 1) PM_{10} studies, including the body of evidence from both individual-city and multi-city studies (Samoli et al., 2003 (APHEA 2)) and to a lesser extent NMMAPS (Dominici et al., 2003a);[32]
>
> 2) Studies demonstrating a distributed lag effect, in particular Zanobetti and Schwartz (2003) showing a doubling of the mortality effect relative to the effect in studies using one-day lags;
>
> 3) Multi-city $PM_{2.5}$ studies – Burnett and Goldberg (2003) and the Schwartz (2003) time-series analysis of the Six-Cities data; and
>
> 4) The meta-analysis of time series studies by Stieb, Judek and Burnett (2002).

Expert C also placed greater weight on multi-city studies, including Burnett and Goldberg, 2003; Schwartz, 2003 and to a lesser extent, NMMAPS. In forming his judgments, he

[29] Expert B included other multi-city studies (Katsouyanni et al., 2003 (APHEA 2); Schwartz., 2003; Schwartz et al., 1996; Burnett and Goldberg, (2003) in the body of work providing convincing evidence of a relationship between short-term exposures to PM and mortality.

[30] Expert B differed from D and E in his selection of particular NMMAPS model results; Expert B put equal weight on the general linear model and revised GAM results while Experts D and E preferred the revised GAM result. These differences partly account for the overall lower values expressed in Expert B's distribution relative to D and E (See Figure 4).

[31] Some experts used the factor 0.6, the overall U.S. average. Others wanted to use a slightly higher value, (0.65) that they thought was more representative of urban areas.

[32] Expert A placed less emphasis on NMMAPS than other experts had, expressing concern about three issues: 1) potentially restrictive use of a common set of modeling assumptions across individual cities where weather and seasonal patterns might differ substantially and where optimization for particular cities would be warranted; 2) the inability of the study to look at patterns of exposure other than every sixth day; and 3) the lack of accounting for possible distributed lag effects.

relied primarily on the Burnett and Goldberg, (2003) study of eight Canadian cities because of its use of $PM_{2.5}$ as an exposure measure and because of the methods used to control for weather.[33]

Causality for Short-Term Effects. As was the case for effects of long-term exposure, the experts' distributions for short-term effects of $PM_{2.5}$ exposure on mortality tended to reflect their views on the likelihood of a causal relationship between short-term exposures to $PM_{2.5}$ and increased mortality. Although most of the experts assigned a greater probability to the likelihood of increased mortality resulting from short-term exposures than they did for long-term exposures, all of the experts again placed at least a five percent probability on the true value of the effect being zero (i.e., no causal relationship; see the qualitative results section for a more detailed discussion of their rationales).

The impact of an expert's causality views on his elicited distribution is most apparent in the case of Expert B. Expert B had assigned only 50 percent likelihood to a causal relationship between mortality and long-term exposures but was roughly 75 percent sure that a causal relationship existed for short-term exposures and mortality. While his median estimate was zero for the effect of long-term exposures, both the 25^{th} and 50^{th} percentiles of his estimates for the effect of short-term exposures were above zero, even when taking into account the possible existence of a population threshold.

Expert C's probability of causality varied with the baseline PM concentration. As his distributions illustrate, he thought a causal relationship was highly likely at the upper end of the range of daily average $PM_{2.5}$ concentrations but highly unlikely at the mid to lower end of the range. Because the causality question essentially required that the experts base their responses on the full PM range specified in the question, he set his likelihood value at 50 percent, midway between zero at the low end and a value close to 100 at the high end.

Shape of the C-R Function for Short-Term Effects. Experts A, D, and E assumed that the C-R function was log-linear, consistent with the available epidemiological evidence. Expert E qualified his assumption of log-linearity by noting that he believes the C-R function to be basically monotonic over the range, but that it is not possible to distinguish the precise mathematical form of the function; in light of this limitation, the log-linear model is a reasonable choice.

Expert B agreed that the relationship between mortality and $PM_{2.5}$ concentration was log-linear, but primarily at higher concentrations above a threshold level. (He also believed it likely that the slope of the response would change with decreasing concentration but that he did not believe there exists sufficient data to define the nature of that change.) As in the case of long-term exposures, Expert B acknowledged that few data existed to support the identification of the specific level at which a human population threshold might exist, but felt such a model was justified on biological and mechanistic grounds. He defined an uncertain threshold, with a distribution ranging between background levels (about 4 $\mu g/m^3$) to 25 $\mu g/m^3$, with a modal value of 15 $\mu g/m^3$.

[33] Expert C expressed uncertainty about assuming $PM_{2.5}$ is the sole contributor to the mortality associated with PM_{10}. He suggested that we do not have sufficient evidence to rule out the role of the coarse $PM_{10-2.5}$ fraction.

Expert B's distribution in Figure 9 is his distribution of uncertainty about the C-R coefficient above the (uncertain) threshold. His 'effective' distribution for the full range of daily $PM_{2.5}$ concentrations is shown for comparison with the combined expert distribution in Figure 10 later in this section. As described in the Analytical Methods section, his effective distribution is estimated by probabilistically combining his uncertain threshold distribution and his reported C-R coefficient distribution above the threshold. The effect of incorporating the threshold has the effect of shifting the overall distribution downward.

Expert C also believed that a one-day increase in $PM_{2.5}$ would likely have a decreasing impact on mortality with a decreasing baseline daily $PM_{2.5}$ concentration. He argued that the linear models primarily reflect limitations in the statistical methods available to investigators and that they are not likely to be consistent with the underlying biology. Consequently, he specified a non-linear C-R function, specifying distributions for the instantaneous slope of the curve at four discrete points in the $PM_{2.5}$ range for this assessment. The four distributions are shown in Figure 9. He assumed that his estimate of the slope, and the uncertainty distribution about it, would be changing continuously between each set of points.

Approaches to Developing Distributions for Short-Term Effects

Most of the experts took an informal approach to weighing the evidence from the studies they relied on, considering the uncertainty represented by the statistical error reported in the studies and by the limitations of the studies that they identified. (The one exception is Expert E, who tended to work explicitly with means and standard errors to develop his first set of estimates.) However, even within this small group, experts took quite different approaches to the choice and weight given to particular studies.

Although the elicitation team encouraged experts to begin answering the quantitative questions by considering the minimum and maximum values their distributions might take, the experts often preferred to begin with the results of particular studies and to make adjustments of those values to account for sources of potential bias and uncertainty. Given the weight of epidemiological and other evidence for the effects of short-term $PM_{2.5}$ exposures and the degree of confidence individuals expressed in particular studies; beginning with the data was clearly more intuitive and comfortable for the experts.[34]

The experts did easily set the minimum and 5th percentiles of their distributions. Experts A, B, D, and E set minimums and 5th percentiles at zero, consistent with their views on the likelihood of a causal relationship between short-term spikes in $PM_{2.5}$ and mortality. Expert C's distributions are tied less directly to his discussion of causality (he gave an overall estimate of 50 percent for the full range asked about in the assessment), given that he expressed decreasing degrees of confidence in a causal relationship with decreasing daily $PM_{2.5}$ concentrations. Expert C did consistently specify minimum and 5th percentile values of zero across the $PM_{2.5}$ concentration range, however.

[34] We recognize that theory holds that this heuristic of "anchoring and adjustment" is likely to lead to overconfident judgments but found that it was, in practice, very difficult to avoid.

Setting the values for the other percentiles was a more complicated endeavor for the experts, who employed very different approaches. As noted earlier, Experts B, D, and E each relied strongly on NMMAPS reanalysis results (Dominici et al., 2003a) as the starting point for their estimates. Their percentile estimates reflect varying degrees of explicit adjustment to the NMMAPS results for three factors (plus minor adjustments for weight given to the results of other studies):

1) conversion from a PM_{10} to $PM_{2.5}$-based effect;

2) upward adjustment to allow for possible distributed lag effects; and

3) upward adjustment for overaggressive (i.e., over-conservative) control for seasonality or temporal confounding in the NMMAPS analytical approach.

Of the three experts, Expert E was the only one to adjust explicitly for each of these three factors, beginning first with the mean and standard error from the NMMAPS reanalysis (using the GAM with stringent convergence criteria); adjusting upward (roughly a factor of two) to account for distributed lag effects; adjusting upward again slightly to account for possible overaggressive statistical control for seasonality; and finally converting these values to reflect the role of $PM_{2.5}$. Expert E also adjusted his 95^{th} percentile value upward to reflect some weight given to other studies (e.g. Schwartz, 2003 - six cities and Burnett and Goldberg, 2003) that reported substantially higher percent increases in mortality. Expert B adjusted for conversion from a PM_{10} to $PM_{2.5}$-based effect, and also adjusted his 75^{th} and 95^{th} percentiles upward to allow for possible distributed lag effects and overaggressive control for seasonality, but did not change his central estimates. Expert D, who also started with the GAM reanalysis results, only adjusted for conversion from a PM_{10} to $PM_{2.5}$-based effect. He acknowledged the evidence for distributed lag effects but did not want to make an explicit adjustment to his estimate. He preferred that an assumption about the size of the distributed lag effect be incorporated afterward, as part of the benefits analysis where these values would be applied. He also did not believe it necessary to adjust his estimates for overaggressive control for seasonality.

Expert C also relied primarily on one study as the underlying basis for his quantitative results, although he preferred the $PM_{2.5}$ (one-day lag) results from the Burnett and Goldberg, (2003) study in eight Canadian cities. While he expressed reluctance to rely primarily on one study, he cited his confidence in the statistical analysis of the study - in particular, in its control for weather. He also felt it important to rely on a multi-city study using $PM_{2.5}$ directly as a measure of exposure. He argued there was still considerable uncertainty about the causal role of the $PM_{10-2.5}$ fraction and that the composition of that fraction might vary between cities. He suggested that the NMMAPS study be considered only as qualitative evidence for a $PM_{2.5}$ effect. The evidence from other studies for a distributed lag effect was discussed, but he did not find it sufficiently convincing to adjust his estimates.

Expert A's approach to characterizing uncertainty about the slope of the C-R function differed markedly from those of the other experts in terms of both his efforts to undertake two independent approaches to answering the question and the breadth of studies he considered.

Expert A began by using two sets of evidence, one from PM_{10} studies and the other from $PM_{2.5}$ studies, to develop independent estimates of his possible median. Both approaches reached as similar estimate, a two percent increase in mortality per 10 $\mu g/m^3$ increase in daily $PM_{2.5}$. The first approach involved PM_{10} studies including the APHEA 2 study of 29 European cities (Samoli et al., 2003) and NMMAPS. Beginning with a single day effect of about 0.7 percent per 10 $\mu g/m^3$ increase, he adjusted downward to 0.55 percent to reflect some weight on NMMAPS. He doubled this estimate to account for multi-day lag, citing Schwartz, 2000b and Zanobetti and Schwartz, 2003 showing a two- to three-fold increase in the number of deaths when the analysis was extended out 30-40 days. Like the other experts, he multiplied this result by a factor of about 1.5 to adjust from PM_{10} to yield a central estimate of about 1.7 percent increase in mortality per 10 $\mu g/m^3$ increase in daily $PM_{2.5}$.

Expert A's second approach involved $PM_{2.5}$ time-series studies. Expert A equally weighted the one-day lag results from $PM_{2.5}$ multi-city studies by Burnett and Goldberg, (2003) and Schwartz (2003) to yield a central estimate of about 1.2 percent, which, when doubled for multiple-day effects, produced an estimate of about 2.4 percent. Expert A noted that both studies have limitations for extrapolation to a mortality effect for the entire U.S. They are not representative of all cities and regions in the U.S., for example, lacking cities where people spend more time outdoors or use more air conditioning.

Expert A set his 95th percentile at 3.2 percent to include the upper bound of the Schwartz estimates (using both natural and penalized splines), adjusted for distributed lag effects. He ultimately established his lower 25th percentile at a level that he felt gave some weight to NMMAPS (adjusted for multi-day lag and for $PM_{2.5}$) and adjusted his 75th percentile to give weight to the original central estimates of the Schwartz (2003) study.

Sources of Uncertainty – Effects of Short-Term Exposures

Experts were asked at several times during the interview, particularly in the preliminary questions, to list the key sources of potential bias and/or uncertainty in current evidence they might rely on to estimate the potential impact on total mortality of a one-day, 10 $\mu g/m^3$ increase in short-term exposure to $PM_{2.5}$. In the context of the quantitative discussion, each was asked to list the top five sources of bias and/or uncertainty and to think about how each factor would affect his estimates.

The tables summarizing the factors identified by each expert may be found in Appendix F. The experts listed many of the same factors; for some they were sources of bias and for others, they were sources of uncertainty. If an expert had a clear sense for the direction a particular factor might have on published results, he was likely to list it as a source of bias; otherwise he identified it as a source of uncertainty.

Some of the common factors that experts thought might be a source of bias in the analytical results of current studies included:

- **Treatment of weather in the analysis.** This would include, for example, concern about proper control for temperature and relative humidity. Most

experts felt that weather had been adequately dealt with in time series analyses. Expert C expressed the greatest concern, citing a 2003 HEI committee report stating that "neither the appropriate degree of control for time or appropriate specification for the effect of weather has been identified for the time-series studies." His concern was that inadequate control might lead to a positive bias in the mortality estimates.

- **Treatment of seasonality or control for seasonal effects.** Three experts believed that NMMAPS investigators had been conservative in their control for seasonality leading to results that probably underestimated the effect of PM_{10} or $PM_{2.5}$. One expert who relied on NMMAPS for his quantitative estimates listed this factor as a source of uncertainty.

- **Analytical or investigator bias.** Several of the experts believed that modeling approaches and choice of assumptions could be prone to subtle biases in favor of finding positive outcomes. Most experts were referring to time-series results in single cities although some thought some multi-city studies (not including NMMAPS) might also be affected. No one offered insight into the likely magnitude of this potential source of bias. Some simply listed this issue as a source of uncertainty.

- **Accounting for distributed lag effects.** Three of the five experts felt that any estimate of mortality should include effects over multiple days and adjusted their estimates accordingly.

The more common sources of uncertainty included:

- **Co-Pollutants.** Experts identified uncertainty related to the role of co-pollutants and whether any of these were either confounders or effect modifiers;

- **The effect of PM composition and chemistry on toxicity.** Though this was recognized as a source of uncertainty, the limited data available to describe any differential toxicity of specific PM components precluded experts from giving quantitative estimates of the effect of this uncertainty on the PM C-R coefficient.

- **Measurement error.** The experts used this term to cover many issues relating to how well the measurements used in the studies correlated with "true" exposures to individuals. These included the use of central site monitors for personal exposures; and the impact of regional differences in activity patterns, air conditioning use, and housing construction (impact on indoor/outdoor penetration). One expert cited measurement error as a potential source of bias as well as uncertainty.

Although not always explicitly mentioned by individual experts, the strength of the causal relationship was an important source of uncertainty as reflected in their willingness to assign at least 5 percent probability to the possibility that the percent increase in mortality associated with short-term exposures might be zero.

As discussed in the long-term exposure section, the largely qualitative structure of these questions did not always promote as clear a quantitative analysis of these factors as would have been desirable. Experts did adjust explicitly for conversion of PM_{10} to $PM_{2.5}$ results and for distributed lag effects if they felt it appropriate, but were reluctant or unable to offer estimates of the magnitude of the impact of other potential sources of bias and uncertainty. Nevertheless, their estimates of the percent increase in mortality per 10 $\mu g/m^3$ increase in $PM_{2.5}$ do appear to reflect these uncertainties to some degree as the distributions are generally broader than statistical uncertainty alone would dictate.

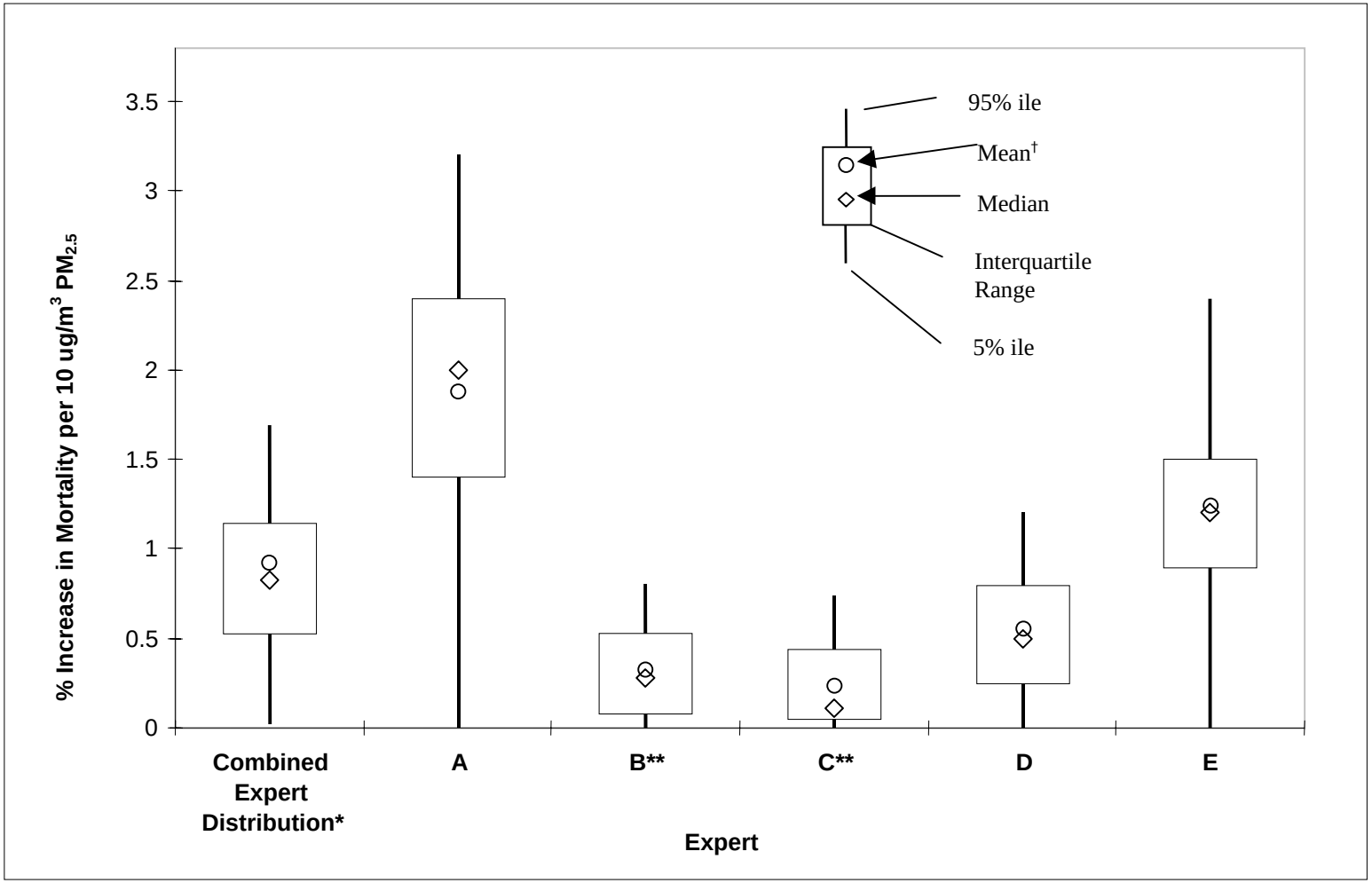
Combined Results of Experts – Effects of Short-term Exposure

As described in the analytical methods section, we combined the experts' judgments using equal weights. We then evaluated the distribution in two ways, comparing it to the judgments of individual experts and evaluating the influence of individual experts. In order to provide perspective on the amount of uncertainty reflected in the combined judgments of the experts on our panel, we compared the combined expert distribution to the mean and standard errors reported in the primary time-series studies on which the experts relied for their judgments.

The combined expert distribution describing the collective uncertainty in the mortality effects of short-term exposures appears in the first box plot of Figure 10. The individual expert distributions are arrayed to the right. Both Expert B's and C's distributions are their 'effective' distributions for the range of daily $PM_{2.5}$ values specified in the question. These were calculated based on the assumption of an underlying uniform distribution of daily $PM_{2.5}$ concentrations (i.e., all baseline concentrations within the range specified in the protocol are equally likely). Numerical values for these figures can be found in Appendix D.

Figure 10

Comparison of Combined Expert Judgment to Experts' Individual Judgments about the Percent Increase in Daily Non-Accidental Mortality Associated with a One-day 10 μg/m^3 Increase in24-hour Average $PM_{2.5}$ Concentration (U.S. Baseline: Background to 60 μg/m3)



* The distributions for Expert B and expert C are their expected distribution for the $PM_{2.5}$ range specified. Prior to combining their results with other experts, Monte Carlo sampling was used to was used to calculate the expected values of Experts B and C's distributions for the full range assuming an underlying uniform distribution background to 60 μg/m3 (see Analytical Methods section for details).

** The experts' judgments were combined assigning equal weight to each expert.

†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

The combined expert distribution has a median value of a 0.83 percent increase in mortality per 10 $\mu g/m^3$ increase in daily average $PM_{2.5}$, a mean of 0.84 and a 90 percent confidence interval of 0.02 to 1.69. Despite the relatively non-symmetrical distributions from Experts B and C in particular, the combined distribution is quite symmetrical. The means for these experts were higher than their median suggesting some skewing of their distributions. It also reflects the effect of truncating the lower halves of their distributions at zero.

The section on the effects of long-term exposures discussed how collapsing the experts' distributions into one distribution for the full range of $PM_{2.5}$ exposures obscures the influence of Experts B and C whose C-R coefficients are conditional on the baseline concentration. The same issue is relevant for the effects of short-term exposures. Figure 11 illustrates the concentration dependence of the combined distribution by plotting the distribution at four baseline annual average $PM_{2.5}$ concentrations: 8, 20, 40, and 60 $\mu g/m^3$. The progressive downward shift in the distribution for the C-R coefficient as one moves from a baseline $PM_{2.5}$ concentration of 60 to a baseline $PM_{2.5}$ concentration of 8 reflects Expert B's and Expert C's higher probabilities of zero effect of PM on mortality at lower concentrations.

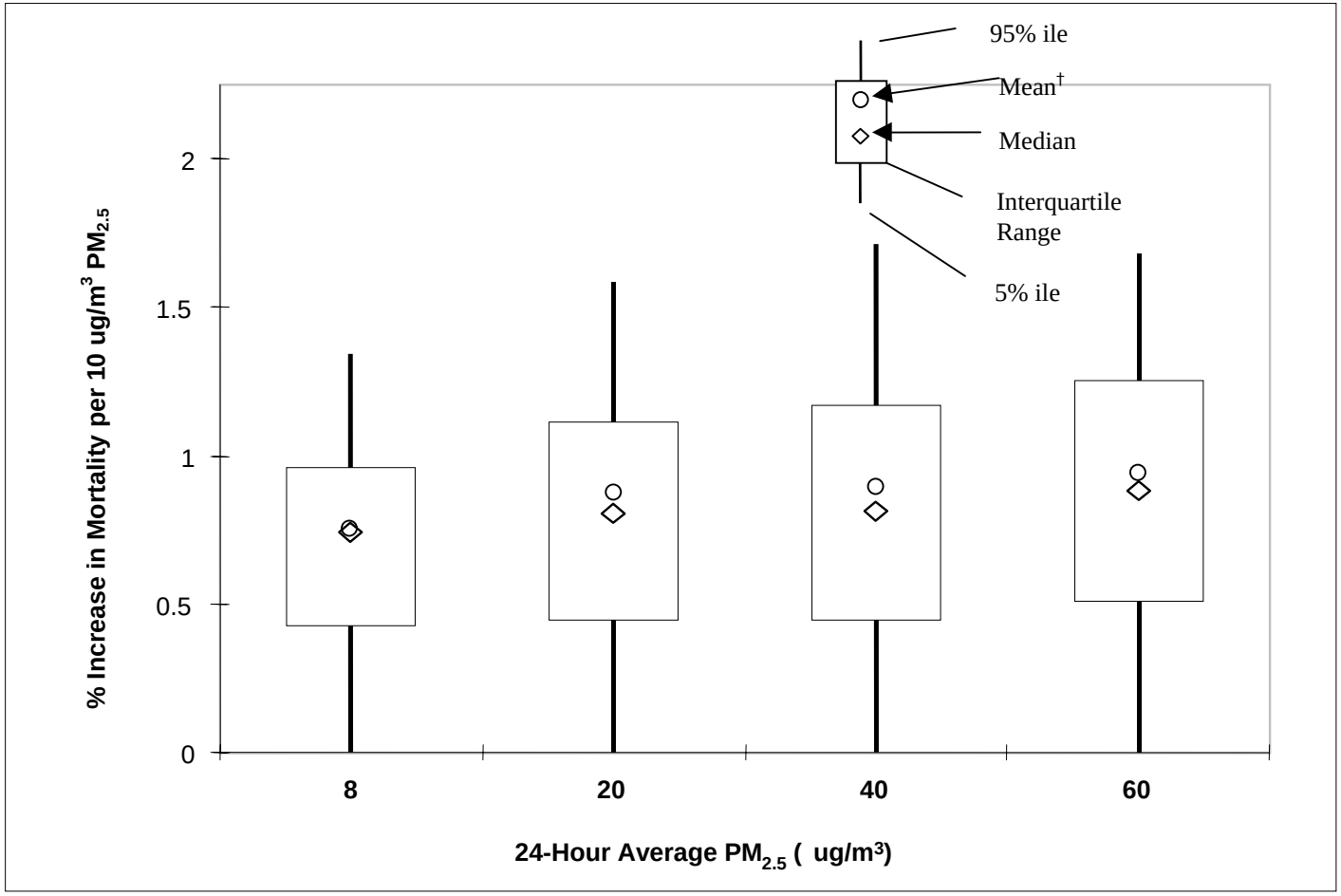
As in the case of the long-term exposure question, the concentration dependence of the combined distribution limits the applicability of this specific combination to other analyses. Because B's and C's distributions are conditional on concentration, the form of the underlying distribution used to generate their "effective" distributions has implications for using the combined distribution in benefits analysis. We have not compared alternative baseline distributions as we did for long-term exposures in Figure 6, but the same argument applies in this case. The equal weighting assumption in the combined distribution only holds if the underlying distribution used to adjust Expert B and C's distribution is the same as that used in the ultimate analysis. Otherwise, the equal weighting assumption is not likely to hold. This potential error can be avoided by applying each expert's distribution separately in a benefits analysis and then pooling the results with equal weights.

As Figures 9 and 10 show, individual experts' characterizations of their uncertainty about the effects of increases in short-term exposures to PM2.5 are dramatically different (compare for example Expert A versus Expert B or C). We conducted a sensitivity analysis to evaluate the influence of individual experts on the combined result by sequentially eliminating one expert at a time from the calculation of the combined distribution. The results of this analysis are shown in Figure 12; the combined expert distribution including all five experts is shown first, followed by the sequence of five combinations, each omitting one expert.

The impact of particular experts is more marked for the short-term exposure analysis than for the long-term exposure. Eliminating Expert A resulted in the largest percent drop at all percentiles: values at the upper percentiles declined by 40 percent and by 30 percent at the lower percentiles. Eliminating Expert B and C led to increases in the estimate of the C-R coefficient of approximately 15-20 percent. Details results of the sensitivity analysis may be found in tables in Appendix D.

Figure 11

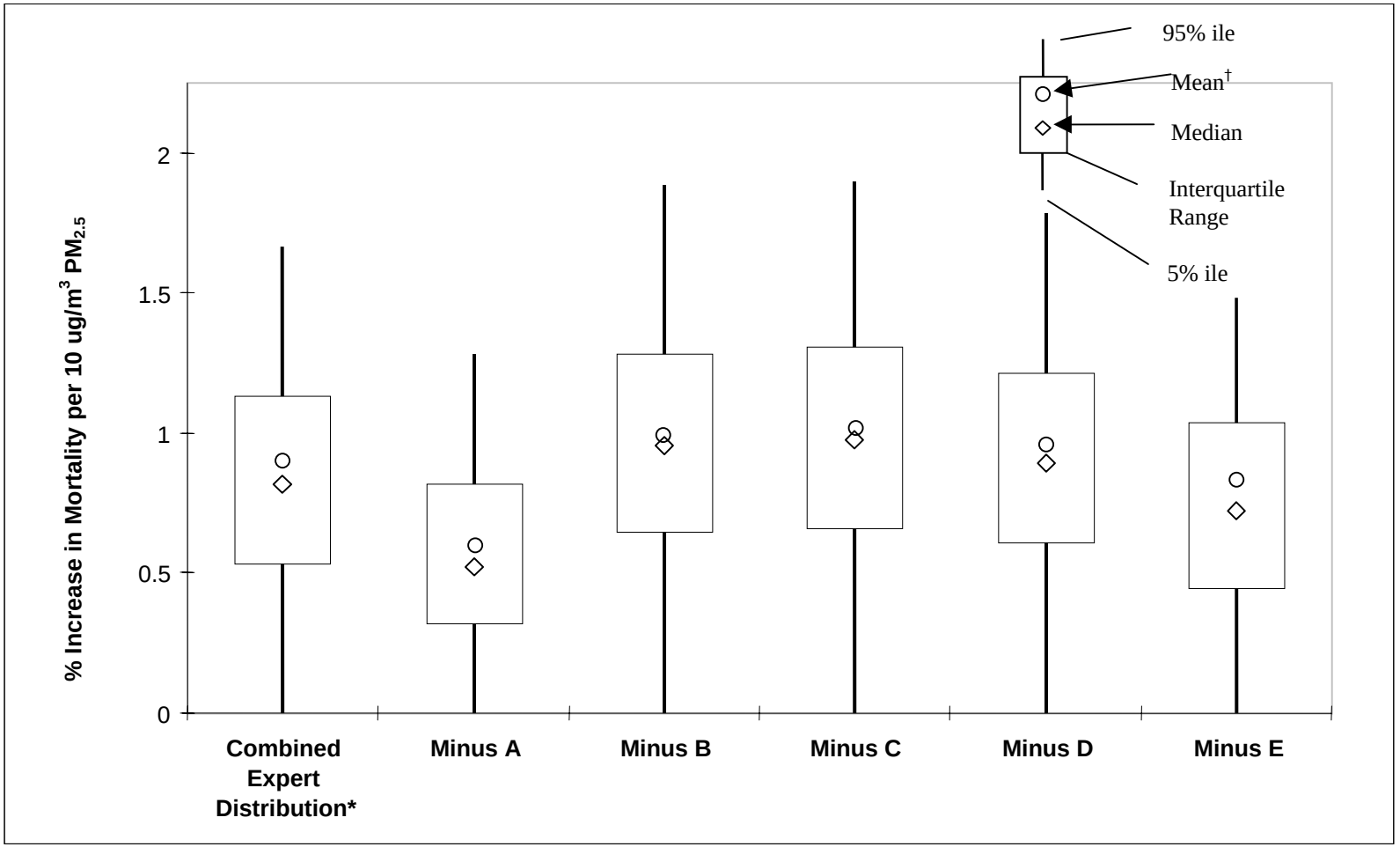
Combined Expert Judgment Distributions for Percent Increase in Annual Non-Accidental Mortality Associated with a 10 µg/m^3 Increase at Specific Baseline 24-Hour Average $PM_{2.5}$ Concentrations*



*The experts' judgments were combined assuming equal weight given to each expert.
†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

Figure 12

Sensitivity Analysis of Individual Experts on Combined Expert Distribution for Percent Increase in Non-Accidental Mortality Associated with a One-Day 10 µg/m^3 Increase in 24-hour Average $PM_{2.5}$



*The experts' judgments were combined assuming equal weight to each expert.

†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

In order to offer some perspective on the combined distribution we plotted the experts' combined distribution alongside some of the principal studies cited by the experts as the basis for their quantitative estimates. We have presented the distributions from the studies in two ways based on the differing assumptions made by experts about the role of multi-day or distributed lag in estimating the magnitude of the mortality effect of short-term exposures. Figure 13 compares the experts' combined distribution to the one-day lag results of the Schwartz (2003) analysis of the Six-cities data, the Burnett and Goldberg, (2003) analysis of the eight Canadian cities and the Dominici et al. (2003a) analysis of the NMMAPS data. The NMMAPS results were adjusted to reflect $PM_{2.5}$ exposures using a conversion factor based on the average national $PM_{2.5}$/ PM_{10} ratio as discussed in the individual results. In Figure 14, the original study estimates have been adjusted upward by a factor of two in accordance with adjustments made by Experts A, E, and to a more limited extent, Expert B, to account for distributed lag effects. The U.S. EPA used a similar two-fold adjustment to the Schwartz, 1996 results (the pre-revised GAM version of the 2003 paper) in its alternative analysis for the Non-Road rule.

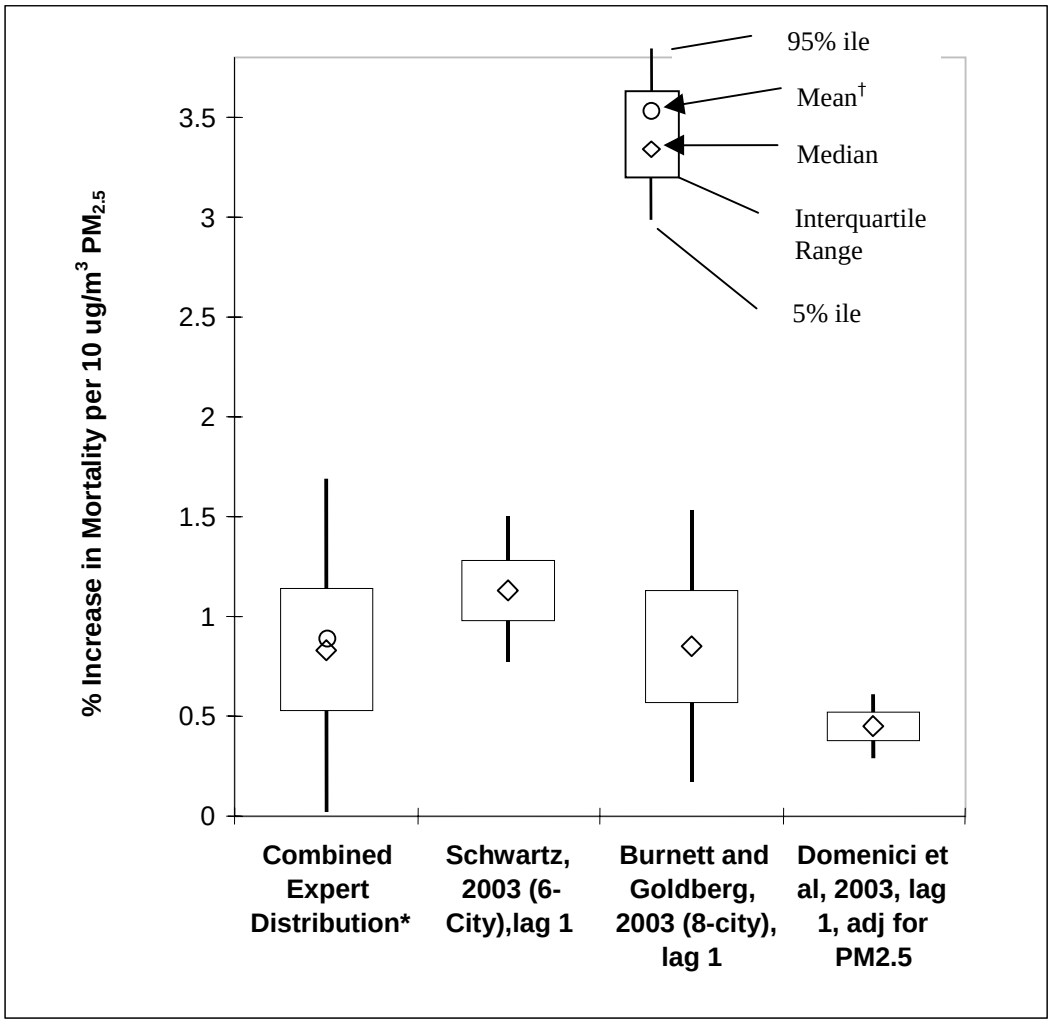
The combined expert distribution falls within the bounds of the range of study results clearly reflecting the confidence most experts placed on, the published data. It is more similar in magnitude to the one-day lag results than to the distributed lag effects estimates from the Burnett and Goldberg (2003) study, although its median is the same as the mean NMMAPS result adjusted for distributed lag and $PM_{2.5}$. Despite the variation apparent in the individual expert distributions, the combined distribution is relatively narrow; we did not observe the phenomenon reported in other studies (Evans et al., 1994) where disparate judgments lead to very broad non-informative distributions.[35]

At the same time, the experts' combined distribution is broader, reflecting more uncertainty, than most of the published distributions used in our comparison. The one exception is the Burnett and Goldberg, (2003) distribution, adjusted for possible distributed lag effects, but only one expert (A) relied to any degree on this study so it received relatively little weight in the final distribution. Ultimately, the combined expert distribution reflects the influence of numerous factors not only these studies and the experts' assumptions about distributed lag; therefore, it is important to be cautious about drawing conclusions about the quality or calibration of their collective uncertainty assessment.

[35] We are aware we cannot rule out overconfidence bias.

Figure 13

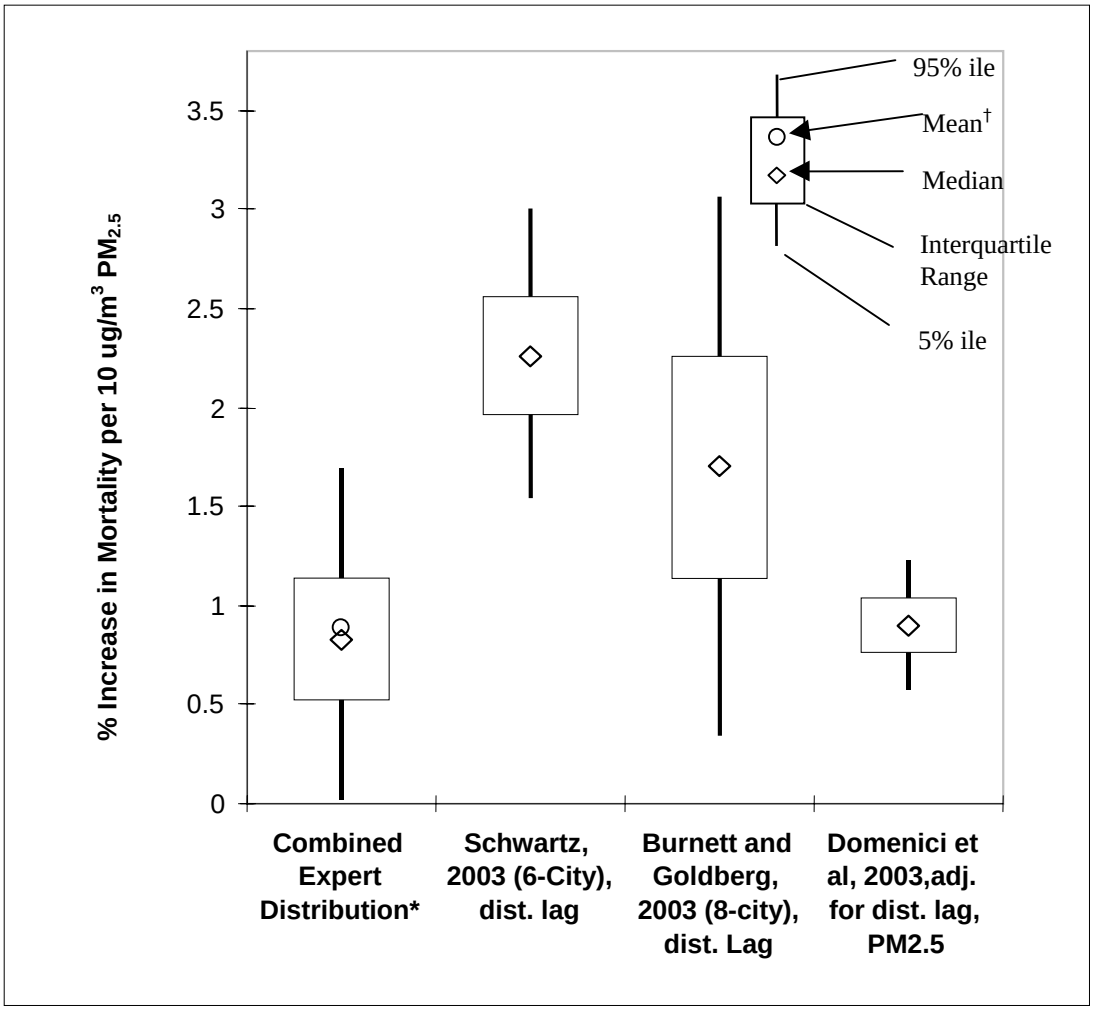
Percent Increase in Daily Non-Accidental Mortality Associated with a One-day 10 μg/m^3 Increase in 24-hour Average $PM_{2.5}$: Comparison of Combined Expert Judgment Distribution To Results from Selected Studies with 1-Day Lag



Notes:
Schwartz, 2003. Six cities study, PM2.5. Reanalysis using GAM with stringent convergence criteria.
Burnett and Goldberg, 2003. Eight Canadian cities, $PM_{2.5}$. Reanalysis using GAM with stringent convergence criteria
Domenici et al., 2003. NMMAPS reanalysis using GAM with stringent convergence criteria, adjusted to reflect $PM_{2.5}$ using $PM_{2.5}/PM_{10}$ ratio=0.6.
* Individual expert judgments combined using equal weights, as described in text
†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

Figure 14

Percent Increase in Daily Non-Accidental Mortality Associated with a One-day 10 μg/m3 Increase in 24-hour Average $PM_{2.5}$: Comparison of Combined Expert Judgment Distribution To Results from Selected Studies with Distributed Lag



Notes:
Schwartz, 2003. Six cities study, $PM_{2.5}$. Reanalysis using GAM with stringent convergence criteria.
Burnett and Goldberg, 2003. Eight Canadian cities, $PM_{2.5}$. Reanalysis using GAM with stringent convergence criteria.
Domenici et al., 2003. NMMAPS reanalysis using GAM with stringent convergence criteria, adjusted to reflect $PM_{2.5}$ using $PM_{2.5}/PM_{10}$ ratio=0.6.

All individual study results adjusted by factor of two to account for possible distributed lag effects (Schwartz, 2000; Zanobetti and Schwartz, 2003) consistent with some expert opinions as noted in text.
* Individual expert judgments combined using equal weights, as described in text.
†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

4. CONCLUSIONS

The pilot assessment yielded several useful insights about using expert judgment to evaluate uncertainty in the $PM_{2.5}$/mortality relationship and identified some key areas on which to focus in subsequent efforts. We group our conclusions into two general categories: insights from evaluation of the quantitative results and insights from an evaluation of this pilot assessment process. In evaluating the pilot elicitation process, we summarize both what elements of the project appeared to work well and what elements may need improvement. We close with some recommendations to consider for future expert judgment elicitation work.

Quantitative Results

The goal of this expert judgment pilot project was to evaluate the use of expert subjective judgment as a method for providing a more complete characterization of state of scientific uncertainty in the C-R relationship between increased exposures to $PM_{2.5}$ and increased mortality. The current EPA approach to characterizing uncertainty in the $PM_{2.5}$/mortality C-R function relies primarily on the statistical error reported in selected epidemiological studies. In this assessment, the experts were asked to consider and reflect in their judgments the impact of several other sources of uncertainty that have been identified but that are not typically represented quantitatively into uncertainty estimates. Examples of these other uncertainties include the strength of the causal relationship, uncontrolled confounding, effect modification, errors in exposure measurement, the role of PM components, and the role of co-pollutants. The results of this pilot assessment reflect clearly the influence of the primary epidemiological studies the experts used in forming their judgments. However, they also suggest that experts did incorporate other sources of uncertainty beyond statistical error in their predictions of the $PM_{2.5}$–mortality coefficients for long- and short-term exposures.

In developing their estimates of the percent increase in annual, non-accidental mortality associated with a 1 $\mu g/m^3$ increase in annual average $PM_{2.5}$, the experts had high praise for both the original studies and reanalyses of the Six Cities data (Dockery et al., 1993; Krewski et al. 2000) and the ACS cohort data (Pope et al., 1995; Krewski, 2000; and Pope et al. 2002). However, most appeared to placed greater weight on the ACS analyses, in particular the most recent study including the extended years of follow-up (Pope et al., 2002). The mean of the combined distribution (0.4 percent change in mortality per 1 $\mu g/m^3$ change in annual mean $PM_{2.5}$) is about 33 percent lower than the average risk estimate for mortality reported in that study for exposure over both time periods evaluated (1979-1983 and 1999-2000). However, the mean of the combined distribution closely matches the lower risk estimate reported in Pope et al., 2002 based on exposures in the first time period only. The concordance of these two values does not necessarily reflect a preference for the latter risk estimate from the ACS study on the part of the experts. Instead, the lower mean reflects several other factors affecting the various experts' judgments including the greater probability they all placed on a C-R coefficient of zero, and the use of a log-linear threshold model and a non-linear threshold model by two of the experts.

We were initially concerned that some of the individual expert distributions and the combined expert distributions reflected virtually no weight on the Six Cities study results, despite statements from all the experts acknowledging the quality of the Six Cities study design. To evaluate that concern, we developed a series of distributions based on different levels of weight or probability assigned to the Krewski et al., 2000 reanalysis of the Dockery et al., 1993 Six Cities and the Pope et al., 2002 reanalysis of the ACS studies (data not shown). The analysis suggested that the experts' judgments were consistent with putting a modest amount of weight (10-25 percent) on the Six Cities study. Weighting by inverse variance is equivalent to placing about 11 percent probability on the Six Cities study.

To provide some perspective on the amount of overall uncertainty reflected in the combined distribution, we compared the variance of the combined expert distribution to one estimated from an inverse variance weighting of the Six-Cities study and ACS study results. The variance of the combined distribution is nearly three times that of the pooled estimate of variance from the two studies. Such a comparison suggests that the spread of the combined distribution represents more than just uncertainty arising from statistical error reported in the major studies. This result is encouraging. However, in absence of knowledge about the "true" C-R relationship, it is not possible to know whether this expression of uncertainty represents a well-calibrated representation of the experts' state of knowledge.

The experts' characterizations of uncertainty about the percent increase in daily mortality associated with a 10 $\mu g/m^3$ increase in 24-hour average $PM_{2.5}$ are difficult to compare to any one or two specific studies. Three experts relied primarily on NMMAPS data but differed in their decisions to adjust the results to account for distributed lag effects. Of the two remaining experts, one relied primarily on the Canadian Eight City study (Burnett et al. 2000; Burnett and Goldberg, 2003) while the other used all of these studies as well as results from APHEA2 (Samoli et al., 2003) and Schwartz (2003). The combined distribution is very much a hybrid of the results of these studies, different experts' assumptions about the existence of thresholds, the shape of the dose response relationship, and the likelihood of a causal connection between exposure to $PM_{2.5}$ and mortality.

The questions regarding the degree of overlap between the mortality effects estimated by the cohort and time-series studies highlighted a potentially important source of uncertainty for the full benefits analysis. The questions were originally intended to help the experts eliminate any effects of short-term exposure from their estimates of the effects of long-term exposures but none of the experts used them for that purpose. All of the experts essentially estimated a "cohort" effect for their estimates of long-term mortality effects, reflecting their reliance on evidence from cohort studies. The quantitative estimates the experts gave regarding the fraction of the cohort effect that represents time-series effects and the percent of time-series effects that are not captured by the cohort studies could be used to isolate an effect of long-term exposures. However, the uncertainty experts expressed about answering these questions was significant and may not be adequately represented in the simple quantitative estimates they were asked to give. Thus, we caution against using these results for this purpose.

Our interviews suggest that it may be premature to focus much effort on the issue of differential toxicity in PM components, given the limited literature base on which experts can

currently draw. With few exceptions, the experts did not find existing evidence compelling enough to cause them to adjust the magnitude of or uncertainty in their C-R estimates. It is important to note that their responses to the differential toxicity questions were at least partly limited by the placement of the questions in a follow-up section to the elicitation; the questions did not receive as much time and attention. EPA may wish to evaluate whether there exist other experts, or other data, that would allow a more complete exploration of this issue in any subsequent elicitation efforts.

The combined expert judgment distributions presented in this study must be considered as illustrative of the panel's collective uncertainty and are not recommended for use in policy analysis. The equal weighting assumption in the combined distribution only holds if the underlying $PM_{2.5}$ distributions used to create the single distributions for Expert B and C are the same as the distributions used in the particular benefits analysis in which they are applied. Otherwise, depending on the form of the underlying distribution, implicitly more or less weight may be given to Experts B and C in the combined distribution, thereby distorting the benefits analysis.

The preferred approach is to use the individual experts' uncertainty distributions in conjunction with the specific $PM_{2.5}$ distribution for a particular application and then pool the results using equal weights. In this approach, the analyst allows each expert's C-R function to be estimated in the benefits model independently (i.e., there is no combining of expert judgments into one function). The benefits model (e.g., BenMAP) is used to derive the total mortality incidence corresponding to the judgments of each expert and to combine (or pool) the estimates into an aggregate value before taking an average of the mortality incidence. This "pooled" approach has been used by EPA to model other benefit endpoints that have multiple C-R function (due to multiple studies). This method lessens the amount of alteration that is required for Experts B and C (although some adjustments must still be made), which reduces the influence of the "effective" distributions from these experts on the combined mortality function.[36]

Insights from the Expert Elicitation Process

In the absence of calibration measures, one of the ways to assess the ultimate value of the results of this pilot assessment is to evaluate the process. For example, did the pilot study employ a structure, supporting materials, and a process that enabled experts to make judgments that would be likely to be well calibrated? We offer below some conclusions on the strengths as well as the weaknesses of the process.

[36] Expert B's and C's still require some adjustment prior to applying their results in a benefits assessment model. For Expert B, the analyst could use probabilities derived from the triangular distribution characterizing uncertainty about the threshold level to calculate an expected value response at each location based on the observed baseline $PM_{2.5}$ level; the probability of that baseline level exceeding the potential threshold value; and Expert B's elicited C-R coefficients. In order to "fill-in" Expert C's C-R function for intermediate baseline PM2.5 values, an analyst could linearly interpolate between the responses for each pair of points provided by Expert C, e.g. 10 to 15 or 15 to 20. This could be used to produce a set of log-linear C-R functions for Expert C conditional on the baseline $PM_{2.5}$ concentration. Total incidence of mortality for Expert C would be the sum of the conditional estimates over the range of baseline air concentrations.

Strengths

The expert group represented a reasonable range of expertise and opinion. They held divergent views on causality, thresholds, C-R functions, and the importance of confounders, among others. Their discussion of these issues and their approaches to expressing their views quantitatively provide insight into both differences of opinion in the scientific community and how they may be dealt with in the elicitation of subjective judgments.

Despite the small size of the expert panel, no one individual expert dominated the outcome of the combined expert distributions for long or short-term exposures. Our sensitivity analysis found that the eliminating individual experts could cause the combined distribution shift by as much as 30-40 percent in one direction or another but usually by less.

The pilot tests of the protocol played an important role in the development of the final elicitation protocol. They helped clarify ambiguities in early drafts of the questions, identify data that might be useful for the briefing book, and gauge the length of time required for the interviews.

The protocol questions generally touched on most of the major issues the experts felt were relevant to answering the quantitative questions. The protocol was flexible enough to allow experts to raise additional specific concerns as well. For example, although the protocol did not have a specific question about the experts' views on distributed lag models, almost all of the experts raised the issue in discussing the "total" mortality from short-term increases in $PM_{2.5}$.

Another strength of this pilot study was that it allowed for alternative approaches to specifying the underlying C-R function. The design of the quantitative questions was flexible enough to allow the experts to posit alternative models to characterize the $PM_{2.5}$/mortality relationship (log-linear, log-linear-threshold, non-linear threshold). Although we ultimately had to express the combined results within the context of a log linear framework for comparative purposes, it was important for the experts to be able to answer the question in the way they felt most appropriate. We do have concerns about the degree of uncertainty expressed in some of the alternative models given the evidence offered by the experts. Future protocols for any subsequent elicitation exercises on this topic should be designed to more systematically explore an expert's choice of C-R function and the evidence supporting his or her choice.

The experts varied somewhat in their comfort levels with statistical manipulations of data, but all the experts understood probability concepts well. They were able to assign probabilities to C-R coefficient values with relative ease. The use of tools, such as probability wheels, to illustrate probabilistic concepts was unnecessary for this group of experts.

Potential Concerns

In reviewing the pilot study protocol, the SAB-HES raised concerns about the size of the panel. They point to the many factors which must be balanced in the selections of expert panels (Hawkins and Graham, 1988). They state that they are "… concerned about whether the judgments of such a limited group can reasonably be interpreted as representing a fair and balanced view of the current state of knowledge." Although the decision analysis field tends to

use relatively small sample sizes (i.e., typically 5-10 experts), some members of the project team are not comfortable with obtaining a combined distribution from such a small number of experts in the absence of more extensive evaluation of the degree to which the expert panel is likely to be statistically representative of the overall population of relevant experts on the question of interest. The panel size seems reasonable given the pilot-scale nature of this analysis and the expedited schedule for completion of the project, and it is not unusually small when compared with expert elicitation projects for questions of similar complexity. However, the panel of experts used in this elicitation may not have represented some important views or may have over-represented minority opinions, and it offered limited flexibility for incorporating opinions from multiple technical disciplines. A larger sample size may provide better representation of the range of opinions on this topic both across and within different disciplines. Therefore, future elicitation projects should emphasize identification of a broader pool of experts and selection of a panel likely to be representative of the breadth of respected scientific opinion on the issues being addressed in the future elicitation.

The elicitation interviews were long. To allow for a complete review of the introductory materials for the project as well as full responses to the preliminary, quantitative, and follow-up questions, the elicitation required an intense eight-hour day. Both elicitors and the experts experienced some fatigue, particularly toward the end of the day.

Not all experts were available to the team for a full day. Two of the interviews had to be condensed into shorter time frames (5-6) hours, one of which required 1-2 hours of follow-up on a second day. The limited time prompted us to limit discussion of background material and some individual questions.

We observed differences in the knowledge of the scientific literature and/or in the degree of preparation for the interviews. Preparation ranged from little preparation to a full day spent in careful review of the relevant literature, the briefing book, and the protocol. Involvement in the pilot project did not specify a level of preparation although the experts were sent the briefing book and protocol to review in advance of the elicitation. However, the level of preparation did appear to affect experts' familiarity with the breadth of the relevant literature (both those showing positive and negative effects of $PM_{2.5}$) and the details of specific studies. Whether the impact of overall knowledge, preparation or both, the experts varied considerably in how well they could articulate the impact of potential confounders, effect modifiers, or other uncertainties on the magnitude and strength of the statistical findings reported by those studies.

Most of the experts relied in some form or another on the heuristic of "anchoring and adjustment" to develop their uncertainty distributions. In this process, the expert begins his estimates with, or "anchors" on, a particular study or set of studies, then develops confidence intervals to account for various factors that influence his judgment. Some expert judgment research has shown that use of this approach leads to overconfidence, and thus poor calibration, because individuals fail to adjust confidence intervals adequately for what they do not know. The protocol did ask the experts to begin development of their distributions by discussing the maximum and minimum values they believed possible, a technique that can help experts decrease the impact of anchoring and adjustment. However, given most experts' expressed confidence in the epidemiologic studies, their reliance on the data is not surprising.

The structure of the elicitation protocol often made it difficult to know exactly how experts' responses to the largely open ended preliminary questions influenced their later probabilistic estimates of uncertainty in the $PM_{2.5}$-mortality relationship. The influences were most clear when the preliminary questions were more focused, for example those asking about the likelihood of a causal relationship, the shape of the C-R function, and the existence of thresholds. We found it generally difficult to ascertain how experts' general concerns about specific confounders, effect modifiers, exposure errors, $PM_{2.5}$ components or co-pollutants translated into quantitative responses. This difficulty is a common consequence of the "aggregated" versus a structured disaggregated approach to answering the quantitative questions.

Although the preliminary questions that asked for quantitative responses (e.g. causality, degree of overlap between cohort and time-series studies, cessation lag, impact of differential toxicity) offered the potential to establish a clear quantitative influence on the experts' ultimate estimates of the C-R coefficient, difficulties with the interpretation of some of these questions limited our ability to trace these influences. For example, the question about the likelihood of causal relationships implicitly applied to the full range of baseline $PM_{2.5}$ exposure (e.g. 8 to 20 $\mu g/m^3$ for annual average PM); an assumption with which all experts did not agree. Perhaps because these questions were found in the warm-up or follow-up sections of the protocol, their quantitative relationship to the later elicited judgments about the mortality effects of long-term and short-term exposures was not as clearly defined as it should have been. Also, the issue of distributed lag in time-series studies should have been explicitly addressed in the protocol.

Some experts had concerns about the key assumptions underlying the quantitative questions about the long- and short-term exposures. For the effects of long-term exposures, several experts questioned the implicit assumptions about the characteristics of the population over time, including basic demographics as well as their state of health or frailty. Some experts were unclear, at least initially, about what assumptions they would need to make about cessation lags in order to answer the questions. Some would have liked to have more information about the nature of the regulations that might be implemented and how they would affect $PM_{2.5}$, its components, and co-pollutants.

In any elicitation, the quality of an expert's judgment depends in part on his willingness to engage seriously in the process of considering the relevant data and sources of uncertainty. A clear understanding of the importance of the questions being asked as well as the overall analytical context for their use can be important in enlisting an expert's support. In this pilot, two of the five experts questioned whether uncertainty about the C-R function for $PM_{2.5}$ was the most important source of uncertainty to address in a benefits analysis. They cited uncertainty about assumptions for the cessation lag for different types of mortality, the ultimate impact of regulation on emissions and air quality, changes in population demographics and health status over time, and other factors that seemed to be as or more important. Although it is not clear that these concerns specifically affected the quality of the judgments these experts gave, any future elicitations in this area may want to devote more time to addressing such concerns with experts as part of a workshop held prior to conducting the interviews.

Finally, motivational bias among experts is a potential concern in any expert judgment assessment where the political stakes for the outcome are high. The elicitation team made reasonable attempts during each interview to probe experts' responses and obtain rational

justifications for their positions. However, it is not possible to rule out the potential for motivational bias in the expert responses.

5. RECOMMENDATIONS

Based on IEc's experience conducting this pilot-scale expert judgment assessment and our review of the results obtained, we developed following recommendations for improving the design of future expert judgment studies focusing on the PM/mortality relationship:

- Recruitment of experts needs to stress the importance of the full commitment to the time necessary for the elicitation, including preparation and subsequent follow-up. A clear, up-front emphasis by EPA on the importance of these results for the improvement of future benefits analyses could also increase the level of expert engagement in the process.

- The experts need to be convinced that the elicitation protocol is asking the right question. Evidence should be presented as part of the recruitment process, the initial briefing process, or in a workshop that demonstrates to experts that the questions merit their time and careful attention.

- The key assumptions underlying both the preliminary and the final quantitative questions need further clarification. While we believe that the elicitation team was able to respond to concerns experts raised about the questions and underlying assumptions (e.g., the mix of PM species), it would have been preferable to provide the necessary detail in the protocol. Including more information about EPA's benefits analysis methodology in the introductory briefing could help experts better understand the scope and goal of the elicitation.

- Additional thought should be given to how the elicitation protocol could best address experts' desires to specify alternatives to the continuous log-linear C-R function. The advantage of the approach taken in the pilot was that the experts could work with the published $PM_{2.5}$-mortality coefficients even with alternative models. However, a more systematic approach to exploring alternative model specifications with experts in the protocol would facilitate both the collection of information about the C-R function and planning for the application of elicited results in a benefits analysis.

- The briefing book should ideally provide more useful summaries and analyses that are readily accessible during the interviews. The material the experts are asked to assimilate is vast. Although it would require more work, summaries/analyses of the following kind could be helpful:

    - Tabular summaries of the quantitative results of epidemiological results for easy comparison between studies. These should also be

available on a computer to facilitate alternative approaches to aggregating the results.

- Comparison of the data available to address particular quantitative issues like the impact of distributed lag models

- Summaries of studies addressing the effect of particular confounding variables, effect modifiers, or other uncertainties or biases that experts are asked specifically to quantify.

- Summaries/analysis of the data characterizing $PM_{2.5}/PM_{10}$ ratios.

- EPA should consider a structured approach to elicitation in which the key assumptions, such as those developed in the preliminary questions, are more clearly linked to the expert judgment. While it is not clear from the expert judgment literature that experts perform better using disaggregated methods, the process intuitively allows for a more systematic and transparent accounting of the factors that experts believe to be important.

- EPA should consider the use of a workshop or other form of pre-elicitation conference to address up front some of the issues encountered in the pilot assessment, such as:

  - Defining the purpose and context for the specific elicitation;

  - Explaining the role of expert judgment in the process;

  - Clarifying the questions and underlying assumptions;

  - Discussing and critiquing the evidence, both positive and negative, with respect to key issues;

  - Identifying analyses and summaries that may be useful for the briefing book; and

  - Helping in the formulation of a disaggregated structure for obtaining inputs to the quantitative estimates.

  The goal of such a workshop would not be to force a consensus but rather to help experts avoid errors in judgment that might arise from uneven preparation and understanding.

- Finally, EPA should evaluate the feasibility of developing and validating a set of calibration questions to help assess the accuracy and precision of the experts' judgments about the $PM_{2.5}$/mortality relationship. In practice, we recognize that it may be difficult to identify a set of independent questions that will be suitable for all technical disciplines. Without additional research into how to calibrate experts from various domains of expertise within the

context of multidimensional questions like $PM_{2.5}$-mortality relationships, questions will remain about how much confidence to place in the judgments obtained.

6. REFERENCES

Abbey, D.E., B.L. Hwang, R.J. Burchette, T. Vancuren, and P.K. Mills. 1995. "Estimated Long-Term Ambient Concentrations of PM(10) and Development of Respiratory Symptoms in a Nonsmoking Population." *Archives of Environmental Health* 50(2): 139-152.

Abbey, D.E., N. Nishino, W.F. McDonnell, R.J. Burchette, S.F. Knutsen, W.L. Beeson, and J.X. Yang. 1999. "Long-Term Inhalable Particles and Other Air Pollutants Related to Mortality in Nonsmokers." *American Journal of Respiratory and Critical Care Medicine.* 159: 373-382.

Abt Associates Inc. 2000. *Final Heavy-Duty Engine / Diesel Fuel Rule: Air Quality Estimation, Selected Health and Welfare Benefits Methods, and Benefit Analysis Results*. Prepared for the U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC. December.

Abt Associates Inc. 2003. *Proposed Nonroad Landbased Diesel Engine Rule: Air Quality Estimation, Selected Health and Welfare Benefits Methods, and Benefit Analysis Results*. Prepared for U.S. EPA, Office of Air Quality Planning and Standards. Research Triangle Park, NC. April.

Brauer, M., J, Brumm, S. Vedal, and A.J. Petkau. 2002. "Exposure misclassification and threshold concentrations in time analyses of air pollution health effects." *Risk Analysis*. December, 22(6): 1183-1193.

Burnett R.T., J. Brook, T. Dann, C. Delocla, O. Philips, S. Cakmak, R. Vincent, M.S. Goldberg, and D. Krewski. 2000. "Association between particulate- and gas-phase components of urban air pollution and daily mortality in eight Canadian cities." *Inhal Toxicol.* 12(4):15-39.

Burnett, R.T. and M.S. Goldberg. 2003. "Size-fractionated particulate mass and daily mortality in eight Canadian cities." In: Revised analyses of time-series studies of air pollution and health. Special report. Boston, MA: Health Effects Institute; pp. 85-90.

Clemen, R.T. 1996. *Making Hard Decisions: An Introduction to Decision Analysis*. Second Edition. (Belmont: Duxbury Press).

Clemen, R.T. and R.L. Winkler. 1999. "Combining Probability Distributions From Experts in Risk Analysis," *Risk Analysis*. 19: 187-203.

Cooke, R.M. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. (New York: Oxford University Press).

Dockery, D.W., C.A. Pope, X.P. Xu, J.D. Spengler, J.H. Ware, M.E. Fay, B.G. Ferris and F.E. Speizer. 1993. "An association between air pollution and mortality in six U.S. cities." *New England Journal of Medicine* 329(24): 1753-1759.

De Leon, S.F., G.D. Thurston, and K. Ito. 2003. "Contribution of respiratory disease to nonrespiratory mortality associations with air pollution." *American Journal of Respiratory and Critical Care Medicine*. 167(8): 1117-1123.

Devlin, R.B.(a); A.J. Ghio, H. Kehrl, G. Sanders, W. Cascio. 2003. "Elderly humans exposed to concentrated air pollution particles have decreased heart rate variability." *European Respiratory Journal*. May, 21(40): 76-80.

Dominici, F., A. McSermott, S. Zeger, and J.M. Samet. 2002. "On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health." *American Journal of Epidemiology*. 156(9): 193-203.

Dominici, F., M. Daniels, A. McDermott, S.L. Zeger, and J.M. Samet. 2003a. "Shape of the exposure-response relation and mortality displacement in the NMMAPS database." In: Revised analyses of time-series studies of air pollution and health. Special report. Boston, MA: Health Effects Institute; pp. 91-96.

Dominici, F., A. McDermott, S.L. Zeger, J.M. Samet. 2003b. "Airborne particulate matter and mortality: timescale effects in four US cities." *American Journal of Epidemiology*. June 15, 157(12): 1055-65.

Dominici, F., A. McDermott, S.L. Zeger, J.M. Scott, and M. Jonathan. 2003c. "National maps of the effects of particulate matter on mortality: exploring geographical variation." *Environmental Health Perspectives*. January, 111 (1): 39-44.

Dye, J.A., J.R. Lehmann, J.K. McGee, D.W. Winsett, A.D. Ledbetter, J.I. Everitt, A.J. Ghio, and D.L. Costa. 2001. Acute pulmonary toxicity of particulate matter filter extracts in rats: coherence with epidemiologic studies in Utah Valley residents. Environ Health Perspect. Jun;109 Suppl 3:395-403

Evans, J.S., G.M Gray, R.L. Sielken, Jr., A.E. Smith, C. Valdez-Flores, and J.D. Graham. 1994. "Use of Probabilistic Expert Judgment in Uncertainty Analysis of Carcinogenic Potency," *Regulatory Toxicology and Pharmacology*. 20:25-36.

Gold, D.R., A. Litonjua, J. Schwartz, E. Lovett, A. Larson, B. Nearing, G. Allen, M. Verrier, R. Cherry, and R. Verrier. 2000. Ambient pollution and heart rate variability. *Circulation* 101(11):1267-73.

Hennekens, C.H. and J.E. Buring. 1987. *Epidemiology in Medicine*. (Boston: Little, Brown, and Company).

Hawkins, N.C. and J.D. Graham. 1990. "Expert Scientific Judgment and Cancer Risk Assessment: A Pilot Study of Pharmacokinetic Data," *Risk Analysis*. 8: 615-625.

Hoek, G., B. Brunekreef, A. Verhoeff, J. van Wijnen, and P. Fischer. 2000. "Daily mortality and air pollution in The Netherlands," *J Air Waste Manag Assoc.* 50(8):1380-9.

Ibald-Mulli, A., H.E. Wichmann, W. Kreyling, and A. Peters. 2002. "Epidemiological evidence on health effects of ultrafine particles." *Journal of Aerosol Medicine*. Summer, 15 (2): 189-201.

Krewski, D., R.T. Burnett, M.S. Goldberg, K. Hoover, J. Siemiatycki, M. Jerrett, M. Abrahamowicz, and W.H. White. 2000. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. Special Report to the Health Effects Institute, Cambridge MA, July.

Kunzli, N., S. Medina, R. Kaiser, P. Quenel, F. Horak, Jr., and M. Studnicka. 2001. Assessment of deaths attributable to air pollution: should we use risk estimates based on time series or on cohort studies? Am J Epidemiol 153(11):1050-5

Laden, F., L.M. Neas, D.W. Dockery, and J. Schwartz. 2000. "Association of fine particulate matter from different sources with daily mortality in six U.S. cities." *Environ Health Perspect.* 108(10):941-7.

Liao, D., J. Cai, W.D. Rosamond, R.W. Barnes, R.G. Hutchinson, E.A. Whitsel, P. Rautaharju, and G. Heiss. 1997. Cardiac autonomic function and incident coronary heart disease: a population-based case-cohort study. The ARIC Study. Atherosclerosis Risk in Communities Study. American Journal of Epidemiology, 145(8):696-706.

Liao, D., J. Creason, C. Shy, R. Williams, R. Watts, and R. Zweidinger. 1999. Daily variation of particulate air pollution and poor cardiac autonomic control in the elderly. Environ Health Perspect 107:521-5

Lipfert, F.W., H.M. Perry, Jr., J.P. Miller, J.D. Baty, R. Wyzga, and S.E. Carmody. 2000. The Washington University - EPRI Veterans' Cohort Mortality Study: Preliminary Results. *Inhalation Toxicology*. 12:41-73.

McDonnell, W.F., N. Nishino-Ishikawa, F.F. Petersen, L.H. Chen, and D.E. Abbey. 2000. Relationships of mortality with the fine and coarse fractions of long-term ambient PM10 concentrations in nonsmokers. *J Expo Anal Environ Epidemiol.* 10(5):427-36.

Moolgavkar, S. and H.F. Hutchinson. 2003. "Air pollution and daily mortality in two U.S. counties: season-specific analyses and exposure-response relationships." *Inhalation toxicology*. August, 15(9): 877-907.

Morgan, M.G. and M. Henrion. 1990. *Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis*. (New York: Cambridge University Press).

Morgan, M.G., S.C. Morris, M. Henrion, D.A. Amaral, and W.R. Rish. 1984. "Technical Uncertainty in Quantitative Policy Analysis - A Sulfur Air Pollution Example," *Risk Analysis*. 4: 201-216.

National Research Council. 1998. Research Priorities for Airborne Particulate Matter: Immediate Priorities and a Long-Range Research Portfolio. National Academy Press. Washington, DC.

National Research Council (NRC). 2002. Estimating the Public Health Benefits of Proposed Air Pollution Regulations. The National Academies Press: Washington, D.C.

Peters, A., A. Doring, H.E. Wichmann, and W. Koenig. 1997. Increased plasma viscosity during an air pollution episode: a link to mortality? Lancet 349(9065):1582-7.

Peters, A., E. Liu, R.L. Verrier, J. Schwartz J, D.R. Gold, and M. Mittleman M *et al*. 2000a. Air pollution and incidence of cardiac arrhythmia. Epidemiology 11(1):11-7.

Peters, A., S. Perz, A. Doring, J. Stieber, W. Koenig, and H.E. Wichmann. 2000b. Activation of the autonomic nervous system and blood coagulation in association with an air pollution episode. Inhal Toxicol 12(2):51-61.

Peters, A., D.W. Dockery, J.E. Muller, and M.A. Mittleman. 2001a. Increased particulate air pollution and the triggering of myocardial infarction. Circulation 103(23):2810-5.

Peters, A., M. Frohlich, A. Doring, T. Immervoll, H.E. Wichmann, and W.L. Hutchinson *et al*. 2001b. Particulate air pollution is associated with an acute phase response in men; results from the MONICA-Augsburg Study. Eur Heart J 22(14):1198-204.

Pope, C.A., III, R.T. Burnett, M.J. Thun, E.E. Calle, D. Krewski, K. Ito, and G.D. Thurston. 2002. "Lung Cancer Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution." *Journal of the American Medical Association.* 287(9):1132-41.

Pope, C.A., III, M.J. Thun, M.M. Namboodiri, D.W. Dockery, J.S. Evans, F.E. Speizer, and C.W. Heath, Jr. 1995. "Particulate Air Pollution as a Predictor of Mortality in a Prospective Study of U.S. Adults." *American Journal of Respiratory Critical Care Medicine* 151: 669-674.

Pope CA III, R.L. Verrier, E.G. Lovett, A.C. Larson, M.E. Raizenne, R.E. Kanner *et al*. (1999). Heart rate variability associated with particulate air pollution. Am Heart J 138(5 Pt 1):890-9.

Ramsay, T.O., R.T. Burnett, and D. Krewski. 2003. "The effect of concurvity in generalized additive models linking mortality to ambient particulate matter." *Epidemiology*. January, 14(1): 18-23.

Samet, J.M., S.L. Zeger, F. Dominici, F. Curriero, I. Coursac, D.W. Dockery, J. Schwartz, and A. Zanobetti. 2000. The National Morbidity, Mortality and Air Pollution Study: Part II: Morbidity, Mortality and Air Pollution in the United States. Research Report No. 94, Part II. Health Effects Institute, Cambridge MA, June 2000.

Samoli, E., G. Touloumi, A. Zanobetti, A. Le Tertre, C. Schindler, R. Atkinson, J. Vonk, G. Rossi, M. Saez, D. Rabczenko, J. Schwartz, and K. Katsouyanni. 2003. "Investigating the dose-response relation between air pollution and total mortality in the APHEA-2 multicity project." *Occup Environ Med.* 60(12):977-82.

Schlesinger R.B., F. Cassee. 2003. “Atmospheric secondary inorganic particulate matter: The toxicological perspective as a basis for health effects risk assessment.” *Inhalation Toxicology*. March 1, 15(3): 197-235.

Schwartz, J. 2000a. "Assessing Confounding, Effect Modification, and Thresholds in the Association Between Ambient Particles and Daily Deaths" *Environ. Health Perspectives* 108:563-568.

Schwartz, J. 2000b. "The Distributed Lag Between Air Pollution and Daily Deaths" *Epidemiology* 11:320-326.

Schwartz, J. 2000c. Harvesting and Long-term Exposure Effects in the Relation between Air Pollution and Mortality. American Journal of Epidemiology 151: 440-448.

Schwartz, J. 2003. “Daily deaths associated with air pollution in six US cities and short-term mortality displacement in Boston.” In: Revised analyses of time-series studies of air pollution and health. Special report. Boston, MA: Health Effects Institute; pp. 219-226.

Schwartz, J., D.W. Dockery and L.M. Neas. 1996. "Is Daily Mortality Associated Specifically With Fine Particles" *Journal of the Air & Waste Management Association*. 46: 927-939.

Schwartz, J., F. Laden, and A. Zanobetti. 2002. “The concentration-response relation between PMSUB2.5 and daily deaths.” *Environmental Health Perspectives*. October 1, 110(10): 1025-1029.

Seaton, A., A. Soutar, V. Crawford, R. Elton, S. McNerlan, and J. Cherrie *et al*. (1999). Particulate air pollution and the blood. Thorax 54(11):1027-32.

Stieb, D.M., S. Judek, and R.T. Burnett. 2002. “Meta-analysis of time-series studies of air pollution and mortality: Effects of gases and particles and the influence of cause of death, age, and season.” *Journal of the Air and Waste Management Association.* 52(4): 470-484.

US Environmental Protection Agency, 1997. *The Benefits and Costs of the Clean Air Act, 1970 to 1990*. Prepared for US Congress by US EPA, Office of Air and Radiation/Office of Policy Analysis and Review, Washington, DC

US Environmental Protection Agency, 1999. *Benefits and Costs of the Clean Air Act: 1990-2010*; *EPA Report to Congress*. US EPA, Office of Air and Radiation and Office of Policy. Washington, DC. Document No. EPA-410-R-99-001. November.

US Environmental Protection Agency, 2000. *Regulatory Impact Analysis for the Heavy-Duty Standards/Diesel Fuel Rulemaking*. Prepared by: Innovative Strategies and Economics Group, Office of Air Quality Planning and Standards, Research Triangle Park, NC December.

US Environmental Protection Agency, 2002. *Latest Findings on National Air Quality: 2001 Status and Trends*. Prepared by: Emissions, Monitoring, and Analysis Division, Office of Air Quality Planning and Standards, Research Triangle Park, NC September.

US Environmental Protection Agency, 2003a. Fourth External Review Draft of Air Quality Criteria for Particulate Matter: Volumes I and II. June.

US Environmental Protection Agency, 2003b. Preliminary Nonroad Landbased Diesel Engine Rule: Air Quality Estimation, Selected Health and Welfare Benefits Methods, and Benefit Analysis Results. Office of Air Quality Planning and Standards. April.

US Environmental Protection Agency, 2003c. Review of the National Ambient Air Quality Standards for Particulate Matter, OAQPS Staff Paper – First Draft.

Utell, M.J., M.W. Frampton M.W., W. Zareba, R.B. Devlin, and W.E. Cascio. 2002. "Cardiovascular effects associated with air pollution: Potential mechanisms and methods of testing." *Inhalation Toxicology*. December 1, 14(12): 1231-1247.

Vedal, S., M. Brauer, R. White. J. Petkau. 2003. "Air pollution and daily mortality in a city with low levels of pollution." *Environmental Health Perspectives*. January, 111(1): 45-51.

Walker, K.D., P. Catalano, J.K. Hammitt, and J.S. Evans. 2003. "Use of expert judgment in exposure assessment: Part 2. Calibration of expert judgments about personal exposures to benzene." *J Expo Anal Environ Epidemiol.* 13(1):1-16.

Walker, K.D., J.S. Evans, D. MacIntosh. 2001. "Use of expert judgment in exposure assessment. Part 1. Characterization of personal exposure to benzene."*J Expo Anal Environ Epidemiol.* 11(4):308-22.

Whitfield, R.G. and T.S. Wallsten. 1984. *Estimating the Risks of Lead-Induced Hemoglobin Decrements Under Conditions of Uncertainty: A Methodology, Pilot Judgments, and Example Calculations*. Argonne National Laboratory Report ANL/EESTM-276. Argonne, IL.

Whitfield, R.G., T.S. Wallsten, R. L. Winkler, H.M. Richmond, and S.R. Hayes. 1991. *Assessing the Risks of Chronic Lung Injury Attributable to Long-Term Ozone Exposure*. Argonne National Laboratory Report ANL/EAIS-2. NTIS/DE91016814. Argonne, IL. July.

Wilson, A. 2003. "Elicitation of Expert Judgment on health effects of fine particulates," Unpublished Manuscript. Harvard School of Public Health.

Winkler, R. L., T.S. Wallsten, R.G Whitfield, H.M. Richmond, S.R. Hayes, and A.S. Rosenbaum. 1995. "An assessment of the risk of chronic lung injury attributable to long-term ozone exposure," *Operations Research*. 43(1): 19-28.

Wolff, S.K., N.C. Hawkins, S.M. Kennedy, and J.D. Graham. 1990. "Selecting Experimental Data for Use in Quantitative Risk Assessment: An Expert Judgment Approach," *Toxicology and Industrial Health*. 6:275-291.

Zanobetti A., J. Schwartz, E. Samoli, A. Gryparis, G. Touloumi, J. Peacock. H. R. Anderson, A. Le Tertre, J. Bobros, M. Celko, A. Goren, B. Forsberg, P. Michelozzi, D. Rabczenko,

S.P. Hoyos, H.E. Wichmann, and K. Katsouyanni. 2003. "The temporal pattern of respiratory and heart disease mortality in response to air pollution," *Environmental Health Perspectives*. 111(9): 1188-93.