Uncertainty in Mortality Response to Airborne Fine Particulate Matter: Combining European Air Pollution Experts

Jouni T. Tuomisto[1], Andrew Wilson[2], John S. Evans[2], and Marko Tainio[1]

[1]Centre for Environmental Health Risk Analysis, National Public Health Institute, Kuopio, Finland; [2]Department of Environmental Health, Harvard School of Public Health, Boston, USA;

Abstract

The authors have performed a structured expert judgment study of the population mortality effects of fine particulate matter ($PM_{2.5}$) air pollution. The opinions of six European air pollution experts were elicited. The ability of each expert to probabilistically characterize uncertainty was evaluated using 12 calibration questions -- relevant variables whose true values were unknown at the time of elicitation, but available at the time of analysis. The elicited opinions exhibited both uncertainty and disagreement. It emerged that there were significant differences in expert performance. Two combinations of the experts' judgments were computed and evaluated -- one in which each expert's views received equal weight; the other in which the expert's judgments were weighted by their performance on the calibration variables. When the performance of these combinations was evaluated the equal-weight decision-maker exhibited acceptable performance, but was nonetheless inferior to the performance-based decision-maker.

In general, the experts agreed with published studies for the best estimate of all-cause mortality from $PM_{2.5}$; however, as would be expected, they gave confidence intervals that were several times broader than the statistical confidence intervals taken directly from the most frequently cited published studies. The experts were rather comfortable with applying epidemiological results from one geographic region to another. However, there was more uncertainty and disagreement about issues of timing of the effect and about the relative toxicity of different constituents of $PM_{2.5}$. Even so, the experts were in fairly good agreement that an appreciable fraction of the long-term health effects occurs within a few months from the exposure and that combustion-derived particles are more toxic than $PM_{2.5}$ on average, while secondary sulphates, nitrates and/or crustal materials may be less toxic. These assessments bring very valuable and relevant information to air pollution risk assessment.

Introduction

Airborne fine particulate matter ($PM_{2.5}$) is one of the major environmental health problems in modern western societies. A recent EU-funded project Clean Air for Europe (CAFE) estimated that $PM_{2.5}$ causes 350,000 premature deaths per year and that the monetary costs are on the order of 190,000 to 700,000 million € per year in the EU alone (Watkiss, Steve et al. 2005). Therefore, there is great pressure for protecting public health by reducing $PM_{2.5}$ emissions.

However, several major uncertainties hinder effective policy-making. These relate to the true potency of $PM_{2.5}$ to cause health problems in the complex mixture of air pollution; the differential potencies of particles of different chemical composition or from different source

classes; the interpretation of the epidemiological literature where long-term cohort studies show higher effects than short-term time-series studies; the mechanistic understanding of the biological processes; and the relationship between the exposure time window and the effect time window. There is a series of reports related to research priorities in this field (National Research Council 2004).

There is very active epidemiological, toxicological, and exposure research on $PM_{2.5}$, and our understanding of this pollutant is expected to greatly improve within the next decade. However, policies that directly or indirectly address $PM_{2.5}$ emissions or exposures are being developed and implemented continuously (e.g., European Commission 2005). Therefore, it is of utmost importance that the best current understanding is available to decision-makers. Elicitation of expert judgment is a method to obtain data about the current understanding and interpretation of the most recent research. When carefully designed, the expert judgment procedure can be viewed as measurement instrument that produces data amenable to the scientific method. Cooke has argued elsewhere that quantifying uncertainty via expert judgment should therefore be undertaken as a scientific endeavour which 1) is transparent to peer review; 2) does not pre-judge experts; 3) discourages assessments at variance to the experts' true beliefs; and 4) enables empirical quality control of final results (Cooke 1991).

The Harvard Kuwait public health project performed a comprehensive analysis of the health impact of the 1991 Kuwait oil fires, which were set ablaze by Iraq just before the first Gulf War. As a part of this exercise, an expert elicitation regarding the health effects of the smoke from the Kuwaiti oil fires was conducted. Six European experts participated in this exercise – responding to an extensive series of quantitative questions and engaging in qualitative discussions about the mortality response to air pollution by fine particulate matter ($PM_{2.5}$). The questions concerned the effects of $PM_{2.5}$ in general and also asked about effects in specific locations, among particular populations, and about both the time lag between exposure and effect of the relative toxicity of particles from various sources. The analysis stressed the quantification of uncertainty in the health impact predictions in conformity with the above desiderata. The elicitation was conducted in the Spring/Summer of 2004.

We present the results of the elicitation of European experts. The purpose of this paper is to quantify the uncertainty in $PM_{2.5}$ mortality response based on assessments of these experts, and to compare different strategies for combining judgment. Other papers describe in detail the experts' reasoning for their answers (Wilson, Tuomisto et al., 2006a) and issues related specifically to the mortality in Kuwait (Wilson, Tuomisto et al., 2006b).

Methods

Elicitation protocol

The elicitation protocol is described in detail in previous articles related to this work (Wilson, Tuomisto et al. 2006a, Wilson, Tuomisto et al. 2006b). Therefore, only a brief summary is given here.

Experts were selected using peer nomination. The top-ranking researchers on air pollution health research (based on article count ranking) were asked to nominate European experts on air pollution epidemiology and/or prominent researchers capable of carefully interpreting the relevant scientific literature. Ten experts with the highest numbers of nominations (but only one per institute) were asked to participate, and six accepted. The participating experts and

their affiliations are listed below. The order in which they are presented does not correspond to the numbering in the subsequent analysis of the data.

- Dr. Bert Brunekreef (PhD) is Professor of Environmental Epidemiology and Head of the Division of Environmental and Occupational Health at the Institute for Risk Assessment Sciences (IRAS) at the University of Utrecht, Netherlands.

- Dr. Annette Peters (PhD) is Assistant Professor and Head of the Epidemiology of Air Pollution Health Effects Research Unit of the Institute of Epidemiology at the GSF National Research Center for Environment and Health, Neuhenberg, Germany.

- Dr. Nino Künzli (MD, PhD), formerly Assistant Professor at the Institute for Social and Preventive Medicine at the University of Basel (Switzerland), is Associate Professor at the University of Southern California Keck School of Medicine, United States.

- Dr. H. Ross Anderson (MD, PhD) is Professor and Head of the Department of Public Health Sciences at St. George's Hospital Medical School at the University of London, England.

- Dr. Ken Donaldson (PhD) is Professor of Respiratory Toxicology on the Faculty of Medicine at the University of Edinburgh, Scotland.

- Dr. Juha Pekkanen (MD, PhD) is Head of the Unit of Environmental Epidemiology in the Division of Environmental Health at the National Public Health Institute of Finland.

The experts were provided with the questions to be asked, additional material related to air pollution concentrations and population characteristics in the relevant areas, and a "briefing book" – i.e., a compact disk with approximately 100 scientific articles related to the issue. All this material was available to the experts before and during the interviews.

The elicitation was preceded by training session in which all experts were gathered together, discussed the issues and underwent a practice elicitation exercise. Methods of scoring and combination were explained in detail. Two experts could not participate in the main workshop, and a similar event was organised for them before their interviews.

All experts were interviewed individually by a normative expert (R. Cooke) and a substantive expert (either A.Wilson or J. Tuomisto). Experts quantified their uncertainty by giving 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentiles of their distributions. To gain insight in the experts' reasoning, experts were queried as to the reasons underlying their assessments, and these were written up by the project team. In addition, each expert wrote a description of the reasoning underlying their estimate of number of premature deaths due to the Kuwait oil fires. The experts were paid for their participation at current consulting rates.

Questions of interest

The protocol included eight questions about mortality after short- or long-term exposures to PM, two questions related to differences in the toxicity of different constituents of PM, and two questions related directly to health effects of the Kuwait oil fires. Some questions pertained to effects in the US, other to effects in Europe, and some to effects in the Mexico

City Metropolitan Area (MCMA). In addition, there were 12 calibration questions (aka seed variables) used for measuring the experts' performance. The true values of these seed variables were unknown to both the experts and the analysts at the time of the elicitation, but were recovered after the elicitation. The questions of interest are listed below.

Q1. What is your estimate of the true, but unknown, percent change in the annual, non-accidental mortality rate in the adult US population resulting from a permanent 1 $\mu g/m^3$ reduction in long-term annual average $PM_{2.5}$ (from a population-weighted baseline concentration of 18 $\mu g/m^3$) throughout the US?

Q2. What is your estimate of the true, but unknown, percent change in the annual, non-accidental mortality rate in the adult European population resulting from a permanent 1 $\mu g/m^3$ reduction in long-term annual average $PM_{2.5}$ (from a population-weighted baseline concentration of 20 $\mu g/m^3$) throughout the EU?

Q3. What is your estimate of the true, but unknown, percent change in non-accidental mortality in the total US population over the one week following a 10 $\mu g/m^3$ increase in $PM_{2.5}$ levels on a single day throughout the US?

Q4. What is your estimate of the true, but unknown, percent change in non-accidental mortality in the total MCMA population over the one week following a 10 $\mu g/m^3$ increase in $PM_{2.5}$ levels on a single day throughout the MCMA?

Q5. What is your estimate of the true, but unknown, percent change in non-accidental mortality in the total European population over the one week following a 10 $\mu g/m^3$ increase in $PM_{2.5}$ levels on a single day throughout the EU?

Q6. What is your estimate of the true, but unknown, percent change in non-accidental mortality in the total US population over the three months following a 10 $\mu g/m^3$ increase in $PM_{2.5}$ levels on a single day throughout the US?

Q7. What fraction of the 50^{th} percentile answer you gave in Question 1 (US, long-term) was due to effects which would express themselves within 1 week of the change in exposure?

Q8. What fraction of the 50^{th} percentile answer you gave in Question 1 (US, long-term) was due to effects which would express themselves within three months of the change in exposure?

Q9. What is your estimate of the true, but unknown, percent change in the total annual, non-accidental mortality rate in the adult US population resulting from a permanent 1 $\mu g/m^3$ reduction of [most toxic] in long-term annual average $PM_{2.5}$ (from a population-weighted baseline concentration of 18 $\mu g/m^3$) throughout the US?

Q10. What is your estimate of the true, but unknown, percent change in the total annual, non-accidental mortality rate in the adult US population resulting from a permanent 1 $\mu g/m^3$ reduction of [least toxic] in long-term annual average $PM_{2.5}$ (from a population-weighted baseline concentration of 18 $\mu g/m^3$) throughout the US?

Q11. What is your estimate of the true, but unknown, percent change in the non-accidental mortality rate in the Kuwaiti national population alive at the beginning of the 1991 oil fires resulting from the nine-month excursion in $PM_{2.5}$ concentration patterns due to the oil fires?

Q12. What is your estimate of the true, but unknown, number of premature deaths in the Kuwaiti national population alive at the beginning of the 1991 oil fires resulting from the nine-month excursion in $PM_{2.5}$ concentration patterns due to the oil fires?

Combining expert judgments

Cooke has argued elsewhere that expert judgment may be seen as a new type of scientific instrument. As with any new scientific instrument, we must first calibrate it on objects whose features are already known, before applying it to the unknown. The Aristotelian's response to Galilei's telescope was to the effect: "Even if the contraption works when pointed to earthly objects, that doesn't mean it also works when pointed to the heavens." The fact that such arguments cannot be defeated does not discharge us from the duty of doing the best we can.

However there are legitimate questions about whether, when, and if so how to combine expert judgments. For example, Morgan and Henrion (1990, pp. 164) give the following advice:

> "Having concluded that there are legitimate differences of opinion among experts, the analyst should then examine the extent to which this range of opinions has important consequences for results. If the range of opinions has no significant consequences for the model outcome, combining expert opinions to obtain some representative average view is clearly justified…If, on the other hand, the range of experts' opinions has major consequences for the model outcome, …, then in most cases the experts' opinions should not be combined just to produce some "average" result."

While we agree with Morgan and Henrion that the full range of expert opinions should be presented to the decision maker, we have found that after having seen the individual opinions decision makers often seek to synthesize the judgments. The approach that we have used for combining expert judgement is based on weighted averaging of the individual assessments. In the equal weight combination, all experts have equal weights. Cooke's "classical model" for combining expert judgements constructs a weighted combination of expert probability assessments. These weights are based on two key performance measures, calibration and informativeness, which are assessed on 'seed variables' whose true values are known post hoc. These values are not known to the experts at the time of assessment; in this study, the values were also not known to the elicitors at the time of the assessment, as the data necessary to compute these values was retrieved months later. The calibration questions used in this analysis are given later.

Calibration corresponds to evaluating the likelihood of the statistical hypothesis that an expert's probabilities are correct. In the language of statistics, this is the probability (or "p-value") of incorrectly rejecting this hypothesis, given the realizations. Thus, low values for the calibration score (near zero) indicate low support for the hypothesis that the expert's probability statements are accurate; high values (near one) indicate high support. Information, or informativeness, measures the degree to which the experts' distributions are concentrated.

The calibration score is a "fast" function; i.e., differences of several orders of magnitude are observed among relatively small groups of experts with, say 12 calibration variables. On the other hand, information is a "slow" function; differences in experts' information scores are typically within a factor of three. In combining expert judgments, these calibration and information scores are multiplied and normalized; hence in combining experts, the calibration

score dominates over information score . Information serves to modulate between more or less equally well-calibrated experts.

The performance-based structured expert judgement methodology has been applied in many risk and reliability studies. The model for combining expert judgments is called "classical" because of its relation to classical hypothesis testing. More detailed definitions of calibration and informativeness are given below.

The performance weights are based on the theory of proper scoring rules. Scoring is an assignment of a numerical value to probability assessments on the basis of observation. A scoring rule is called strictly proper if a subject receives his best expected score if and only if his stated assessment corresponds to his true opinion. See, for example, (Cooke 1991 chapter 9). Under a strictly proper scoring rule an expert can maximize his/her long run expected score by, and only by, stating his/her true beliefs. There is thus no advantage in trying to tailor one's assessments so as to achieve maximal score – e.g., by reporting narrower confidence intervals than they actually believe are appropriate.

The classical model computes "performance-based" weighted combinations, and uses the performance measures to assess the quality of other combinations. In particular, the performance of the equal weight combination is assessed.

Calibration

We have asked for each expert's uncertainty over a number of calibration variables; these variables are chosen to resemble the quantities of interest, and to demonstrate the experts' ability as probability assessors. An expert states *n* fixed quantiles of his/her subjective distribution for each of several uncertain quantities. There are *n+1* 'inter-quantile intervals' into which the actual values may fall. Let

$$p = (p_1,\ldots,p_{n+1}) \qquad (1)$$

denote the theoretical probability vector associated with these intervals. Thus, if the expert assesses the 5%, 25%, 50%, 75% and 95% quantiles for the uncertain quantities, then n = 5 and p = (5%, 20%, 25%, 25%, 20%, 5%). The expert believes there is 5% probability that the realization falls between his/her 0% and 5% quantiles, a 20% probability that the realization falls between his/her 5% and 25% quantiles, and so on.

Suppose we have such quantile assessments for *m* seed variables. Let

$$s = (s_1,\ldots s_{m+1}) \qquad (2)$$

denote the empirical probability vector of relative frequencies with which the realizations fall in the interquantile intervals. Thus

s_1 = (# realizations less than or equal to the 5% quantile) / m

s_2 = (# realizations strictly above the 5% quantile and less than or equal to the 25% quantile) / m

s_3 = (# realizations strictly above the 25% quantile and less than or equal to the 50% quantile) / m

And, so on.

If the expert is well calibrated, he/she should give intervals such that – in a statistical sense – 5% of the realizations of the calibration variables fall into the corresponding 0% to 5% intervals, 20% fall into the 5% to 25% intervals, etc.

We may write:

$$2mI(s,p) = 2m\sum_{i=1}^{m+1} s_i \ln(\frac{s_i}{p_i}) \quad (3)$$

where I(s,p) is the Shannon relative information of *s* with respect to *p*. For all s,p with $p_i > 0$, i = 1,…, m +1, we have I(s,p) ≥ 0 and I(s,p) = 0 if and only if s=p (see Kullback 1959).

Under the hypothesis that the uncertain quantities may be viewed as independent samples from the probability vector p, 2m*I(s,p)* is asymptotically Chi-square distributed with m degrees of freedom: $P(2mI(s;p) \le x) \approx \chi^2_m(x)$ where χ^2_m is the cumulative distribution function for a Chi-square variable with *m* degrees of freedom. Then

$$CAL = 1 - \chi^2_m(2mI(s,p)) \quad (4)$$

is the upper tail probability, and is asymptotically equal to the probability of seeing a disagreement no larger than I(s,p) on *n* realizations, under the hypothesis that the realizations are drawn independently from *p*.

CAL is a measure of the expert's calibration. Low values (near zero) correspond to poor calibration. This arises when the difference between s and p cannot be plausibly explained as the result of mere statistical fluctuation. For example, if m = 10, and we find that 8 of the realizations fall below their respective 5% quantile or above their respective 95% quantile, then we could not plausibly believe that the probability for such events was really 5%. This phenomenon is sometimes called "overconfidence." Similarly, if 8 of the 10 realizations fell below their 50% quantiles, then this would indicate a "median bias." In both cases, the value of CAL would be low. High values of CAL indicate good calibration.

Informativeness

Information is measured as Shannon's relative information with respect to a user-selected background measure. The background measure will be taken as the uniform (or loguniform) measure over a finite "intrinsic range" for each variable. For a given uncertain quantity and a given set of expert assessments, the intrinsic range is defined as the smallest interval containing all the experts' quantiles and the realization, if available, augmented above and below by K%.

The relative information of expert *e* on a given variable is:

$$I(e) = \sum_{i=1}^{n+1} p_i \ln(\frac{p_i}{r_i}) \quad (5)$$

where r_i are the background measures of the corresponding intervals and *n* the number of quantiles assessed. For each expert, an information score for all variables is obtained by summing the information scores for each variable. This corresponds to the information in the

expert's joint distribution relative to the product of the background measures under the assumption that the expert's distributions are independent. Roughly speaking, with the uniform background measure, more informative distributions are obtained by choosing quantiles that are closer together, whereas less informative distributions result when the quantiles are farther apart.

Equal-weight and performance-based decision maker

The probability density function for the equal-weight "decision maker" is constructed by assigning equal weight to each expert's density. If *E* experts have assessed a given set of variables, the weights for each density are 1/E; hence for variable *i* in this set the decision maker's density is given by:

$$f_{eqdm,i} = \left(\frac{1}{E}\right) \sum_{j=1\cdots E} f_{j,i} \qquad (6)$$

where $f_{j,i}$ is the density associated with expert *j's* assessment for variable *i*.

The performance-based "decision maker's" probability density function is computed as a weighted combination of the individual expert's densities, where each expert's weight is based on his/her performance. Two performance-based "decision makers" are supported in the software EXCALIBUR. The "global weight decision maker" is constructed using average information over all calibration variables and, thus, one set of weights for all questions. The "item weight decision maker" is constructed using weights for each question separately, using the experts' information scores for each specific question, rather than the average information score over all questions.

In this study the global and items weights do not differ significantly, and we focus on the former, calling it simply "performance-based decision maker". The performance-based decision maker (Table 4) uses performance-based weights that are defined, per expert, by the product of expert's calibration score and his/her overall information score on calibration variables, and by an optimization procedure.

For expert *j*, the same weight is used for all variables assessed. Hence, for variable *i* the performance-based decision maker's density is:

$$f_{gwdm,i} = \frac{\sum_{j=1\cdots E} w_j f_{j,i}}{\sum_{j=1\cdots E} w_j} \qquad (7)$$

The cut-off value for experts that were given zero weight was based on optimising: the value that gave the highest performance score to the decision-maker was selected. The optimising procedure is the following. For each value of α, define a decision maker DM_α, which is computed as a weighted linear combination of the experts whose calibration score is greater than or equal to α. DM_α is scored with respect to calibration and information. The weight that this DM_α would receive if he were added as a "virtual expert" is called the "virtual weight" of DM_α. The value of α for which the virtual weight of DM_α is the greatest is chosen as the cut-off value for determining which experts to exclude from the combination.

Seed variables

Seed variables fulfil a threefold purpose, namely to enable: 1) the evaluation of each expert's performance as a probability assessor; 2) the performance-based combination of the experts' distributions; and 3) assessment of the relative performance of various possible combinations of the experts' distributions.

To do this, performance on seed variables must be seen as relevant for performance on the variables of interest, at least in the following sense: If one expert gave narrow confidence bands widely missing the true values of the seed variables, while another expert gave similarly narrow confidence bands which frequently included the true values of the seed variables, would these experts be given equal credence regarding the variables of interest? If the answer is affirmative, then the seed variables fall short of their mark. Evidence indicates that performance on 'almanac items' ("How many heretics were burned at Montsegur in 1244?") does not correlate with performance on variables from the experts' field of expertise (Cooke, Mendel et al. 1988). On the other hand, there is some evidence that performance on seed variables from the field of expertise does predict performance on variables of interest (Qing 2002).

The seed questions for this study were based on PM_{10} data and mortality data from London and Athens for the period 1997 – 2001. The London PM_{10} measurement stations used were Hillington, Eltham, Brent, Bloomsbury, and Bexley. The Athens stations used were Agia Paraskevi, Lykovrisi, Marussi, Thrakomakedones, and Zografou.

Half of the seed variables concerned PM concentrations in London and Athens (S1-S6); the other half concerned mortality in these two cities (S7-S12). The experts have somewhat different backgrounds and there is no reason to expect that they should perform identically on these two sets. These seeds were selected because assessing exposure and mortality are arguably equally important in judging uncertainty in dose-response relations.

Questions on PM_{10} reference station exceedances in London and Athens

S1. On how many days in 2001 did the daily average PM_{10} concentration exceed 50 $\mu g/m^3$ at at least one of the above London stations?

S2. On how many days in 2001 did the daily average PM_{10} concentration fall below 30 $\mu g/m^3$ at all of the above London stations?

S3. On how many days in 1997 did the daily average PM_{10} concentration exceed 50 $\mu g/m^3$ at at least one of the above London stations?

S4. On how many days in 1997 did the daily average PM_{10} concentration fall below 30 $\mu g/m^3$ at all of the above London stations?

S5. On how many days in 2001 did the daily average PM_{10} concentration exceed 50 $\mu g/m^3$ at at least one of the above Athens stations?

S6. On how many days in 2001 did the daily average PM_{10} concentration fall below 30 $\mu g/m^3$ at all of the above Athens stations?

Questions on mortality in London

The annual average for the year 2000 of daily PM_{10} concentration, averaged over the 5 stations in London was 18.4 $\mu g/m^3$. The highest weekly concentration averaged over these same 5 London stations was 33.4 $\mu g/m^3$.

S7. What is the ratio of the number of non-accidental deaths in the week (7 days starting from January 1st) of 2000 with the highest average PM_{10} concentration to the weekly average number of non-accidental deaths in 2000?

S8. What is the ratio of the number of cardiovascular deaths (ICD10 Cause I) in the week (7 days starting from January 1st) of 2000 with the highest average PM_{10} concentration to the weekly average number of cardiovascular deaths in 2000.

Questions on mortality in Athens

The annual average for the year 2001 of daily PM_{10} concentration, averaged over the 5 Athens stations was 45.5 $\mu g/m^3$; the highest weekly concentration averaged over the Athens stations was 63.2 $\mu g/m^3$; the lowest weekly concentration averaged over these same 5 Athens stations was 16.5 $\mu g/m^3$.

S9. What is the ratio of the number of non-accidental deaths in the week (7 days starting from January 1st) of 2001 with the highest average PM_{10} concentration to the weekly average number of non-accidental deaths in 2001?

S10. What is the ratio of the number of cardiovascular deaths in the week (7 days starting from January 1st) of 2001 with the highest average PM_{10} concentration to the weekly average number of cardiovascular deaths in 2001?

S11. What is the ratio of the number of non-accidental deaths in the week (7 days starting from January 1st) of 2001 with the lowest average PM_{10} concentration to the weekly average number of non-accidental deaths in 2001?

S12. What is the ratio of the number of cardiovascular deaths in the week (7 days starting from January 1st) of 2001 with the lowest average PM_{10} concentration to the weekly average number of cardiovascular deaths in 2001?

Results

Variables of interest

The results of the variables of interest and the detailed description of the reasoning of each expert have been published elsewhere (Wilson, Tuomisto et al. 2006a; Wilson, Tuomisto et al. 2006b). Therefore, this paper will focus on issues related to the calibration and combination of the expert judgements.

To get a picture of the degree of homogeneity within the expert group, range graphs showing all assessments on each item are useful. Figure 1 shows all expert assessments with the 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles of the estimates. Figure 2 shows all assessments of the calibration (seed) variables. Equal-weight and performance-based decision-makers are shown on these figures as well.

For the percent change in long term mortality following a 1 $\mu g/m^3$ reduction in ambient $PM_{2.5}$, (questions Q1 and Q2 for the US and Europe, respectively), the equal-weight decision maker's quantiles are more spread out than the performance-based decision maker's are, and the medians tend to be higher. Numerical output for these decision makers is given in Tables 4 and 5.

Questions Q3, Q4, and Q5 show similar information for the percent change in mortality over one week following a 10 $\mu g/m^3$ rise in $PM_{2.5}$ on one day in the US, Europe, and Mexico City, respectively. The pattern is similar to that of Q1 and Q2.

Questions Q6 and Q7 concern the fraction of the long-term mortality effect (for the US) that is expressed within 1 week or 3 months, respectively. Overall, the experts disagree more about these questions than with more straightforward questions Q1-Q5. We also see that the time from exposure to effect is shorter when estimated by using performance-based decision-maker in comparison to the equal-weight decision-maker.

For the questions relating to long term effects, the experts' central 90% confidence intervals, generally show considerable overlap. For short term effects we observe disjoint 90% confidence bounds, and this becomes more marked for three-month effects, for relative toxicity questions (Q6 and Q7), and for calibration variables. For the mortality calibration variables (S7 through S12), there was considerable agreement in the median value, but substantial disagreement regarding the uncertainty. Overall, the degree of consistency in this data is comparable to that observed in other studies.

Combination of expert judgements

In this study, experts give their uncertainty assessments on calibration variables in the form of 5%, 25%, 50%, 75% and 95% quantiles. As noted above, two weighting schemes, equal-weight decision maker and performance-based decision maker were used to combine the results.

The published confidence intervals are derived under a number of assumptions, in particular, that the Cox proportional hazard model is true for the data analysed. Further, these confidence intervals reflect only statistical sampling fluctuations. The experts, not-surprisingly, acknowledge that the uncertainty in interpreting epidemiological results involves many sources in addition to sampling uncertainty, and reflect this in their subjective confidence intervals.

Calibration. Table 1 includes nine columns. The first one gives the labels of experts – including the decision-makers based on expert combinations. The second column shows the calibration scores for all experts. The combined opinion of the six experts, i.e., 'equal,' is better calibrated than any individual expert, and expert B's calibration score is better than the other experts. It will be noted that experts B, C, and D had calibration scores corresponding to p-values above 5%. Calibration scores on the order of 0.001, or lower, fail to confer the requisite level of confidence in these expert's results.

Informativeness. Columns 3 and 4 in Table 1 show information scores for all experts. The scores in column 3 consider all variables – both substantive and calibration. In contrast, those in column 4 reflect only the 12 calibration variables. The overall information scores for the calibration and substantive variables are generally similar. The one exception is expert D –

who has a much lower information score on the calibration variables than he/she does on the substantive variables.

These information scores cannot be compared across studies, as the relative information is taken with respect to a background measure that is determined by the experts' responses themselves. They can, however, be compared to other information scores within this study. Note that the patterns of scores for the experts are similar, regardless of whether computed from all variables or restricted to the calibration variables. Expert D is generally the least informative, though this expert did feel confident on specific items. Of course, the individual experts have different knowledge bases and will be more/less informative on various items, reflecting these differences.

Most experts are less informative on the seed variables than on the variables of interest. This reflects the fact that published risk coefficients serve to anchor the mortality assessments, causing the median assessments to cluster – with the result that experts' confidence bands cover a large portion of the intrinsic range. It may be noted in this regard that the long and short-term questions are closely related to issues widely-discussed in the literature, for which published uncertainty bounds are available.

We see that the relative information with respect to the equal weight combination is in the order of one half of the information in the experts' individual assessments.

Weighting scores. Weights for each expert and for the equal- and performance-based "decision makers" are shown in columns 5, 8, and 9. The numbers in column 5 reflect the products of columns 2 and 4; the ratio of highest to lowest combined score is about 3000. If column 5 were normalized and used to form weighted combinations, experts B, C, and D would be influential. This, however, does not impose the strictly proper scoring rule constraint, according to which there must be a calibration cut-off, beneath which experts are unweighted. When the cut-off is imposed, expert C is unweighted.

Robustness. Robustness analysis addresses the question, to what extent the results of the study would be affected by loss of a single expert or calibration variable. Robustness is an issue whenever we optimise performance (robustness analysis for the equal-weight decision maker is omitted.) Tables 2 and 3 show the robustness analysis for the performance-based decision maker; that is, they show how the scores would change if calibration variables (Table 2) or experts (Table 3) were removed from the analysis one at a time. The last two columns show the relative informativeness of each expert with respect to the equal weight decision-maker. This gives a summary measure of how much the experts agree among themselves.

The relevant comparison is between the columns 6 and 7 of Tables 1, 2 and 3. We see that the effect of "perturbing" the model by removing an expert or an item, is small relative to the differences among the experts themselves. In case of removing expert B, there is a significant change, but even this is within the inter-expert differences.

Table 6 shows combined expert judgements of Q1 about the effect of a long-term $PM_{2.5}$ exposure on mortality. These are compared with selected epidemiological cohort studies. The median estimates are in good agreement. This reflects the fact that all experts based their estimates of Q1 on one of these two key studies.

Discussion

This study performed an expert elicitation exercise with European air pollution experts to produce an up-to-date probabilistic characterization of knowledge about the effect of $PM_{2.5}$ exposure on mortality. The questions asked were about the effects of $PM_{2.5}$ in general and also related to specific locations, populations, and timing of exposure. The assessments were in agreement with the current epidemiological literature on long-term and short-term effects. The experts were rather comfortable with applying the U.S. cohort coefficients in Europe, as the estimated medians and confidence bounds were similar for both areas.

However, there was clearly more disagreement among these experts on issues such as the timing of health effects after $PM_{2.5}$ exposure, and potencies of the most or least potent subfractions of $PM_{2.5}$. This is not surprising, as the toxicological literature on mechanisms is only now starting to bring understanding to these unresolved matters. In any case, the experts were in fairly good agreement that a substantial part of the long-term health effects occurs within a few months of the exposure and that combustion-derived particles are clearly more toxic than $PM_{2.5}$ on average, while secondary sulphates, nitrates and/or crustal material may be less toxic.

There is another recent expert elicitation performed in the U.S. The results of that study show fair agreement on the main questions on long-term effect (Q1) (Industrial Economics 2004). A more detailed comparison of these two elicitations will be presented in another paper (Wilson, Tuomisto et al. 2006c).

Many of the seed questions were very difficult. Even good knowledge on typical concentrations of PM_{10} in urban areas is not sufficient for a calibrated answer. The expert had to be well aware of the magnitude of the typical daily variation of the concentrations, and variation from one measurement station to another. In addition, questions about exceedances on ANY or ALL of the five measurement stations require careful consideration of multiplication of several correlated probabilities, a task in which most people have difficulties. Therefore, the calibration scores may be low even if the expert has excellent knowledge on the issue in general.

The questions about mortality ratios also contained several complicated considerations even when the total number of deaths per year was known to the expert. First, the expert had to estimate the variation of weekly mortality during the year, and the variation of PM_{10} in relation to that. Second, although there are typical peaks at certain times of the year of PM_{10} on one hand and mortality on the other hand, it is not obvious that the highest (or lowest) PM_{10} concentration on that particular year occur at the typical time. Third, the effect of the high/low PM_{10} concentration must be incorporated into the expected mortality. And fourth, the random variation that may affect the result depends on the expected number of deaths (the average number was known) and must also be estimated.

Overall, it seemed that overconfidence (seen as small confidence bounds compared with other experts and the realisation), rather than biased median estimates, was the main reason why some experts got lower calibration scores than others. It was noticed during the workshop that some experts were overconfident when assessing the illustrative teaching questions, but after having seen their own scores, clearly improved their performance in the actual elicitation. This emphasises the importance of experience with probability assessments in addition to expertise on the substance.

Choice of combination. In choosing between the equal-weight and performance-based decision makers the following should be borne in mind: 1) both the equal-weight and the performance-based decision makers show acceptable statistical performance; 2) the performance-based decision maker is significantly more informative than the equal-weight decision-maker; 3) the overall scores as reflected in the unnormalised weight, is better for the performance-based decision maker than for the equal-weight decision maker; and 4) the robustness of the performance-based decision maker is quite satisfactory. Thus we conclude that there is a scientific basis for preferring the performance-based decision-maker.

In conclusion, this study showed that European air pollution experts generally agree that current epidemiological studies reflect a causal relationship between PM exposure and mortality, and that weighted combinations of the central estimates from these studies provide unbiased estimates of the mortality impacts of PM exposure. At the same time, they believe that the published confidence intervals underestimate the total uncertainty in estimating these impacts and expand their subjective confidence intervals accordingly.

Combination of judgments across experts – whether equal or performance weighted – provides improved ability to estimate the values of calibration variables and, in theory, should provide improved estimates of quantities of interest – such as the mortality impacts of exposure to fine particulate matter. Performance based combination can be shown to be theoretically preferable to equal weighted combination.

But the most important results of the study may be development of a benchmark characterization of current knowledge and uncertainty about the mortality impacts of PM exposure – a benchmark that can be used to assess future improvements in scientific understanding of this issue. Further, assessments of timing and potency of different constituents bring very valuable information – which is not currently relied upon -- to air pollution risk assessment.

References

Cooke, R., M. Mendel, et al. (1988). "Calibration and Information in Expert Resolution - a Classical Approach." Automatica **24**(1): 87-93.

Cooke, R. M. (1991). Experts in Uncertainty, Oxford University Press.

Dockery, D. W., C. A. Pope, III, et al. (1993). "An association between air pollution and mortality in six U.S. cities." The New England Journal of Medicine **329**(24): 1753-1759.

European Commission (2005). Communication from the Commission to the Council and the European Parliament. Thematic strategy on air pollution. Brussels, European Commission**:** 1-18.

Frijters, M., R. M. Cooke, et al. (1999). Expert Judgment Uncertainty Analysis for Inundation Probability (in Dutch). Utrecht, Ministry of Water Management, Bouwdienst, Rijkswaterstaat.

Industrial Economics (2004). An expert judgment assessment of the concentration-response relationship between $PM_{2.5}$ exposure and mortality. Cambridge, MA, Industrial Economics, Inc**:** 1-94.

Krewski, D., R. T. Burnett, et al. (2000). Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. Cambridge, MA, Health Effects Institute.

Kullback, A. S. (1959). Information theory and statistics. New York, John Wiley and Sons, Inc.

Laden, F., J. Schwartz, et al. (2006). "Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities Study." American Journal of Respiratory and Critical Care Medicine **in press**.

National Research Council (2004). Research priorities for airborne particulate matter : IV continuing research progress. Washington, DC, National Academies Press.

Pope, C. A., III, R. T. Burnett, et al. (2002). "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution." JAMA: The Journal Of The American Medical Association **287**(9): 1132-1141.

Pope, C. A., M. J. Thun, et al. (1995). "Particulate Air-Pollution as a Predictor of Mortality in a Prospective-Study of US Adults." American Journal of Respiratory and Critical Care Medicine **151**(3): 669-674.

Qing, X. (2002). Risk analysis for real estate investment. Dept of Architecture. Delft, TU Delft.

Watkiss, P., P. Steve, et al. (2005). Baseline Scenarios for Service Contract for carrying out cost-benefit analysis of air quality related issues, in particular in the clean air for Europe (CAFE) programme, AEA Technology Environment**:** 1-122.

Wilson, A., J. Tuomisto, et al. (2006a). "Quantifying uncertainty in the relationship between population mortality and airborne particulate matter." manuscript.

Wilson, A., J. T. Tuomisto, et al. (2006b). "Mortality impact of the 1991 Kuwait oil fires: quantitative estimates of uncertainty via expert judgment." manuscript.

Table 1. Calibration, informativeness, and weights of the six experts and the combined performance-based decision-maker (PDM), with and without optimisation, and equal-weight decision-maker (EDM). All 12 seed variables were used in calculations. All experts with a calibration score less than the significance level (0.0855) found by the optimisation procedure are unweighted when creating the PDM with optimisation. An expert's unnormalised weight in column 5 is shown in parenthesis if it was below cut-off. The last column sums to 1 when PDM weights are excluded; they were derived from other normalisations.

Expert or Decision Maker	Calibration (p-value)	Mean Relative Information		Weight Un-normalized	Information Relative to EDM		Weight Normalized	
		All Variables	Calibration Variables		All Variables	Calibration Variables	without DM	with DM
A	0.00051	1.75	1.68	(0.0009)	0.89	0.98	0	0.0014
B	0.12	1.59	1.49	0.18	0.60	0.68	0.90	0.29
C	0.081	1.39	0.88	(0.071)	0.62	0.62	0	0.11
D	0.085	0.80	0.23	0.020	0.484	0.45	0.10	0.032
E	0.00003	2.32	2.67	(0.00008)	1.23	1.58	0	0.00013
F	0.00063	1.49	1.24	(0.0008)	0.62	0.64	0	0.0013
EDM	0.65	0.81	0.54	0.35	- -	- -	- -	0.56
PDM	0.58	1.01	0.81	0.47	- -	- -	- -	0.70
PDM no optimization	0.35	0.962	0.73	0.26	- -	- -	- -	0.49

Table 2. Robustness on items with the performance-based decision-maker (PDM).

Excluded Item	Relative Information of Perturbed PDM		Calibration of Perturbed PDM	Relative Information with respect to Original PDM	
	All Variables	Calibration Variables		All Variables	Calibration Variables
S1	0.95	0.62	0.67	0.33	0.25
S2	0.87	0.60	0.67	0.24	0.24
S3	0.99	0.80	0.54	0.041	0.047
S4	0.96	0.70	0.086	0.21	0.18
S5	0.98	0.79	0.54	0.074	0.084
S6	0.94	0.72	0.60	0.11	0.12
S7	0.93	0.65	0.67	0.25	0.19
S8	1.18	0.88	0.37	0.39	0.33
S9	0.84	0.51	0.70	0.23	0.20
S10	1.04	0.83	0.70	0.037	0.041
S11	1.53	1.39	0.32	0.45	0.51
S12	1.55	1.43	0.32	0.45	0.52

Table 3. Robustness on experts with the performance-based decision-maker (PDM).

Excluded Expert	Relative Information of Perturbed PDM		Calibration of Perturbed PDM	Relative Information with respect to Original PDM	
	All Variables	Calibration Variables		All Variables	Calibration Variables
A	0.91	0.81	0.58	0.0070	2.7 E-07
B	0.85	0.39	0.061	0.56	0.46
C	0.99	0.81	0.58	3.1 E-05	1.2 E-07
D	0.86	0.47	0.31	0.34	0.29
E	0.89	0.81	0.58	0.00068	7.5 E-08
F	0.97	0.80	0.58	0.0076	0.0058

Table 4. Uncertainty distribution for variables of interest and seed variables, performance-based decision-maker. All variables are on uniform (not log-uniform) scale.

Id	Question	Fractile					Realisation
		5%	25%	50%	75%	95%	
USlong	Q1	0.060	0.42	0.60	1.0	3.8	--
EUlong	Q2	0.029	0.14	0.62	1.1	3.9	--
USshort	Q3	0.016	0.12	0.33	0.54	0.74	--
MXshort	Q4	0.029	0.25	0.47	0.68	0.90	--
EUshort	Q5	0.016	0.088	0.23	0.35	0.70	--
US3mo	Q6	0.0013	0.0096	0.026	0.043	0.080	--
USin1wk(frac)	Q7	0.00002	0.032	0.40	0.59	0.75	--
USin3mo(frac)	Q8	0.05	0.1	0.52	0.69	0.80	--
Difftxhi	Q9	0.062	0.68	1.1	2.4	8.0	--
Difftxlo	Q0	0	0.018	0.25	0.36	1.0	--
Kuwaitperc	Q11	0.030	0.13	0.30	0.46	4.5	--
Kuwaitnr	Q12	4.0	16	35	54	780	--
L01>50	S1	2.6	9.8	20	40	68	10
L01<30	S2	50	100	170	250	310	300
L97>50	S3	2.6	9.8	20	40	68	22
L97<30	S4	50	100	170	250	310	240
A01>50	S5	45	100	230	300	340	240
A01<30	S6	15	30	50	62	120	27
Lnahi	S7	0.88	1.0	1.0	1.0	1.3	0.91
Lcvhi	S8	0.93	1.0	1.0	1.0	1.2	0.96
Anahi	S9	0.83	1.0	1.0	1.0	1.4	1.0
Acvhi	S10	0.90	1.0	1.0	1.0	1.3	1.1
Analo	S11	0.62	0.95	0.98	0.99	1.2	1.1
Acvlo	S12	0.62	0.93	0.96	0.98	1.2	1.1

Table 5: Uncertainty distribution for variables of interest and seed variables, equal-weight decision-maker.

Id	Question	Fractile					Realisation
		5%	25%	50%	75%	95%	
USlong	Q1	0.018	0.46	0.97	1.6	4.5	--
EUlong	Q2	0.034	0.41	0.98	1.6	4.6	--
USshort	Q3	0.0088	0.071	0.13	0.52	1.6	--
MXshort	Q4	0.010	0.19	0.49	0.88	1.9	--
EUshort	Q5	0.013	0.077	0.14	0.46	1.6	--
US3mo	Q6	0.00074	0.0070	0.013	0.042	0.16	--
USin1wk(frac)	Q7	0.0014	0.012	0.054	0.19	0.63	--
USin3mo(frac)	Q8	0.053	0.14	0.25	0.38	0.74	--
Difftxhi	Q9	0.087	0.87	1.9	3.7	10	--
Difftxlo	Q10	0	0	0.11	0.43	2.0	--
Kuwaitperc	Q11	0.0022	0.22	0.85	1.7	6.5	--
Kuwaitnr	Q12	1.7	18	82	250	8000	--
L01>50	S1	2.2	10	19	36	81	10
L01<30	S2	9.4	82	220	260	320	300
L97>50	S3	2.5	11	19	35	82	22
L97<30	S4	9.4	84	200	250	320	240
A01>50	S5	53	150	190	270	330	240
A01<30	S6	7.0	22	39	56	130	27
Lnahi	S7	0.84	1.0	1.0	1.0	1.3	0.91
Lcvhi	S8	0.92	1.0	1.0	1.1	1.2	0.96
Anahi	S9	0.77	1.0	1.0	1.1	1.5	1.0
Acvhi	S10	0.86	1.0	1.0	1.1	1.4	1.1
Analo	S11	0.54	0.88	0.97	0.99	1.3	1.1
Acvlo	S12	0.55	0.82	0.96	0.99	1.3	1.1

Table 6. Comparison of estimates of all cause mortality, percent increase per 1 μg/m^3 increase in $PM_{2.5}$ (question Q1 and selected epidemiological studies). ACS = American Cancer Society Study, SCS = Harvard Six Cities Study.

Study (reference)	Median/best estimate	95% quantile	5% quantile	Difference 95%-5%	Ratio 95%/5%
ACS (Pope, Thun et al. 1995)	0.64	0.89	0.38	0.60	2.79
ACS reanalysis (Krewski, Burnett et al. 2000)	0.68	0.92	0.42	0.59	2.60
ACS update (Pope, Burnett et al. 2002)	0.58	0.94	0.22	0.85	6.72
SCS (Dockery, Pope et al. 1993)	1.25	1.92	0.50	1.70	6.00
SCS reanalysis (Krewski, Burnett et al. 2000)	1.34	2.01	0.57	1.71	5.13
SCS update (Laden, Schwartz et al. 2006)	1.50	2.17	0.78	1.66	3.63
Equal-weight decision-maker	0.97	4.54	0.02	4.52	257
Performance-based decision-maker	0.60	3.80	0.06	3.74	63.7

Figure legends

Figure 1. Questions of interest as assessed by the six experts and equal-weight and performance-based decision-makers. The box shows the median and interquartal range, and whiskers show 90 % confidence interval. The vertical lines in Q1, Q2, Q9, and Q10 show the union of the 95 % confidence intervals of ACS and SCS reanalyses (Krewski, Burnett et al. 2000). Question 11 shows the distributions in relation to each expert's median; experts used different effect windows, and therefore these assessments cannot be directly combined.

Figure 2. Seed variables as assessed by the six experts and equal-weight and performance-based decision-makers. The box shows the median and interquartal range, and whiskers show 90 % confidence interval. The vertical lines show the realisations.

Figure 1.

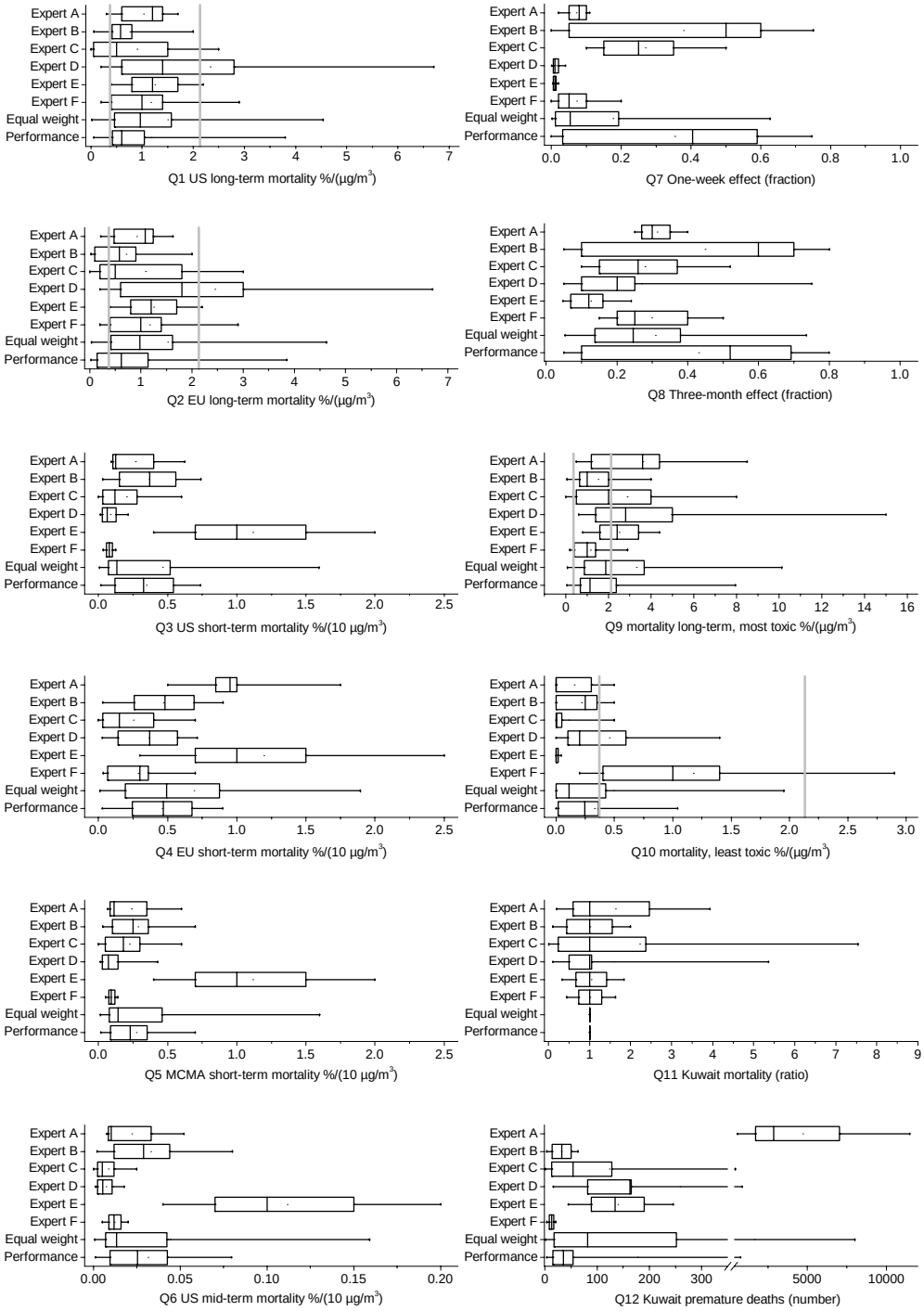
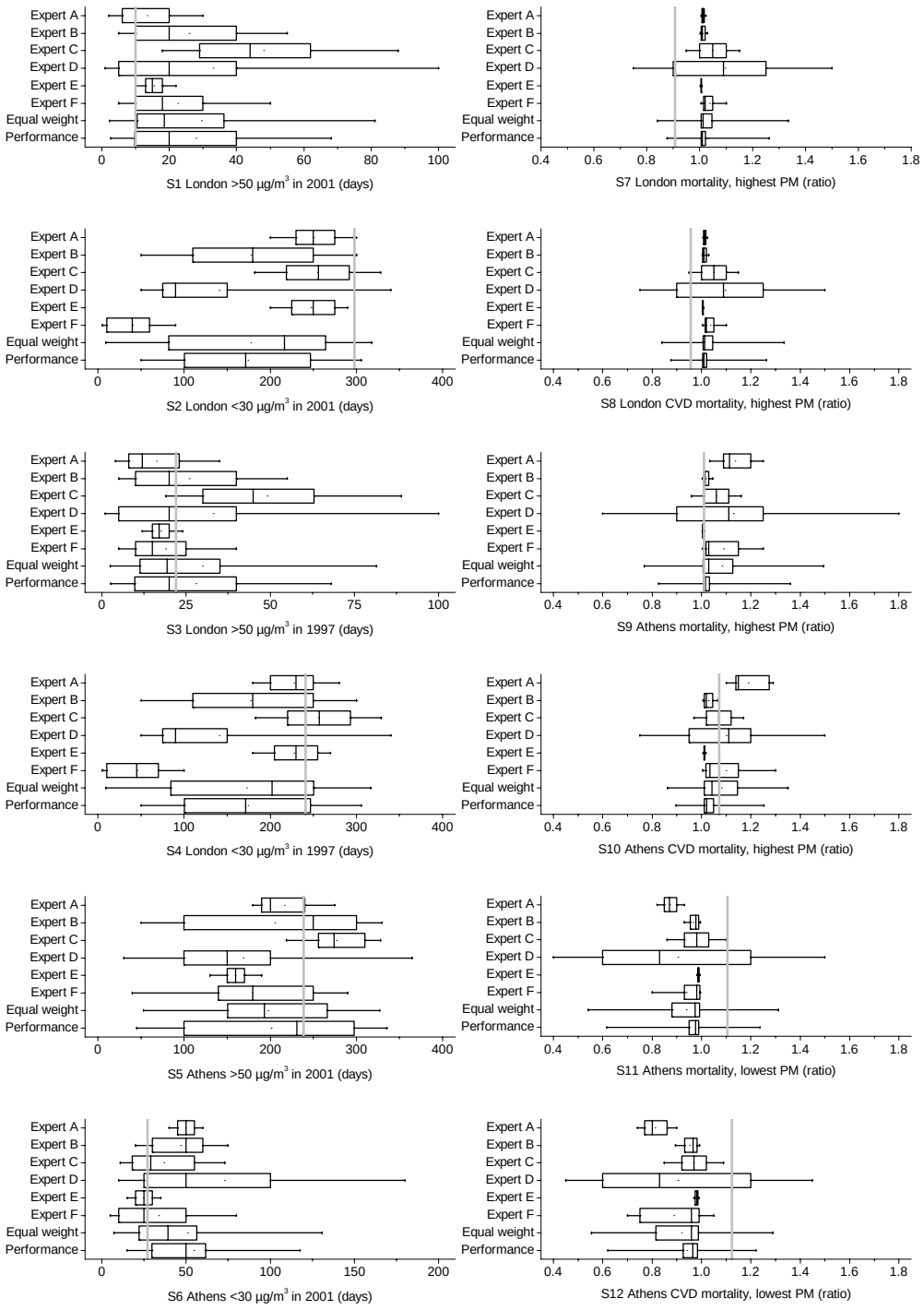

Expert A
Expert B
Expert C
Expert D
Expert E
Expert F
Equal weight
Performance
Q1 US long-term mortality %/(µg/m3)
Q7 One-week effect (fraction)
Q2 EU long-term mortality %/(µg/m3)
Q8 Three-month effect (fraction)
Q3 US short-term mortality %/(10 µg/m3)
Q9 mortality long-term, most toxic %/(µg/m3)
Q4 EU short-term mortality %/(10 µg/m3)
Q10 mortality, least toxic %/(µg/m3)
Q5 MCMA short-term mortality %/(10 µg/m3)
Q11 Kuwait mortality (ratio)
Q6 US mid-term mortality %/(10 µg/m3)
Q12 Kuwait premature deaths (number)

Figure 2.


Expert A
Expert B
Expert C
Expert D
Expert E
Expert F
Equal weight
Performance
S1 London >50 µg/m3 in 2001 (days)
S7 London mortality, highest PM (ratio)
S2 London <30 µg/m3 in 2001 (days)
S8 London CVD mortality, highest PM (ratio)
S3 London >50 µg/m3 in 1997 (days)
S9 Athens mortality, highest PM (ratio)
S4 London <30 µg/m3 in 1997 (days)
S10 Athens CVD mortality, highest PM (ratio)
S5 Athens >50 µg/m3 in 2001 (days)
S11 Athens mortality, lowest PM (ratio)
S6 Athens <30 µg/m3 in 2001 (days)
S12 Athens CVD mortality, lowest PM (ratio)