Technical Comments to CASAC on the *Policy Assessment for the Review of the Ozone National Ambient Air Quality Standards, External Review Draft*

Anne E. Smith, Ph.D., Managing Director
Garrett Glasgow, Senior Consultant
NERA Economic Consulting
November 26, 2019

On October 31, 2019, the U.S. Environmental Protection Agency (EPA, or "the Agency") released its *Policy Assessment for the Review of the Ozone National Ambient Air Quality Standards, External Review Draft* (hereafter referred to as the "Draft PA").[1] On December 3-6¸ the Clean Air Scientific Advisory Committee (CASAC) will be convening to, among other matters, develop its comments to EPA on the Draft PA. The following document is intended to provide technical comments and additional information that is of relevance and potential interest to the CASAC, as well as to the Agency. We have prepared these with financial support from a coalition of industry associations.

Our comments are focused specifically on the question of whether the Draft PA's quantitative estimates of health risks and associated discussions of risk-based considerations provide information that is useful and reliable as guidance to decision makers when considering the merits of the current and alternative potential National Ambient Air Quality Standards (NAAQS).

Based on our review of the Draft PA and its underlying data and models:

- We concur with the Agency's finding that the exposure and associated health risk estimates under the current standard of 70 ppb are similar to those estimated when that standard was set in 2015.
- We also concur that the uncertainties associated with quantitative risk estimates have not significantly changed since the 2015 review.

With regard to uncertainties in modeling health risks, we also concur with the following methodological decisions in the Draft PA's quantitative risk assessment:

- It is not unreasonable for the Draft PA to give relatively greater weight to lung function risk estimates based on the E-R model over the MSS model. Although we find that estimates from both approaches are subject to large statistical error ranges and a variety of model uncertainties, the Draft PA provides clear evidence that the MSS-based model estimates are affected by extrapolation beyond the base of scientific evidence more than those from the E-R model.
- We conclude that the Agency has made a reasonable judgment not to have conducted any epidemiologic-based risk calculations, because we have found no evidence that the Agency has yet developed appropriate tools for quantifying the role of the very extensive epistemic uncertainties in those types of risk calculations.

[1] EPA, *Policy Assessment for the Review of the Ozone National Ambient Air Quality Standards, External Review Draft*, EPA-452/P-19-002, Office of Air Quality Planning and Standards, Research Triangle Park, NC, October 2019.

1. Summary of Quantitative Risk Calculations in Draft PA

The Draft HREA's quantitative risk calculations are limited to estimates of individual-level ozone exposures and associated lung-function responses. These are calculated using the Air Pollutants Exposure Model (APEX). In contrast to the 2014 Health Risk and Exposure Assessment or "2014 HREA" (U.S. EPA, 2014), the Draft PA does not provide *any* estimates of risks that are epidemiologic-based.[2] Instead, it relies on data from specific epidemiologic papers only to summarize the distributions of air concentrations of ozone that were observed in each study.[3] That is, the epidemiological papers are evaluated only in the context of the typical "evidence-based" reasoning, and no quantitative risk estimates are calculated.

In these comments, we discuss how uncertainties – both statistical and epistemic ("model") uncertainties – affect the reliability of both APEX-based and epidemiologic-based types of calculations for guidance on the adequacy of the current standard and on potential alternative standards. To help better quantify the magnitude and direction of some key uncertainties on estimates in the Draft PA, we supplement our discussion with additional quantitative sensitivity analyses that we have been able to perform.

The APEX-based risk analysis is as extensive (arguably more extensive) as that which appeared in the 2014 HREA. It follows nearly the same format as in the 2014 HREA, focusing on the same measures of impact. These include the percent of population groups exceeding certain benchmark levels of ozone exposure with various frequencies in a given year (called "exposure risks"), and the percent of population groups experiencing certain percentage decrements in FEV1 with various frequencies in a given year (called "lung function" risks).[4] The population subgroups include all children (ages 5-18), asthmatic children, all adults, and asthmatic adults. (In 2014, adults were analyzed in two age groups that are no longer separately analyzed). In the Draft PA, the discussion and interpretation of the risk results are focused on those for children only, and we do the same in our sensitivity analyses.

Although risk estimates are reported for asthmatic subgroups of all children and all adults, those impacts, when stated as a percent of all individuals, are essentially the same as those for all individuals. This similarity is because the APEX analysis does not alter the assumptions about the behavioral patterns of asthmatics in time and space from those of non-asthmatics.[5] Thus, when we present additional sensitivity analysis results, we provide only results for all individuals, but the reader should be aware that the results we present would be largely the same for the asthmatic subsets.[6]

The Draft PA provides risks estimates for 8 cities compared to 15 cities in the 2014 HREA; 7 of the 8 cities were analyzed in the 2014 HREA.[7] In both HREAs, risks are calculated for air quality scenarios

[2] Epidemiologic-based risk calculations would likely have been performed using the BenMAP model. Even the word "BenMAP" does not appear in the draft.

[3] These data appear in Appendix B.

[4] FEV1 refers to the volume of air that a person can forcibly expel in one second. The Draft PA reports estimates of *decrements* in FEV1 that are categorized into ≥ 10%, 15%, and 20% declines from an individual's normal FEV1 level, and labelled "dFEV1."

[5] Draft PA, p. 3D-53.

[6] This is true because we only present results for the *percent* of affected populations. For any given percent, the *numbers* of affected individuals will always be much smaller for the asthmatics than for all individuals, simply because the populations of asthmatics in each city are (roughly) 1/10th the size of the respective full population.

[7] The new city in this risk analysis is Phoenix.

that just attain the current standard (*i.e.*, 70 ppb), as well as scenarios consistent with 75 ppb and 65 ppb design values. (The 2014 HREA included a couple of other air quality scenarios.)

The calculations of lung function risks are, as in the 2014 HREA, derived from two models, both based on the results of the same set of clinical studies: a population-weighted exposure-response (E-R) model, and one version of the McDonnell-Stewart-Smith (MSS) exposure-response model. The E-R model parameters are identical to those used in the 2014 HREA. The MSS model has been updated since the 2014 HREA: this Draft PA uses MSS-2013 (from McDonnell *et al.*, 2013) while the 2014 HREA used MSS-2012 (from McDonnell *et al.*, 2012).

Many of the inputs required by APEX have been updated since those of the 2014 analysis. In addition to the updated MSS model parameters (which affect only the MSS-based lung function risk estimates), there are at least five other categories of input assumptions that have been updated, including the hour-to-hour ozone levels across each city that are predicted under each ozone design value level and demographic, physiological, activity pattern, and meteorological input assumptions. The latter changes affect all exposure and lung function estimates. Even the APEX code itself has been updated (from version 4.5 used in 2014 to version 5.12a used now).[8]

In regard to epidemiologic risk evidence, the Draft PA notes that the epidemiologic-based risk uncertainties were so large that those portions of the HREA were given little weight in the 2014 decision. It also takes the position, through qualitative discussion, that there has been no progress in reducing those uncertainties since 2014, implying that there is no usefulness in producing those types of risk calculations again at this time.

Given the overlaps in the scenarios and population groups analyzed, there are reasonable points of comparison between the current and prior exposure and lung function risk estimates. Such comparisons are a focus of how the Draft PA makes use of the new HREA results:

- The Draft PA uses the comparability of estimated risks under the current standard versus those in the 2014 HREA to make the case that given that 70 ppb was found to be requisitely protective in 2015, it would remain requisitely protective today unless there were a change in the degree of uncertainty in scientific evidence regarding those risk estimates.

- The Draft PA also makes the case that there has been no significant change in uncertainties in the risk analysis since 2014. It therefore concludes that no change is warranted in the current primary standard.

The Draft PA notes that estimates of exposures exceeding various benchmark levels were the primary metrics used in 2015 when determining that the 70 ppb standard reduced risks sufficiently (from those estimated for the then-current 75 ppb standard) to provide adequate protection of the public health. In recounting that evaluation, the Draft PA indicates that the benchmark exposure metric of greatest concern were those for children and those that involved more multiple exceedances per year for the lower two benchmark levels of 60 ppb and 70 ppb. It also mentions some concern with even one exceedance per year for the highest benchmark level of 80 ppb. The Draft PA does not, however, provide a clear summary table to support its finding that the estimates of these exposure risks are similar to those in the

[8] Although the APEX code has been changed, we were able to run a set of the old input files from the 2014 HREA through APEX 5.12a and obtained dFEV1 results very similar to those in the 2014 HREA. The very minor differences that resulted are probably due to our comparison run including only 10,000 iterations, whereas the original 2014 HREA results were based on 200,000 iterations. Thus, we believe that the changes to the APEX code have not changed what it predicts in the way of benchmark or dFEV1 outcomes *for any given set of input assumptions*. We, therefore, assume that all differences between the 2014 HREA results and those in the Draft PA are due solely to the (many) changes in input assumptions.

2014 HREA. We have done so in tables below, which we find confirm that the Draft PA is correctly summarizing the situation.

Table 1 provides a summary of the average and maximum percentages of all children projected to experience exceedances of the above-mentioned benchmark exposures of greatest interest in the prior ozone NAAQS review. This comparison includes the seven cities that were analyzed in both the 2014 HREA and in the current Draft PA. (Phoenix is the only city in the current Draft PA that cannot be included in this comparison.) Table 2, on the following page, presents the comparisons individually for each of those seven cities.

Across the board, the tables show equivalent or smaller exposure risk estimates in the Draft PA than in the 2014 HREA. One clarification on this statement is warranted. Our tables report the values *exactly as stated in each document*; however, the 2014 HREA reported values that rounded to less than 0.1% as "0" rather than "<0.1." Thus, although average values of "0" from 2014 might appear to be lower than a "<0.1" in the Draft PA, they should actually be interpreted as equivalently small – too small to differentiate.[9]

We cannot at this time determine the specific reasons for the somewhat lower benchmark exceedance risks in the Draft PA. Nevertheless, we have been able to replicate results in both the Draft PA and the 2014 HREA, using the new input files and the 2014 input files, respectively. Thus, the differences must be caused by some combination of the air quality, physiological, behavioral, and meteorological updates to the inputs to APEX.[10] It would have been helpful if the Agency could have provided a breakdown of how each category of update contributed to this overall reduction in benchmark risks at comparable standard levels.

Table 1. Averages of Draft PA's Benchmark Exceedance Risks Compared to Those in 2014 HREA for the 70 ppb Standard for the 7 Cities in Both Documents (% of all children while at elevated exertion)

		At Least Two Exposures				At Least One Exposure	
		60 ppb		70 ppb		80 ppb	
		Average	Maximum	Average	Maximum	Average	Maximum
All Cities	**Draft PA**	0.6 – 1.7	2.8	<0.1	0.1	0 - <0.1	0.1
	HREA 2014	1.5 – 3.2	7.1	0 – 0.1	0.4	0 – 0.1	0 – 0.2

Notes: (1) The 2014 HREA reported values that rounded to less than 0.1 as "0." Thus, all the values of "0" in the table above should be interpreted as "<0.1," as is used in the Draft PA; (2) The 7 cities that are included in the averages in the above table are the cities in Table 2 below.

[9] Confirmation of this point can be seen in Table 2, where the average values are sometimes reported to be "0" even though the maximum value is non-zero. An average cannot be exactly zero if any of its components is non-zero.

[10] We have also checked and found that the differences in results are due to the input file differences and not any changes within the APEX code.

Table 2. City-Specific Benchmark Exceedance Risk Comparisons for the 70 ppb Standard for the 7 Cities in Both Documents (% of all children experiencing while at elevated exertion)

		At Least Two Exposures				**At Least One Exposure**	
		60 ppb		**70 ppb**		**80 ppb**	
		Average	**Maximum**	**Average**	**Maximum**	**Average**	**Maximum**
Atlanta	**Draft PA**	0.6	1.1	<0.1	<0.1	<0.1	0.1
	HREA 2014	2.1	3.3	0	0.1	0.1	0.2
Boston	**Draft PA**	0.8	1.4	<0.1	<0.1	<0.1	<0.1
	HREA 2014	2.2	5.5	0.1	0.4	0.1	0.2
Dallas	**Draft PA**	1.2	2.1	<0.1	0.1	<0.1	<0.1
	HREA 2014	2.2	7.1	0	0.1	0	0.1
Detroit	**Draft PA**	1.7	2.8	<0.1	0.1	<0.1	<0.1
	HREA 2014	1.9	3.6	0	0.1	0	0
Philadelphia	**Draft PA**	0.8	0.9	<0.1	<0.1	<0.1	<0.1
	HREA 2014	1.7	3.3	0	0.1	0	0.1
Sacramento	**Draft PA**	0.6	0.9	<0.1	<0.1	0	0
	HREA 2014	1.5	3.4	0	0.1	0	0
St. Louis	**Draft PA**	1.5	2.6	<0.1	<0.1	<0.1	<0.1
	HREA 2014	3.2	7.0	0.1	0.3	0.1	0.2

Note: The 2014 HREA reported values that rounded to less than 0.1 as "0." Thus, all the values of "0" in the table above should be interpreted as "<0.1," as is used in the Draft PA.

The Draft PA notes that lung function risk estimates were given less weight in the 2015 NAAQS decision than the projected benchmark exposure exceedances, due to their greater degree of model uncertainty Also, it notes that epidemiologic-based risk estimates were given the least weight due to their even greater degree of model uncertainty. While there is qualitative discussion of uncertainties in the Draft PA's Appendix 3D, more could be done to explain these results in the main body, and more could be done to quantify the range of model uncertainties in the lung function risk estimates, as will be discussed in these comments. In Section 2, we discuss the nature of the uncertainties in the lung function risk estimates in the Draft PA. Section 3 then discusses some key uncertainties that justify not having conducted quantitative estimates of epidemiologic-based risks.

2. Uncertainties in Lung Function Risk Calculations

The lung function estimates in the Draft PA come from two models: a "population-based" exposure response model ("E-R") and the "individual level" MSS-based E-R calculation (MSS-2013). Both models are derived from results from the same set of clinical studies. Also, both models are used to generate risk estimates in the same units: dFEV1≥10%, 15% and 20%, at a frequency of ≥1, 2 and 4 times per year. They differ in the following ways:

- The E-R model estimates the percentage of the population experiencing such events directly from APEX-based estimates of numbers of people in each ozone exposure level (based on their daily maximum 7-hour average while at moderate or higher exertion), with increasing probability of such a dFEV1 event occurring as the ozone exposure level rises. The Draft PA's calculations use a 7-hour exposure period, whereas in the 2014 HREA, an 8-hour exposure period was used; otherwise, the risk relationship is unchanged. The shift to a 7-hour period is because this period better matches the 6.6-hour period over which most of the clinical studies were conducted. (Presumably the 2014 HREA used an 8-hour exposure period because that is the averaging period of the ozone NAAQS. However, for purposes of minimizing uncertainty in the model's projections of responses, we would concur that the averaging time of the original data is the appropriate metric to match when projecting future risks.)[11]
- To produce the "individual level" MSS-based calculation, the APEX model generates a set of simulated individuals with different ages, body mass indexes, etc., each of whom experiences a sequence of "events" as they move through time and space over the course of an ozone season. Each event is an activity that takes place in a specific location ("microenvironment") for a duration between 1 and 60 minutes. For example, a simulated individual might commute to work by car for 30 minutes, then work in an office for 60 minutes, and so on. For each simulated individual, this sequence of events is combined with data on the hourly ozone concentrations relevant to the urban area and ozone scenario being studied to calculate that individual's sequence of ozone doses over the year. That calculation uses a complex set of statistically-estimated parameters known as the MSS model parameters. When combined with daily random "intra-individual" variations in dFEV1 (another statistically-estimated parameter), APEX calculates each event in which a simulated person's dFEV1 exceeds 10%, 15% or 20%, and reports the number of days in the simulated ozone season in which each dFEV1 level occurs. These events are divided by the number of total individuals simulated to obtain the percentage of the population affected.

2.1. Model Uncertainty Evident in Contrasting Results of E-R vs. MSS Approaches

The very fact that there are two different types of models that make inferences from the same evidence base (*i.e.*, the clinical studies data) to predict lung function responses under future alternative real world conditions provides an empirical basis for assessing the degree of model uncertainty in the lung function risk estimates.

[11] We find that the estimates of benchmark exceedances are also affected in a similar manner by the shift from 8-hour averaging to 7-hour averaging. That is, use of 8-hour average exposures substantially reduces the percentages of individuals experiencing benchmark exceedances of all types analyzed. The magnitude of the reduction is about 40% for exceedances of the 60 ppb benchmark at least once per year, and the reduction is larger for higher benchmarks and for more exceedances per year. (As one would expect, the averaging period input assumption does not affect MSS-based risk estimates, as the MSS model uses its own ozone exposure estimation method.)

Table 3 below is a copy of Draft PA Table 3-4 that provides a quick perspective on the degree of this model uncertainty. The E-R risk estimates (in the top segment of the table) are much lower than those from the MSS-2013 model (in the bottom segment of the table). For example, MSS-2013 estimates about 4.1% to 7.1% of all children will experience dFEV1≥15% at least once per year under the current standard, while the E-R model predicts it will be only 0.5% to 0.8% – about 90% lower. A similar magnitude of difference can be seen for all of the other comparisons that can be made from this table.

Table 3. Comparison of dFEV1 Results from E-R Model versus MSS-2013 Model (Source: Table 3-4 of Draft PA)

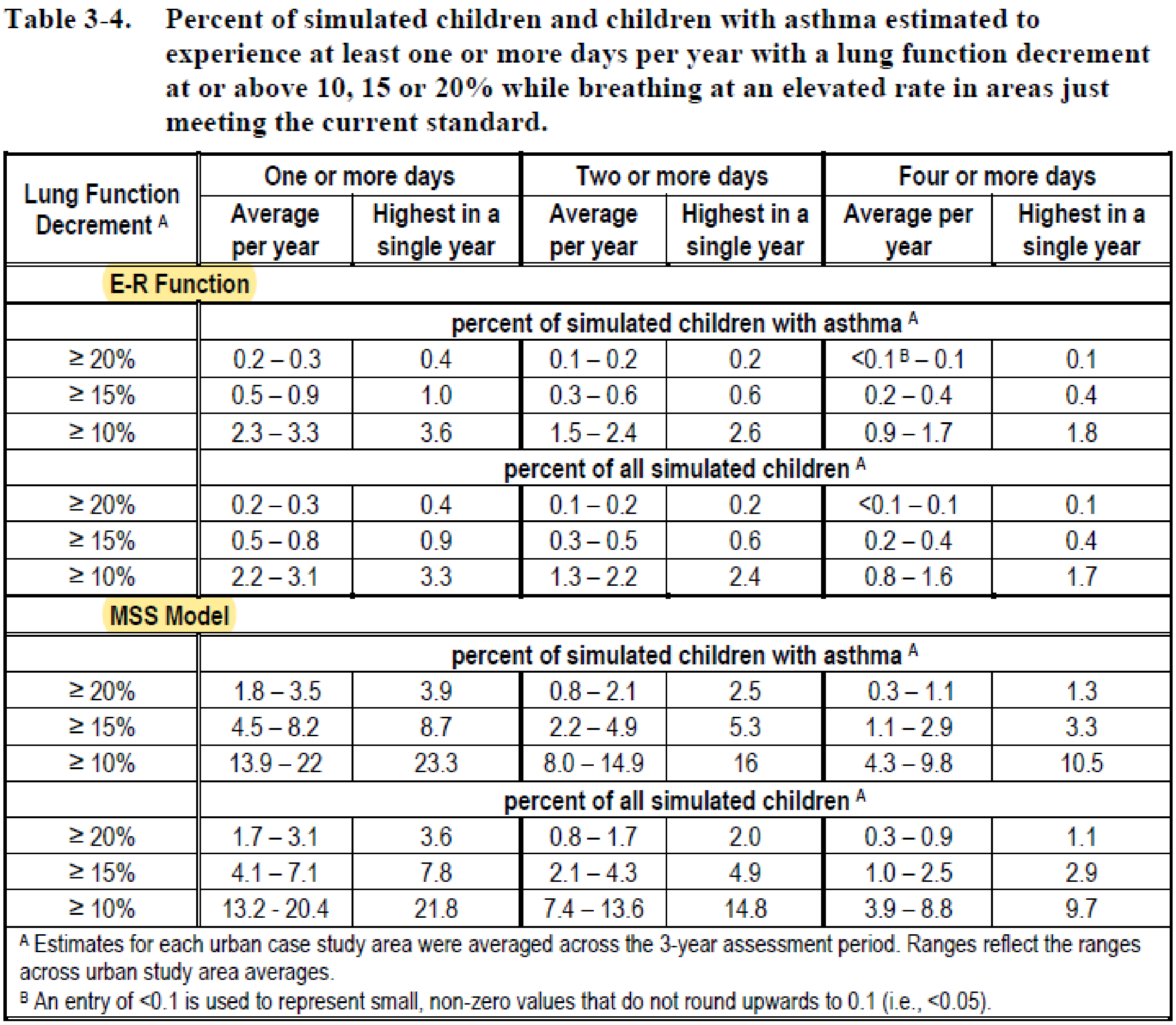


Table 3-4. Percent of simulated children and children with asthma estimated to experience at least one or more days per year with a lung function decrement at or above 10, 15 or 20% while breathing at an elevated rate in areas just meeting the current standard.

Lung Function Decrement[A]	One or more days		Two or more days		Four or more days	
	Average per year	Highest in a single year	Average per year	Highest in a single year	Average per year	Highest in a single year
E-R Function						
	percent of simulated children with asthma[A]					
≥ 20%	0.2 – 0.3	0.4	0.1 – 0.2	0.2	<0.1[B] – 0.1	0.1
≥ 15%	0.5 – 0.9	1.0	0.3 – 0.6	0.6	0.2 – 0.4	0.4
≥ 10%	2.3 – 3.3	3.6	1.5 – 2.4	2.6	0.9 – 1.7	1.8
	percent of all simulated children[A]					
≥ 20%	0.2 – 0.3	0.4	0.1 – 0.2	0.2	<0.1 – 0.1	0.1
≥ 15%	0.5 – 0.8	0.9	0.3 – 0.5	0.6	0.2 – 0.4	0.4
≥ 10%	2.2 – 3.1	3.3	1.3 – 2.2	2.4	0.8 – 1.6	1.7
MSS Model						
	percent of simulated children with asthma[A]					
≥ 20%	1.8 – 3.5	3.9	0.8 – 2.1	2.5	0.3 – 1.1	1.3
≥ 15%	4.5 – 8.2	8.7	2.2 – 4.9	5.3	1.1 – 2.9	3.3
≥ 10%	13.9 – 22	23.3	8.0 – 14.9	16	4.3 – 9.8	10.5
	percent of all simulated children[A]					
≥ 20%	1.7 – 3.1	3.6	0.8 – 1.7	2.0	0.3 – 0.9	1.1
≥ 15%	4.1 – 7.1	7.8	2.1 – 4.3	4.9	1.0 – 2.5	2.9
≥ 10%	13.2 - 20.4	21.8	7.4 – 13.6	14.8	3.9 – 8.8	9.7

[A] Estimates for each urban case study area were averaged across the 3-year assessment period. Ranges reflect the ranges across urban study area averages.

[B] An entry of <0.1 is used to represent small, non-zero values that do not round upwards to 0.1 (i.e., <0.05).

The Draft PA emphasizes estimates of lung function results from the E-R model because its analyses (documented in Draft PA Section D.3.4.2) find that the E-R results are less affected by extrapolation outside the range of exposures in clinical studies. For example, small portions of the E-R lung function exceedances are attributable to ozone exposures that are less than 40 ppb – a range not studied in any of the clinical trials. The Draft PA reports that between 11% and 16% of the E-R-based risks are attributable to 7-hour average ozone exposures <40 ppb under the current standard (for dFEV1≥10% at least once per year).[12] In contrast, between 57% and 67% of the MSS-2013-based risks for this category are predicted to be associated with exposure <40 ppb.[13] This difference suggests that perhaps a majority of the much higher lung function risk estimates from the MSS-2013 model are attributable to exposures at levels below 40 ppb. Those are estimates that entail more extrapolation outside of the evidence base than the lung function decrements predicted to occur when exposures are at higher levels.

It is because of the above findings from its detailed evaluation of the two models' projections that the Draft PA concludes that there is "appreciably greater uncertainty associated with the MSS model estimates than the E-R function estimates due to the significantly greater portion of relatively low concentrations contributing to risk."[14] Having reviewed the sensitivity analyses supporting this conclusion in the Draft PA,[15] we concur that relying on the MSS-2013 model's risk estimates would be giving greater weight to estimates that have less foundation in the available scientific evidence base – and thus, are subject to greater epistemic uncertainty.

Below, we provide our assessment of model uncertainties and statistical errors associated with the MSS and E-R models individually. We show that both models have additional uncertainties that warrant careful consideration when deciding how much weight to assign to the Draft PA's quantitative estimates of lung function risks under alternative air quality scenarios.

2.2. Uncertainties Within the MSS Approach Alone

In addition to the model uncertainty associated with the choice of the E-R versus MSS approaches, there is also model uncertainty associated with risk estimates from the MSS model approach alone. The Draft PA uses the MSS-2013 model, while the 2014 HREA used the earlier MSS-2012 model.

The MSS-2013 model has been described as better at accounting for intra-subject variability and, thus, producing a better model fit than MSS-2012 (Draft PA 3-52). Despite this assertion, there remains model uncertainty even in the choice between the MSS-2012 and MSS-2013 models.[16] Even if one has strong subjective confidence in relying solely on MSS-2013, there is also model uncertainty in risk estimates that APEX derives from that model. The MSS-2013 model differs from MSS-2012 primarily by breaking the intra-individual error term into two portions – one ("*v*1") that does not vary with the ozone exposure level and the other ("*v*2") that increases as the ozone exposure level increases. McDonnell *et al.* (2013) note that *v*1 could be interpreted as a separate, non-ozone-related contribution to variation in lung function. This suggests that when calculating elevated risks related to ozone exposure, one valid approach would be to set the non-ozone-related error term, *v*1, equal to zero (Draft PA 3D-144). Whether to set the non-ozone-related error term equal to zero or not thus represents another source of model uncertainty in the MSS approach.

[12] Draft PA, Table 3D-62, p. 3D-147. These results are also summarized in Draft PA Table 3-6.

[13] Draft PA, Table 3D-63, p. 3D-150. These results are also summarized in Draft PA Table 3-7.

[14] Draft PA, p. 3-67.

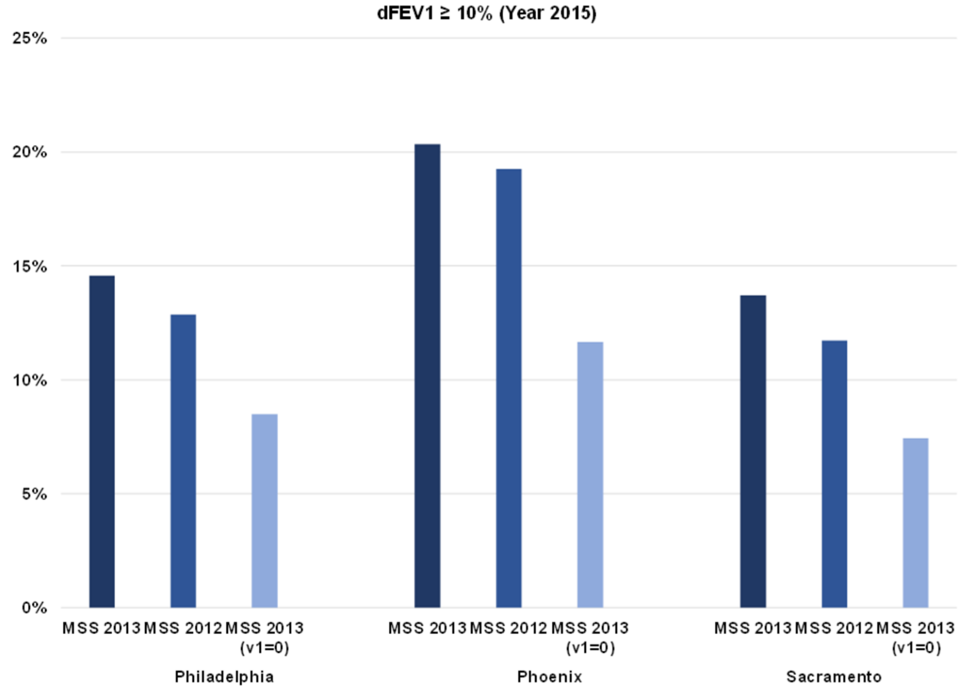
[15] Draft PA, Section 3D.3.4.2, pp. 3D-146 to 3D-156.

[16] In fact, best model fit is not always the standard that the Agency applies when selecting its preferred model. For example, the 2014 HREA used an MSS-2012 model without body mass index (BMI) included, on the grounds that, when included in the model, the BMI variable was statistically insignificant (2014 HREA 6-8). Nevertheless, the model with BMI was the better fitting model *as measured by the AIC* (McDonnell *et al.* 2012, p. 624), and there were clear differences in the risk estimates produced by the BMI and non-BMI versions of the MSS-2012 model (Smith and Glasgow 2015, Glasgow and Smith 2017). We note that the Agency's preference for the MSS-2013 model is based solely on comparing its AIC with that of the MSS-2012 model, while that best-fit criterion was not used for model choice in the 2014 HREA.

NERA has examined the effect of this model uncertainty on the MSS lung function risk calculations by considering the three versions of the MSS model described above: MSS-2012, MSS-2013, and the MSS-2013 model with *v*1 set to zero. To undertake the sensitivity analyses presented below, we used the same version of APEX and the same APEX input files to calculate the MSS lung function risk estimates in the Draft PA. The only differences across the calculations below come from changes to the MSS coefficients in the APEX physiology file, which were edited to match the estimated coefficients for the particular version of the MSS model being examined.[17]

Figure 1 below provides an example of the degree of model uncertainty related to MSS model choice in the MSS lung function risk calculations. This figure presents the estimated risk of a lung function decrement of 10% or greater at least once per year, for children in three of the eight cities examined in the Draft PA (Philadelphia, Phoenix, and Sacramento), for the 2015 ozone season under the scenario in which the 70 ppb ozone standard was just met. Results are presented for the three versions of the MSS model discussed above (MSS-2012, MSS-2013, and MSS-2013 with *v*1 set to zero).[18]

Figure 1. Sensitivity of Draft PA MSS-based risk estimates to version of the MSS model (for children in 3 cities, using 2015 ozone data, dFEV1≥10% at least once per year)



For each city, the leftmost (dark blue) bar is the risk estimate produced by using all of the same inputs as in the Draft PA, including the same version of the MSS model (MSS-2013). As expected, these results

[17] For each version of the MSS model examined, we maintained the EPA's assumptions on how to draw from the MSS error term distributions. Specifically, we specified new error term draws for each day of the ozone season, with the draws bounded at ± 2 standard deviations of the error term distribution.

[18] These estimates are based on 10,000 simulated children, compared to 60,000 in the results in the Draft PA. However, we find that the resulting risk estimates when using all the same input assumptions are close enough to the precise results in the Draft PA to allow a reliable understanding about the sensitivity of those estimates to alternative input assumptions.

are consistent with the results presented in the Draft PA (Draft PA, Table 3D-49).[19] The middle (medium blue) bar for each city shows how this risk estimate changes when the MSS-2012 model that was used in the 2014 HREA is used in place of the MSS-2013 model used in the Draft PA, holding all other inputs equal. This comparison shows that, in this instance, the MSS-2013 model produces higher risk estimates than the MSS-2012 model.

The rightmost (light blue) bar for each city presents the risk estimates when using the MSS-2013 model with $v1$ set equal to zero. A comparison of these risk estimates to the risk estimates based on the MSS-2013 model (with a non-zero $v1$ term) reveals that ad hoc assumptions about the proper way to simulate the effect of the daily intra-individual variation in dFEV1 are a large source of model uncertainty in the MSS lung function risk calculations.

The model uncertainty surrounding the specification of $v1$ is identified in the qualitative risk discussion of the Draft PA (Draft PA, 3D-144). It is recognized as an uncertainty that is likely to overstate risk and described as of "medium" magnitude. Figure 1 provides some quantification of what magnitude should be assigned to the word "medium" – roughly a 70% to 85% overstatement for the dFEV1≥10% at least once per year risk metric.

The MSS-specific model uncertainty described in Figure 1 is fairly similar from year to year and from city to city for a given risk metric. However, it does vary with the risk metric. For example, Figure 2 and Figure 3 (on the next page) present the same sensitivity analyses as were presented in Figure 1, but for higher dFEV1 levels. The results for dFEV1≥15% at least once a year are presented in Figure 2, and the results for dFEV1≥20% at least once a year are presented in Figure 3. Note that the scale of the y-axis is smaller for Figure 2 (and even smaller for Figure 3) in comparison to Figure 1. Because the risk estimates are smaller when considering decrements of larger magnitudes, this change in scale was necessary in order to make the relative differences in the risk estimates between models more visible.

A comparison of Figures 2 and 3 to Figure 1 reveals that the differences in the risk estimates between the MSS-2013 model and the MSS-2012 model are larger as the size of the lung function decrement increases. In some cases, the risk estimate based on the MSS-2012 model is less than half of that obtained when using the MSS-2013 model.

The risk estimates obtained when using the MSS-2013 model with $v1$ set equal to zero are similar to those obtained when using the MSS-2012 model. This indicates that the model uncertainty surrounding the specification of $v1$ is even stronger when considering risk metrics based on larger values of dFEV1. In this case, we see overstatements for the dFEV1≥15% at least once a year and dFEV1≥20% at least once a year risk metrics ranging from 67% to 90%.

[19] The only reason our results are not an exact replication of those in the Draft PA is because ours were generated with 10,000 simulated individuals while those in the Draft PA were generated with 60,000 simulated individuals.

Figure 2. Sensitivity of Draft PA MSS-based risk estimates to version of the MSS model (for children in 3 cities, using 2015 ozone data, dFEV1≥15% at least once per year)

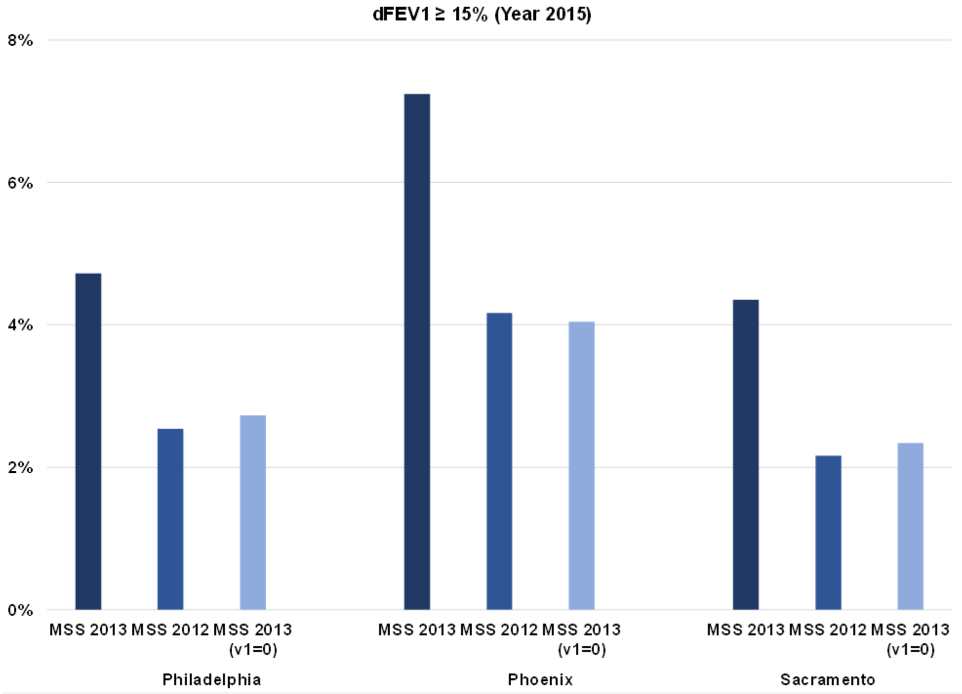
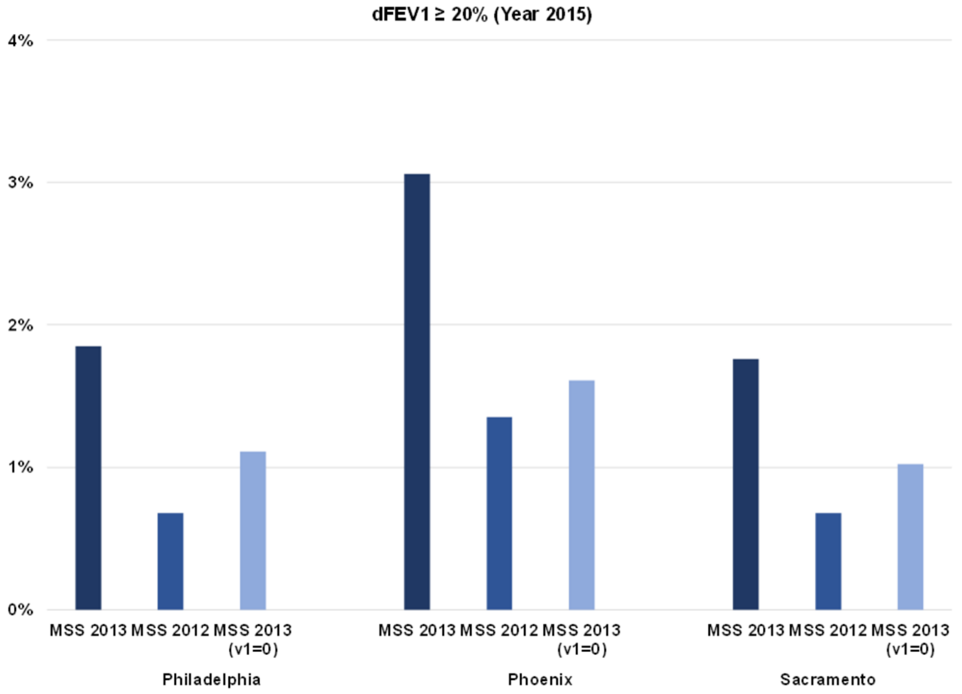


Figure 3. Sensitivity of Draft PA MSS-based risk estimates to version of the MSS model (for children in 3 cities, using 2015 ozone data, dFEV1≥20% at least once per year)



In addition to the model uncertainty described above, the Draft PA's MSS-based risk estimates are also subject to statistical uncertainty. No matter which specification of the MSS model is used, the relevant MSS exposure-response function used in the dFEV1 risk calculations will have been estimated on a sample of clinical observations and, thus, will contain some amount of statistical uncertainty. From this it follows that the HREA's dFEV1 risk estimates based on any given specification of the MSS model will also contain statistical uncertainty (Glasgow and Smith 2017).

We demonstrated the influence of statistical uncertainty on the MSS lung function risk calculations in a report submitted to the 2014 ozone docket (Smith and Glasgow 2015) and a peer-reviewed paper (Glasgow and Smith 2017). The Draft PA acknowledges this additional uncertainty in its qualitative discussion of MSS uncertainties (Draft PA, p. 3D-145), writing:

> *"Glasgow and Smith (2017) evaluated statistical uncertainty in the MSS model employed by APEX. Multiple sets of lung function risk results were generated using random draws of the MSS model coefficients (considering their standard errors) and performing APEX simulations for children ages 5-17 and for 2010 air quality just meeting a design value of 75 ppb in Atlanta. Calculated bounds on the risk estimates could extend to as low as 0% and >35% of children experiencing at least one decrement ≥10%."*

This quote provide a correct, although minimalist, summary of the findings. Specifically, this summary omits several important findings related to the effect of statistical uncertainty on the risk estimates.

Figure 4 (on the next page) is a figure originally included in the NERA report submitted to the 2014 ozone docket (Smith and Glasgow 2015). This figure presents density plots for the risk calculations for dFEV1≥15% at least once, twice, or six times, for the three different age groups considered in the 2014 HREA, for the 2010 ozone season in Atlanta. The variance of these density plots was determined by the statistical uncertainty in the MSS-based risk estimates.[20]

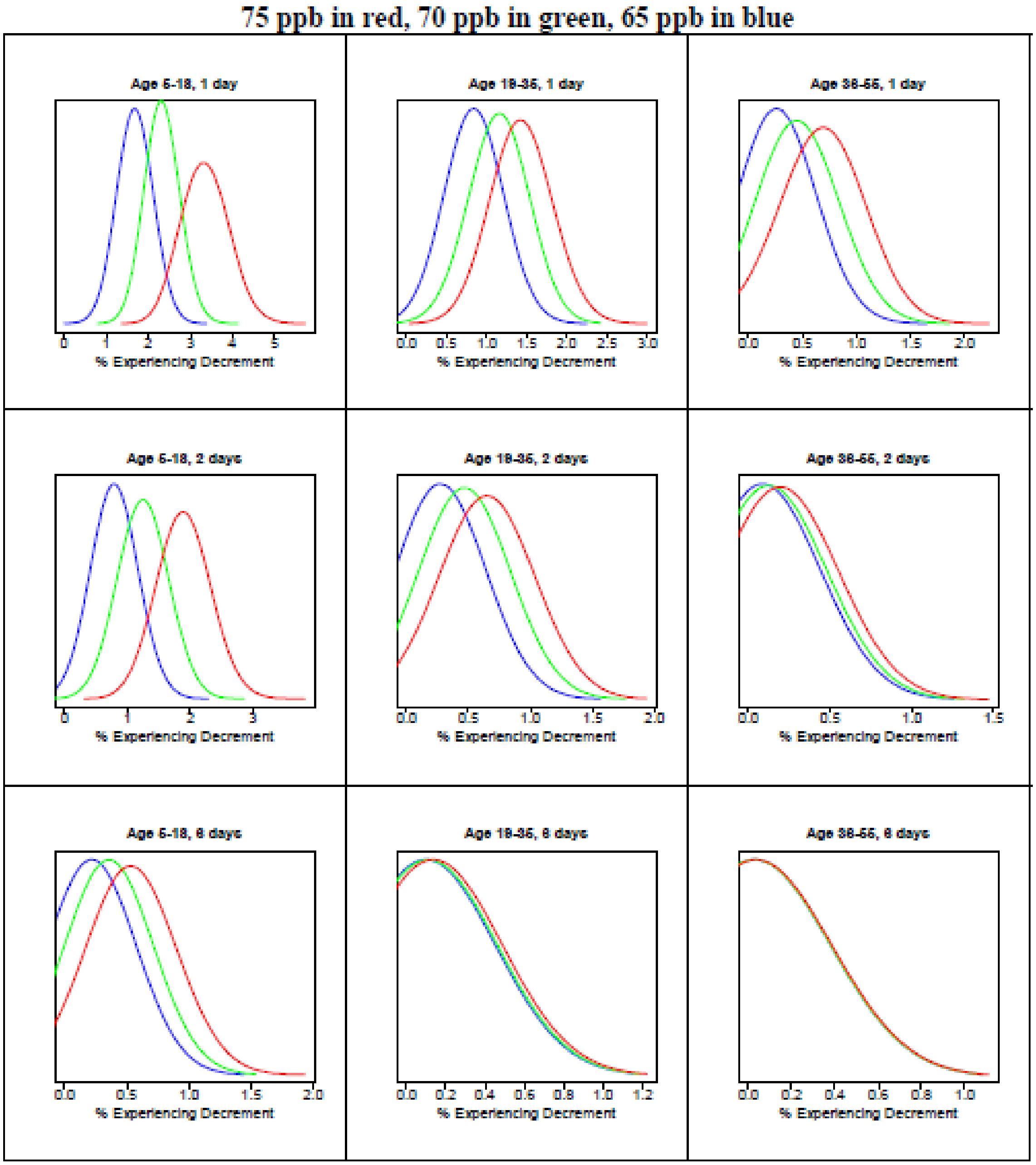
Three different ozone scenarios are presented in each tile of Figure 4 – just attaining a 75 ppb standard (red density plots), a 70 ppb standard (green density plots), and a 65 ppb standard (blue density plots). Examination of these distributions of the risk calculations reveals three important points that were not captured in the summary of the influence of statistical uncertainty as written in the Draft PA.

First, across all age groups and frequencies of decrements (*i.e.*, across all 9 tiles in the figure) one sees significant overlap in the distributions for each ozone scenario. These overlaps suggest uncertainty as to whether a change in ozone standards would lead to a significant reduction in the percentage experiencing this level of lung decrement. For example, for the one-day decrement for ages 5-18 (the upper left corner tile), 43 out of 100 APEX simulations using the 75 ppb ozone scenario were less than or equal to the maximum value estimated by the 100 APEX simulations using the 70 ppb ozone scenario. Similarly, 23 out of 100 APEX simulations using the 70 ppb ozone scenario were less than the maximum value estimated by the 100 APEX simulations using the 65 ppb ozone scenario (Smith and Glasgow 2015).

Second, for many age groups and ozone concentrations the density plots indicate a relatively high probability of risk estimates at or near zero, particularly for the multi-day decrements. This indicates that when statistical uncertainty is considered, one cannot rule out the possibility that the percentage of individuals in these categories experiencing a lung function decrement may be too small to be deemed to represent a public health concern.

[20] We calculated these distributions by first taking 100 draws from the multivariate normal distribution defined by the MSS model coefficients and variance-covariance matrix (Krinsky and Robb 1986). We then calculated the MSS-based lung function risk estimates in APEX for each of these draws and created density plots based on the range of risk estimates resulting from this calculation. Figure 4 presents these density plots.

Figure 4. Statistical Errors Around Point Estimates of Lung Function Risk Using MSS-2012 (for Atlanta, for dFEV1≥15%, 2010 Air Quality Data)
(Source: Figure 5 of Smith and Glasgow, 2015)



Although a method for summarizing the statistical uncertainty in the MSS-based risk estimates exists, and was even cited, the Draft PA does not report any statistical error bounds on these estimates. In this comment period, we have not had time to develop the comparable statistical error bounds for the new APEX results using the MSS-2013 model and the new APEX inputs. However, based on our prior (unpublished) replications of both MSS-2012 and MSS-2013, we are familiar with the attributes of the respective variance-covariance matrices that are used to calculate statistical errors as described in Smith and Glasgow (2015) and Glasgow and Smith (2017). Based on this information, we have no reason to expect that the statistical errors around the MSS-2013 risk estimates will be proportionately different from those presented in Figure 4 and in Smith and Glasgow (2015).

Finally, we want to note our disagreement with the Draft PA's further statement about our findings. In the same qualitative discussion of MSS uncertainties (at p. 3D-145), the Draft PA appears to minimize our reported findings of statistical error, stating:

> *"While the bounds were wide ranging ..., their reported median risk estimate (18.1%) is similar to that estimated in the 2014 HREA. Note, these central tendency risk values are based on using the best estimates of the MSS model coefficients and likely have the least amount of uncertainty."*

We disagree with this statement for two reasons. First, the fact that the median risk estimate in Glasgow and Smith (2017) is similar to that estimated in the 2014 HREA should not be interpreted as evidence that the HREA risk estimate is correct. The distribution of risk estimates described in Glasgow and Smith (2017) was created by first taking a large number of random draws from the multivariate normal distribution defined by the MSS model coefficients and variance-covariance matrix. The mean of these random draws will approximately equal the original MSS model coefficients and, thus, the median of the distribution of risk estimates produced using these draws will be very close to the original HREA risk estimate by design. This is not an independent confirmation of the risk estimates presented in the HREA.

Second, it is not sensible to say that risk values based on the "best estimates" of the MSS model coefficients likely have the least amount of uncertainty. The goal in estimating the MSS model coefficients is to estimate the true MSS model parameters in the population. However, the MSS model coefficients were estimated using a sample of clinical observations; a different sample would have led to different "best estimates" of the MSS model. That is, the estimated coefficients are the best estimates for the particular sample used in estimating the model, but this does not guarantee that they are accurate estimates of the true underlying population parameters. It is this type of statistical uncertainty that is reflected in the standard errors on the coefficients, and that ultimately led to the range of risk estimates in Glasgow and Smith (2017).

2.3. Uncertainties Within the E-R Approach Alone

We have noted above that statistical errors have not been reported for the MSS-based dFEV1 estimates and explained our reasons for inferring that those estimates have very wide confidence ranges. It is also the case that the Draft PA does not provide statistical error bounds on its E-R estimates. This omission is less justified, as EPA has the means to easily calculate them when it calculates the point estimates.

NERA obtained the code that the Agency used to compute the E-R point estimates and their respective 95% confidence ranges. We did this for several of the cities for the dFEV1 estimates for children reported in Table 3D-37 of the Draft PA.[21] We summarize the widths of those confidence intervals relative to the point estimates in Table 4 below, and the associated absolute estimates in Table 5. They are found to be very wide, particularly for the larger dFEV1 reduction estimates:

[21] Although we provide confidence ranges for percent of all children affected, the point estimates for percent of asthmatic children affected are nearly identical throughout the Draft PA. This is because the APEX simulations do not make different assumptions about the activity and microenvironment probabilities of asthmatic versus non-asthmatic children. To the extent that asthmatic children might be less active or stay in lower-ozone environments, one would expect fewer of them to experience the various benchmark exceedances. All else equal, one would also expect lung function occurrences to also be lower than the percent for all children, but uncertainty about their responsiveness to a given ozone exposure level may offset that behavioral effect. Nevertheless, in these comments we present only percent impacts to all children, and the reader should infer that approximately the same percent impacts will be predicted for asthmatic children in the Draft PA.

- For dFEV1≥10%, the E-R point estimates of percent of children have a 95% confidence range that is about 35% to 43% lower than the median point estimates, and about 63% to 65% higher than the point estimates. Stated as percentages of children affected, Table 5 shows median estimates in Table 3D-47 that are roughly 2.5%, as compared to 95% confidence ranges from roughly 1.5% to 4%.

- For dFEV1≥20%, the E-R point estimates of percent of children experiencing at least one occurrence per year have 95% statistical error ranges that are as much as 84% lower than the reported point estimate, and as much as 207% higher. While the upward side of the range is very large in relative terms, the absolute estimates for this decrement are very small, as shown in Table 5: only about 0.2% to 0.3% of children are projected to have such occurrences at the median and, thus, the upper end of the confidence interval for dFEV1≥20% still is projected to occur for less than 1% of all children.

- The asymmetry of the confidence ranges is driven by the fact that the lower bound of the risk relationship cannot extend below zero. For 7-hour average exposures (while under moderate exertion) of less than 40 ppb, there appears to be a discrete amount of probability associated with zero responses. This implies that the mean estimates of E-R-based risk are lower than the median estimates that are presented in the Draft PA.

Table 4. 95% Confidence Bounds on E-R Models' Estimates of dFEV1 Reductions, Stated as Percent Differences from the Point Estimates in Table 3D-27 (for dFEV1 exceedances at least once/year, for selected cities and years, 70 ppb)

		dFEV1≥10%		dFEV1≥15%		dFEV1≥20%	
City	**Year**	**2.5% CI**	**97.5% CI**	**2.5% CI**	**97.5% CI**	**2.5% CI**	**97.5% CI**
Atlanta	2016	-38%	+63%	-72%	+64%	-69%	+189%
Dallas	2015	-37%	+63%	-70%	+63%	-69%	+201%
Dallas	2016	-43%	+65%	-85%	+70%	-83%	+191%
Dallas	2017	-39%	+64%	-74%	+66%	-73%	+203%
Philadelphia	2015	-40%	+64%	-78%	+67%	-77%	+198%
Phoenix	2015	-35%	+65%	-73%	+63%	-73%	+207%
Sacramento	2015	-43%	+65%	-86%	+71%	-84%	+195%

Table 5. 95% Confidence Bounds on E-R Models' Estimates of dFEV1 Reductions, Stated as Absolute Estimates (Percent of Children) (for dFEV1 exceedances at least once/year, for selected cities and years, 70 ppb)

		dFEV1≥10%			dFEV1≥15%			dFEV1≥20%		
City	**Year**	2.5% CI	Median	97.5% CI	2.5% CI	Median	97.5% CI	2.5% CI	Median	97.5% CI
Atlanta	2016	1.5%	2.5%	4.0%	0.2%	0.6%	1.1%	0.1%	0.2%	0.7%
Dallas	2015	1.6%	2.6%	4.3%	0.2%	0.7%	1.1%	0.1%	0.3%	0.8%
Dallas	2016	1.2%	2.1%	3.5%	0.1%	0.5%	0.8%	0.0%	0.2%	0.5%
Dallas	2017	1.5%	2.5%	4.1%	0.2%	0.6%	1.0%	0.1%	0.2%	0.7%
Philadelphia	2015	1.4%	2.3%	3.8%	0.1%	0.6%	1.0%	0.0%	0.2%	0.6%
Phoenix	2015	2.0%	3.0%	5.0%	0.2%	0.8%	1.3%	0.1%	0.3%	0.9%
Sacramento	2015	1.2%	2.1%	3.5%	0.1%	0.5%	0.9%	0.0%	0.2%	0.5%

Thus, statistical errors are quite significant for the E-R model results, as is the case for the MSS model results. The statistical error appears to be larger than the predicted changes in median lung function risk associated with different air quality scenarios. For example, the city-wide average for the current standard (in Table 3D-37) for dFEV1≥10% is 2.4%, while for a 75 ppb design value it is 3.0% (Table 3D-43) and for a 65 ppb design value it is 1.9% (Table 3D-46). These changes in the dFEV1≥10% are smaller in percentage terms than the 95% confidence range around the 2.4% under the current standard.

As always, model uncertainty is an important consideration that often exceeds statistical errors. We have less ability to assess model uncertainty associated with the E-R model under current time constraints. However, we note two points in this regard.

1. There is model uncertainty regarding the appropriate length of the averaging period for assessing lung function decrement risks from a given ozone exposure. In the 2014 HREA, the E-R model was run using the daily maximum 8-hour exposure with moderate exertion. In the Draft PA, the same model (same set of model parameters) was applied to a 7-hour rather than 8-hour exposure. We have run the E-R model with the 8-hour exposure average, too. We find that use of the 8-hour exposure window reduces the percent of affected children by about 20% for dFEV1≥10%, and by as much as 50% for dFEV1≥20% (based on at least one occurrence per year, for children, for the 70 ppb standard). This also indicates that the effect of deciding to use a 7-hour averaging window for the E-R model for this review's risk calculations would have *increased* the E-R lung function risk estimates compared to this in the 2014 HREA if there had been no other APEX input assumption changes.

2. The Final IRP for this review indicated that the Agency was planning to re-estimate the E-R model using a probit relationship, rather than the 90/10 logistic/linear fit that it used in the 2014 HREA. The reason EPA gave for shifting to a probit fit was because "using the logistic fit in E-R functions may overestimate the contribution of risk attributed to low O_3 exposure levels."[22] The Draft PA has actually used the same logistic fit that the Agency used in the 2014 HREA and is silent on the question of the probit fit in the plan. This leaves an unanswered question about whether the current E-R estimates are perhaps overstated relative to a potential alternative way of specifying the E-R model. Clearly, there is model error yet to be quantified, the magnitude of which warrants better understanding. It would have been helpful if the Draft PA had explained why the probit fit was not undertaken, and to have provided more explanation of how a probit-based fit, if completed, might have altered the direction of the risk results.

[22] See Final IRP, p. 5-20.

3. Uncertainties in Epidemiological Risk Calculations

As noted in Section 1 of these comments, the Draft PA does not provide any epidemiologic-based risk calculations. In recounting the basis for setting the current standard in 2015, the Draft PA states that epidemiologic-based risk estimates were given the least weight; at most, they were considered to be supporting evidence that risks at the then-current standard of 75 ppb could be meaningfully reduced by revising the standard to 70 ppb.[23] The reason the epidemiologic-based risk estimates received such little weight in the last review is tied to epistemic ("model") uncertainties.[24] The Draft PA reports that model uncertainties associated with estimating quantitative risks from epidemiologic associations have not been reduced since the 2015 decision, and explains that the Agency focused its efforts on preparing revised estimates of the exposure and lung function risks that received greater weight in the prior decision. In this section, we discuss the key epistemic uncertainties associated with epidemiologic-based risk calculations and focus on one important example of how epistemic uncertainties could have been more directly quantified in the 2014 HREA's mortality risk calculations.

3.1. Epistemic Uncertainties as a General Matter

The Draft PA provides a reasonably complete overview of the key forms of epistemic uncertainty of the standard BenMAP-based approach that EPA has traditionally used for calculating epidemiologic-based risk estimates. The most salient of these are:

- **Exposure measurement errors**. Measurement error is most commonly discussed in terms of having to use central monitors or model-based exposure estimates to predict what people are actually experiencing. Such errors are known to bias estimates of the slope of the concentration-response (C-R) association, and to hinder ability to detect the shape of a potential true C-R that may be indicated to exist by epidemiologic associations. However, the Draft PA points to another aspect of this uncertainty that is far more problematic for using associations to predict actual levels of risk or changes in risk under real-world air quality scenarios. This is the inability to know the correct temporal window of the effective exposure to use when estimating a C-R relationship. Use of an incorrect exposure window may still detect an association if it exists (this requires only that the correct exposures have been correlated with the used exposure estimates). However, if the levels of pollutant in the exposure window that is used differ from the levels in the (unknown) correct exposure window, the *quantitative* estimates of relative risk will be incorrect, and so too will be quantitative estimates of population risks due to that pollutant and of potential changes in risk levels if the pollutant concentrations were to be changed. Uncertainty about the exposure window is of particular concern for estimates of long-term health risk and is discussed further below in that context.

- **Co-pollutant effects**. When more than one pollutant is having an effect on health, it is possible that the quantitative estimate of relative risk for any individual pollutant can be biased upwards or downwards by the effect of other pollutants if they are correlated with each other in time and space. In fact, even the existence of an association can be falsely attributed to the wrong pollutant in this situation. The Draft PA explains that enhanced efforts to test multi-pollutant models improve our confidence that observed associations of health endpoints with ozone are not spurious. However, this enhanced confidence does not extend to the *quantitative* interpretation of the relative risk. This is because differences in the measurement errors for the various pollutants make it impossible to untangle with any confidence *how much* of each association is due specifically to changes in each pollutant. This, in turn, adds to the epistemic uncertainty

[23] Draft PA, p. 3-16.

[24] Draft PA, p. 3-45.

associated with quantitative risk calculations based on these epidemiologic estimates of a C-R function.

There is no indication that EPA has developed any uncertainty analysis tools to better integrate and quantify the impact of such epistemic uncertainties into its epidemiologically-based risk calculation methods. As we have noted in multiple papers, EPA's BenMAP model is not equipped for integrated uncertainty analysis and the policy guidance provided by risk calculations that omit quantitative epistemic uncertainty analysis is unreliable. Attached as an Appendix to these comments is a copy of Dr. Smith's October 2019 comments to CASAC on the external review draft the Policy Assessment for Particulate Matter (*i.e.*, on U.S. EPA, 2019a). The October 2019 comments summarize our research on this topic, most of which is relevant to ozone risk analyses, as well.

Given that EPA has not yet developed integrated uncertainty analysis tools, we conclude that it was a sound decision not to proceed with such risk calculations in the Draft PA. Any epidemiologic risk estimates calculated in the deterministic manner of the 2014 HREA (and in the recent Draft PA for Particulate Matter) would not provide reliable guidance to decision makers. It would only have cluttered the Draft PA with content that was deemed unworthy of any meaningful weight in 2015 ozone NAAQS decision, and would have no improved usefulness now.

3.2. Long-Term Respiratory Mortality Uncertainty Specifically

We noted above that epistemic uncertainties regarding the shape of an epidemiologic-based C-R function can be very large. A specific example of this arose in the 2014 HREA with respect to risk of respiratory mortality from long-term ozone exposure. In that HREA, a new epidemiologic study by Jerrett *et al.* (2009) was adopted for quantitative estimates of respiratory mortality risks. The draft of that HREA applied a linear/no-threshold version of the C-R association in that model. Quite remarkably, this C-R assumption produced quantitative risk estimates that 18.5% of all U.S. respiratory deaths were hastened under the then-recent (*i.e.*, 2007)) ozone levels.[25] However, that draft HREA had ignored results from several alternative models reported in Jerrett *et al.* that indicated the presence of a threshold in mortality risk at a level near the mean of all the ozone observations.[26] The best-fitting model had a threshold at 56 ppb. When the public comments revealed that the risk results were extremely sensitive to the alternative models provided in the paper, the final HREA included sensitivity cases across all of the potential models reported in the paper. The best-fitting model's risk estimates were that 0.7%, not 18.5%, of respiratory mortality was due to ozone conditions.[27]

By *quantifying* this one form of epistemic uncertainty, it became more apparent to decision makers just how unreliable the risk estimates are from such an epidemiologic extrapolation. Those sensitivity analysis results are emblematic of the reasons why epidemiologic-based risk estimates were given little weight in the final decision. However, we note that the 2014 HREA still implied that the risk analysis needed to choose a single model from the many in the paper, and it continued to emphasize the no-threshold model simply because the better-fitting threshold model could not be shown to be *statistically-significantly* better than the no-threshold model under *every* statistical test that could be identified. We noted that when the uncertainty in the risk estimate is so large, it is inappropriate to eliminate this uncertainty by selecting a single model with one deterministic threshold assumption. In Smith and Glasgow (2018), we demonstrated how an integrated uncertainty analysis could be conducted that would provide a probability distribution over this risk estimate that accounted for the relative likelihood that any

[25] 2014 HREA, p. 8-9.

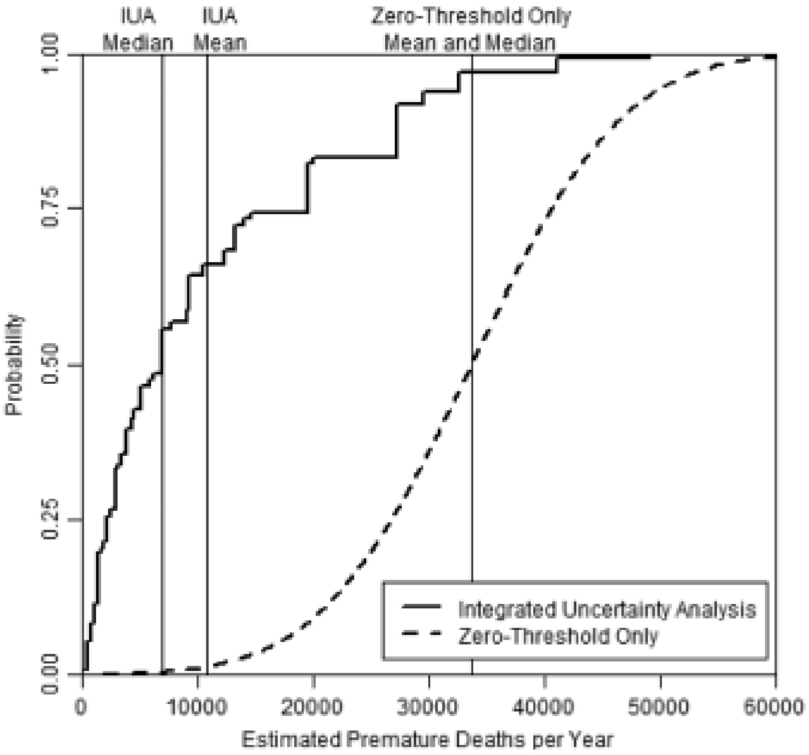
[26] The ozone exposures in this study were the average of the daily 1-hour maxima over the period 1977-2000, which was the general time frame of the cohort follow-up period studied, 1982-2000.

[27] 2014 HREA, p. 8-19. (Divide 1,600 deaths by 240,000 total respiratory deaths reported on p. 8-9).

of the many models in Jerrett *et al.* would be correct. The integrated uncertainty distribution over the respiratory mortality estimates would replace the many competing sensitivity cases with a single probability distribution. It was also shown that this method could generate much different policy guidance than might be inferred from the core risk estimate in the 2014 HREA.

An example of the results is provided in Figure 5 below, which is a copy of Figure 2 from our paper. The figure shows that a probabilistic integration of the many alternative models in Jerrett *et al.* results in mean and median mortality risk estimates far below those from the HREA's deterministic choice. While the confidence ranges are wide on both estimates, the integrated uncertainty probability distribution indicates a 97% probability that the risk is lower than even the mean (*i.e.*, EPA's point estimate) based on the linear/no-threshold model.[28,29]

Figure 5. Cumulative Probability Distribution over Estimated 2007 U.S. Premature Respiratory Deaths Due to Ozone Exposures: Results of Integrated Uncertainty Analysis Compared to Core HREA Estimate Model Assuming No Threshold (Source: Smith and Glasgow, 2018, Figure 2)



We also note that a more recent paper (Turner *et al.*, 2016) provides a similar analysis of long-term respiratory mortality risk to that in Jerrett *et al.* The Turner *et al.* paper uses the same cohort (the American Cancer Society cohort) but makes two changes. First, it extends the cohort follow-up period by

[28] Smith and Glasgow, 2018, p. 170.

[29] While the probability distribution developed in this illustrative example of epistemic uncertainty analysis indicates a 100% chance that there is some non-zero risk, this is because the entire analysis was predicated on a *presumption* that the Jerrett *et al.* associations are causal. Integration of uncertainty on whether there is a causal relationship underlying that paper's findings would further alter the distributions.

four more years (to 2004) – that is, it contains mortality observations for 22 years compared to 18 years in the Jerrett *et al.* database. Second, it uses modeled ozone exposures rather than the average of monitored values that Jerrett *et al.* used. This allowed the cohort covered to be expanded by about 50%, because more locations could be assigned ozone exposures when using modeled rather than monitored values.[30] The expansion of coverage increased deaths through 2000 by about 50%, while the addition of four years of follow up (for both the original and expanded sets of individuals) resulted in an approximate doubling of the number of deaths observed, with attending statistical power benefits.[31]

However, the shift to modeled estimates also created an important limitation to the *quantitative* usefulness of the new Turner *et al.* results. In shifting to modeled ozone exposures, the authors also shifted to assigning ozone exposure estimates from the years 2002-2004. This is an entirely different exposure window, and almost certainly not as appropriate of a window as Jerrett *et al.* used (*i.e.*, 1977-2000). First, Turner *et al.*'s ozone exposure assumptions post-date about 75% of the deaths that they are being used to explain, by somewhere between 2 and 20 years.[32] Given that ozone concentrations were also falling rapidly after 2000, Turner *et al.*'s assumed exposures are also likely substantially lower than the exposures that were occurring before most of the deaths in their data set.

This is a classic example of uncertainty due to exposure window choice that we discussed in Section 3.1. While there is some degree of uncertainty associated with the choice of exposure window in Jerrett *et al.*, those exposures are at least contemporaneous with the follow up period. The choice of exposure window is a major flaw for the Turner *et al.* study and afflicts the reliability of any use of the Turner *et al.* results for quantification of long-term respiratory mortality risk, as we explain below.

Turner *et al.* report that their mean ozone exposure level is 38.2 ppb, with a range from 26.7 ppb to 59.3 ppb. Again, these are estimates of the average during 2002-2004. Jerrett *et al.* report that their mean ozone exposure is 57.7 ppb, with a range of 33.3 ppb to 104 ppb -- average values during 1977-2000. Thus, ozone exposures averaged about 50% higher during the period when most of the death occurred in the Turner *et al.* study (the highest exposure may have been as much as 75% higher). If the relative risks for a given set of deaths are regressed against two alternative exposure assumptions, with the first being across the board higher than the second, the estimated relative risk of the second is higher than that of the first.[33] Thus, relative risks from the Turner *et al.* paper, because they reflect an incorrect exposure window with understated ozone concentration for the deaths that are being explained in that cohort, are inappropriately biased upwards.

Another point that should be noted is that Turner *et al.* continue to find evidence of an effect threshold. For example, Figure E3 of the Turner *et al.* online data supplement provides a spline estimate of the shape of their ozone C-R estimate. The relative risk declines from the minimum exposure until about 35 ppb and remains lower than risks associated with the minimum exposure until about 40 ppb. The authors

[30] That is, Jerrett *et al.*'s cohort included 449,000 individuals while Turner *et al.*'s was expanded by 220,000 to a total of 669,000 individuals

[31] That is, Jerrett *et al.*'s cohort had a 26.5% cumulative mortality through 2000 (for 119,000 deaths) while Turner *et al.*'s cohort had 35.4% cumulative mortality by the end of 2004 (for 237,000 deaths). Of the 237,000 total deaths in the latter case, we estimate that 177,000 occurred through 2000 (including the 119,000 in the Jerrett cohort plus about 58,000 among the 220,000 additional individuals in additional locations.) Thus, we estimate that 75% (177000/237000) of the deaths in the Turner et al, cohort occurred before 2000.

[32] They post-date *all* of the deaths that were also analyzed by Jerrett *et al.*, the last of which were in 2000, and which represent 119,000 of the 237,000 that Turner *et al.* have analyzed – fully 50% of Turner's observations. Considering the mortality rates in the evidence, we estimate that about 75% of all the deaths in Turner's expanded cohort occurred before 2000.

[33] That is, in the second case, the same differences in death rates would be "explained" by a smaller amount of difference in assumed exposure, so that the estimate of the change in risk *per increment* of the pollutant (*i.e.*, the relative risk) would be a larger number.

report that they found evidence of a threshold at 35 ppb.[34] This may sound lower than the best-fitting threshold reported in Jerrett *et al.*, which was at 56 ppb.[35] However, these appear to be the same finding: 35 ppb is just below the mean of the Turner *et al.* ozone data (*i.e.*, 38.2 ppb), and 56 ppb is just below the mean of the Jerrett *et al.* ozone data (*i.e.*, 57.7 ppb). That is, these thresholds occur in the same portion of the distribution of cohort exposures. It is also important to note that in both cases, the evidence for a threshold is approximately in the middle of the range of observations – it is not as if there is an anomaly present where the data are sparse. This threshold must be attributable to the 75% of the deaths that occurred before 2000 in Turner *et al.*'s data set (about two-thirds of which comprise the deaths in Jerrett *et al.*'s data set). Thus, Turner *et al.*'s analysis reaffirms the evidence for a threshold of effects for long-term respiratory mortality at a level that was about 56 ppb, when measured in terms that are generally contemporaneous with the actual deaths.

[34] Turner *et al.*, 2016, p. 1140.

[35] Jerrett *et al.*, 2009, Supplementary Appendix, Table 3S.

4. Conclusions

Based on our review of the Draft PA and its underlying data and models:

- We concur with the Agency's finding that the exposure and associated health risk estimates under the current standard of 70 ppb are similar to those estimated when that standard was set in 2015.
- We also concur that the uncertainties associated with quantitative risk estimates have not significantly changed since the 2015 review.

With regard to uncertainties in modeling health risks, we also concur with the following methodological decisions in the Draft PA's quantitative risk assessment:

- It is not unreasonable for the Draft PA to give relatively greater weight to lung function risk estimates based on the E-R model over the MSS model. Although we find that estimates from both approaches are subject to large statistical error ranges and a variety of model uncertainties, the Draft PA provides clear evidence that the MSS-based model estimates are affected by extrapolation beyond the base of scientific evidence more than those from the E-R model.
- We conclude that the Agency has made a reasonable judgment not to have conducted any epidemiologic-based risk calculations, because we have found no evidence that the Agency has yet developed appropriate tools for quantifying the role of the very extensive epistemic uncertainties in those types of risk calculations.

References

Glasgow G, Smith AE (2017). Uncertainty in the estimated risk of lung function decrements owing to ozone exposure. J Expo Sci Environ Epidemiol 27(5): 535-538.

Jerrett, M, Burnett, RT, Pope, CA, 3rd, Ito, K, Thurston, G, Krewski, D, Shi, Y, Calle, E and Thun, M (2009). Long-term ozone exposure and mortality. N Engl J Med 360(11): 1085- 14 1095.

Krinsky I, Robb AL (1986). On approximating the statistical properties of elasticities. Rev Econ Stat 68: 715-719.

McDonnell, WF, Stewart, PW, Smith, MV, Kim, CS and Schelegle, ES (2012). Prediction of lung function response for populations exposed to a wide range of ozone conditions. Inhalation Toxicology 24(10): 619-633.

McDonnell, WF, Stewart, PW and Smith, MV (2013). Ozone exposure-response model for lung function changes: an alternate variability structure. Inhalation Toxicology 25(6): 348-353.

Smith, AE, Glasgow G (2015). Quantification of uncertainty in EPA's estimates of lung function impacts from ozone exposure." Report prepared for American Petroleum Institute, Washington DC. March. (submitted to 2015 ozone NAAQS docket with API comments).

Smith AE, Glasgow G (2018). Integrated uncertainty analysis for ambient pollutant health risk assessment: A case study of ozone mortality risk. Risk Analysis 38(1): 163-176.

Turner, MC, Jerrett, M, Pope, CA, 3rd, Krewski, D, Gapstur, SM, Diver, WR, Beckerman, BS, Marshall, JD, Su, J, Crouse, DL and Burnett, RT (2016). Long-term ozone exposure and mortality in a large prospective study. American Journal of Respiratory and Critical Care Medicine 193(10): 1134-1142.

U.S. EPA. (2014). Health risk and exposure assessment for ozone. (Final report). Office of Air Quality Planning and Standards, Research Triangle Park, NC. EPA-452/R-14-004a. August.

U.S. EPA. (2019a). Policy assessment for the review of the ozone National Ambient Air Quality Standards, external review draft. Office of Air Quality Planning and Standards, Research Triangle Park, NC. EPA-452/P-19-002. October.

U.S. EPA. (2019b). Integrated review plan for the ozone National Ambient Air Quality Standards. Office of Air Quality Planning and Standards, Research Triangle Park, NC and National Center for Environmental Assessment, Research Triangle Park, NC. EPA-452/R-19-002. August.

APPENDIX

Copy of oral and written comments of Dr. Anne E. Smith submitted to CASAC on the Policy Assessment for Particulate Matter, October, 2019.

Oral Statement to CASAC on the *Policy Assessment for the Review of the National Ambient Air Quality Standards for Particulate Matter, External Review Draft*

Anne E. Smith, Ph.D.
Managing Director, NERA Economic Consulting
October 22, 2019

My name is Dr. Anne Smith. My comments here (and that I have provided in more detail in writing to CASAC) are on the quantitative risk analysis in Section 3.3 and Appendix C of the draft PA.

I have a long track record of research on methods of incorporating uncertainty into quantitative risk analyses, and I have authored 7 articles published in the past 5 years specifically focused on how EPA conducts its risk analyses for criteria pollutants like PM2.5. These papers have not been cited in the draft PA.

Based on these, I conclude that the draft PA's risk analysis fails to provide useful or reliable information to support the science-policy judgment that the Administrator must make for the PM NAAQS.

This is because it fails to incorporate the most important types of uncertainty that affect its calculations.

EPA's risk analysis *does* provide what appear to be uncertainty ranges around its risk estimates, but each range actually reflects only one minor form of uncertainty: noise in the data used by epidemiologists to statistically estimate evidence of a PM-health association. This is called "statistical error."

There are a host of other far larger sources of uncertainty that are not at all reflected in those risk ranges. These missing sources of uncertainty are variously called "model uncertainty," "epistemic uncertainty," or "scientific" uncertainty. They arise because *the essence of risk analysis is extrapolation*.

What do I mean by "extrapolation"? An epidemiological study attempts to infer statistically what *did* happen under one set of circumstances, and a risk analysis then attempts to predict what *would* happen under different circumstances. Even statistically-significant epidemiological associations may not be reliable for the kinds of extrapolations that a risk analyst must make. At best, they can serve as a starting point for the risk analyst's deliberations.

This distinction is recognized in the risk analysis profession's paradigm, which identifies four separate structural elements of a sound risk analysis: (1) hazard identification, (2) exposure assessment, (3) dose-response assessment and (4) risk characterization. Risk characterization, which combines information from the exposure and dose-response assessment steps, is the subject of Section 3.3 and Appendix C.

Note, however, the clear distinction in this paradigm between hazard identification and dose-response assessment.

Within this paradigm, epidemiological studies are, first and foremost, for assessing whether a hazard *exists* from exposure to ambient $PM_{2.5}$. Unfortunately, the Agency does not clearly recognize the additional evaluations that are required when shifting from that hazard identification step to the dose-response assessment step. The Agency's analysts simplistically use coefficients from the epidemiological studies *at face value* as a deterministic dose-response formula. This is not sufficient for an informative characterization of risks.

In my own analyses, I have shown that epistemic uncertainties affecting PM2.5 risk calculations do NOT just widen the range about the risk estimate. Rather, they asymmetrically add more probability on the lower rather than higher side of the risk estimate. This is the sort of insight that is of relevance to a decision maker.

Is there a better method to address epistemic uncertainties in the risk analysis than used in the PA? Yes. It is called "integrated uncertainty analysis". One of my papers recounts how the Agency realized over 40 years ago that integrated uncertainty analysis was needed to support a well-informed NAAQS decision. *And that it would entail using subjective judgments about the epistemic uncertainties*.

It's also important for CASAC to recognize that integrated uncertainty analysis was used in the first PM2.5 HREA back in 1997. It can be done!

One of my papers shows how EPA's elimination of such uncertainty from its next 2 PM HREAs rendered them irrelevant to the next 2 PM NAAQS decision makers. The same fate awaits the risk analysis in the draft PA.

Is the BenMAP tool part of the problem? BenMAP cannot perform integrated uncertainty analysis. Instead, BenMAP focuses all of its energy on doing deterministic risk calculations at a level of geographic and demographic disaggregation that is far out of line with our present degree of knowledge about the true risk relationships.

I have submitted written comments that expand on these points. They also summarize each of my 7 recent articles and explain how they stand as a group in support of my conclusions.

Comments to CASAC on the *Policy Assessment for the Review of the National Ambient Air Quality Standards for Particulate Matter, External Review Draft*

Anne E. Smith, Ph.D.
Managing Director, NERA Economic Consulting
October 21, 2019

On September 5, 2019, the U.S. Environmental Protection Agency (EPA, or "the Agency") released its *Policy Assessment for the Review of the National Ambient Air Quality Standards for Particulate Matter, External Review Draft* (hereafter referred to as the "draft PA").[1] On October 24-25¸ the Clean Air Scientific Advisory Committee (CASAC) will be convening to develop its comments to EPA on the draft PA. The following document is intended to provide additional information from my own research that is of relevance and potential interest to the CASAC, as well as to the Agency. I have prepared these comments with financial support from a coalition of industry associations.

My comments are focused specifically on the question of *whether the section of the draft PA that discusses <u>risk-based considerations</u> provides information that is useful and reliable as guidance to decision makers when considering the merits of the current and alternative potential National Ambient Air Quality Standards (NAAQS).*[2]

This is a research topic on which I have been actively engaged for the past thirty-five years, starting with a study funded by EPA in the early 1980s to explore methods for incorporating uncertainties into NAAQS-focused health risk analyses. (I conducted that study while working in the Decision Analysis Group at SRI International.) I have continued to be actively engaged in developing and demonstrating methods for air pollutant health risk analysis, including preparing technical comments on the risk analyses for many prior NAAQS reviews of $PM_{2.5}$, ozone, SO_2, and NO_2.

Of most relevance to the present deliberations is a set of seven articles that I have written (and have had published in the period 2015-2019) on criteria pollutant heath risk analysis methods. These papers have addressed both the Health Risk and Exposure Assessment (HREA) documents that have been prepared to support in the Administrators' decisions on NAAQS revisions and the risk and benefits estimates that appear in the Regulatory Impact Analyses (RIAs) that must accompany each new major proposed and final rulemaking, such as most NAAQS revisions. This $PM_{2.5}$ review cycle has not produced a separate HREA document, but Section 3.3 of the draft PA contains the same kind of information that is usually first presented in an HREA. For that reason, any reference to methods appropriate for HREAs in my comments below should be viewed as a relevant comment for the contents of Section 3.3 of the draft PA.

My objective in this document is to provide a brief synthesis of the content and conclusions of my articles in the specific context of the current review of the $PM_{2.5}$ primary NAAQS. I provide full references to the

[1] EPA, *Policy Assessment for the Review of the National Ambient Air Quality Standards for Particulate Matter, External Review Draft*, EPA-452/P-19-001, Office of Air Quality Planning and Standards, Research Triangle Park, NC, September 2019.

[2] This discussion is based primarily on Sections 3.3 and Appendix C of the draft PA. It does, of course, also tie into some of the broader questions regarding the causal interpretation of the concentration-response (C-R) functions that EPA has inferred from $PM_{2.5}$ epidemiological studies. Although my focus herein is on the risk analysis in this draft PA, I have previously (in 2009) commented on the Agency's causality framework. Because I consider those comments to be of continued relevance, I am attaching a copy of them in an appendix to this document.

articles in the bullet list below.[3] I list the papers in the order in which I discuss them (rather than by publication date), along with a short name that I will use in the rest of this document to refer to each. Those short names are shown in bold font:

- "Setting Air Quality Standards for $PM_{2.5}$: A Role for Subjective Uncertainty in NAAQS Quantitative Risk Assessments?" *Risk Analysis*, Vol. 38(11), November 2018, pp. 2318-2339. **("HREA History")**

- "Author Synthesis and Response" to Special Issue on Air Pollution Health Risks, *Risk Analysis* Vol. 36(9), September 2016, pp. 1780-1782. **("Special Issue Synthesis")**

- "Enhancing the Characterization of Epistemic Uncertainties in $PM_{2.5}$ Risk Analyses" (with W. Gans), *Risk Analysis*, Vol. 35(3), March 2015, pp. 361-378. **("BenMAP Review")**

- "Response to Commentary by Fann *et al.* on "Enhancing the Characterization of Epistemic Uncertainties in $PM_{2.5}$ Risk Analyses," *Risk Analysis*, Vol. 35(3), March 2015, pp. 381-384. **("Response to EPA")**

- "Integrated Uncertainty Analysis for Ambient Pollutant Health Risk Assessment: A Case Study of Ozone Mortality Risk" (with Garrett Glasgow), *Risk Analysis*, Vol. 38(1), January 2018, pp. 63-176. **("Ozone IUA")**

- "Inconsistencies in Risk Analyses for Ambient Air Pollutant Regulations," *Risk Analysis*, Vol. 36(9), September 2016, pp. 1737-1744. **("Inconsistencies")**

- "Using Uncertainty Analysis to Improve Consistency in Regulatory Assessments Of Criteria Pollutant Standards" (accepted by *Risk Analysis* in September 2019, in press). **("Improving Consistency")**

1. Summary of Key Points

The articles listed above cover several different aspects of issues with and methods for health risk analysis for criteria pollutants. However, each of the articles, and all of them as a group, support my conclusion here that *the risk analysis presented in the draft PA provides insufficient and even misleading guidance for the upcoming $PM_{2.5}$ NAAQS decision*.

The central flaw of this draft PA's risk analysis is its lack of any attempt to quantify or integrate the most important sources of uncertainty inherent in the risk calculations—uncertainties that are variously referred to as "model," "epistemic," or "scientific" uncertainty. Instead, the range around each of the draft PA's reported risk estimates reflects only the statistical error of the respective single concentration-response (C-R) function slope coefficient assumed for that particular risk calculation. That statistical error reflects only the noise in the data of that one epidemiological study; it bears no relationship to additional potential errors in the C-R slope estimate that may be due to misspecification of the C-R shape to be estimated, the manner in which covariates are assumed to influence the PM-risk relationship, potential exposure measurement errors,[4] or potentially different levels of exposure errors for co-pollutants. These latter

[3] Under my copyright agreements, I can send copies of these papers to individuals who request them of me. Such requests can be emailed to me at Anne.Smith@NERA.com.

[4] Exposure mismeasurement is often discussed in terms of whether the estimated ambient concentration of total $PM_{2.5}$ mass that is attributed to the individuals in a study is a correct estimate of their actual ambient total mass exposure. However, there are other important (perhaps more important) errors that fall in the category that are often ignored. One is whether the correct period for exposure is being used, including whether lagged, time-varying and/or cumulative exposures are more appropriate. Another form of mismeasurement error is whether total $PM_{2.5}$ mass is the correct exposure, or some specific $PM_{2.5}$ constituent

forms of potential biases and error have repeatedly been recognized as far larger than the statistical errors. In my own research analyzing these non-statistical uncertainties, I have routinely found that they do not simply widen the range about the risk estimate, but asymmetrically add more probability on the lower rather than higher side of the risk estimate.

Layered on top of the aforementioned host of unquantified model uncertainties is a concern with prediction error.[5] An epidemiological estimate of the relative risk based on a given set of observations—even if highly accurate as an estimate of the average responses of the individuals within that data set that are causally attributable to $PM_{2.5}$—does not necessarily accurately predict what the average population response will be in a future year, at a different future level of exposure to $PM_{2.5}$, or in different locations. However, such extrapolation is the essence of risk analyses based on epidemiological associations: the epidemiological estimation attempts to infer statistically what *did* happen under one set of circumstances, and the risk analysis then attempts to predict what *would* happen under different circumstances. The prediction error of the estimated C-R relationship can be large for individual observations within the original set of observations. However, it is vastly exacerbated when using an estimated C-R relationship for "out of sample" extrapolations to other conditions, which is the essence of the type of health risk analysis reported in Section 3.3 of the draft PA.

By using statistical error as the only measure of uncertainty in its risk estimates, Section 3.3 provides a highly misleading summary of risk-based considerations for decision makers. As written, it could sway decision makers and other readers into a mistaken belief that the ranges around its risk estimates represent a quantitative synopsis of the many uncertainties that the section discusses only qualitatively. The risk analysis in the draft PA is at odds with and completely ignores the available literature on the need for quantitative integrated assessment of the uncertainties that arise when using from epidemiologically-based relative-risk coefficients to predict absolute levels of risk under different circumstances. Lacking any attempt at providing an integrated uncertainty analysis (IUA), the draft PA does not provide the type of decision-relevant information that is core objective of good quantitative risk analysis.

2. The Progressively Declining Usefulness of Risk Analysis in PM NAAQS Decision-making

My November 2018 paper labelled **"HREA History"** tracks the use of and reliance on results from quantitative health risk analysis in the Agency's NAAQS decisions dating back to the 1970s. It first documents how the Agency realized over 40 years ago that integrated assessment of the epistemic uncertainties associated with available scientific evidence was critical to making a well-informed decision on what level of criteria pollutant exposures could be judged to be protective of the public health with an adequate margin of safety. This included recognition by the Agency that such uncertainties could only be fully characterized through the subjective judgments of experts who are able to assess the merits of existing scientific studies.

The **"HREA History"** paper then reviews the HREAs specifically associated with the $PM_{2.5}$ NAAQS, including those for the 1997, 2006, and 2013 $PM_{2.5}$ NAAQS decisions. It tracks how epistemic uncertainty was characterized in the past three $PM_{2.5}$ NAAQS HREAs, and it analyzes how informative each successive HREA was to the respective Administrators' final judgments on the appropriate NAAQS

that is a subset of the total mass. Serious errors in the C-R slope estimate that is then applied in a risk calculation can be created by either of these latter forms of measurement error, even if the researchers have very accurate information on what the concentration of total $PM_{2.5}$ mass was at a given individual's residence and point in time.

[5] Prediction error of an estimated relationship is not the same thing as statistical error of the slope estimate, which is the only form of statistical uncertainty that EPA uses in its risk analyses in the draft PA. The prediction error of a line fit through a set of observations is larger than the statistical error on the estimated slope of that line, and increasingly so for observations farther from the mean observation.

to establish. Consistent with the structure of the current draft PA, the three prior $PM_{2.5}$ NAAQS decisions considered an "evidence-based" approach and a "risk-based" approach.[6]

The paper describes a progressive elimination of quantitative characterizations of epistemic uncertainty in the $PM_{2.5}$ HREAs, and a concomitant elimination of relevance to the decision makers of the information provided by the HREAs. The research shows that these parallel trends are not a coincidence but are clearly causally related. That is, in each successive NAAQS review decision, it is shown that deterministic assumptions that eliminated consideration of critical epistemic uncertainties in the HREAs' risk estimates undercut the decision makers' ability to rely on those respective HREAs. As a result, in both the 2006 and 2013 rulemakings, the justification for the selected NAAQS level was founded entirely on the evidence-based approach, while the "risk-based considerations" reported in the HREA were ignored.

My article recounts how the first $PM_{2.5}$ HREA, finalized in 1997, contained a set of three alternative potential subjective probability distributions over the key scientific uncertainties in the risk calculations – the most important of which was the level below which the C-R functions might not continue.[7] Given the quantification of epistemic uncertainties in that HREA, it was possible to present a consistent discussion of evidence-based and risk-based considerations. The Administrator at the time was able to select NAAQS levels that were coherently related to the information provided in the HREA.

The second $PM_{2.5}$ HREA (for the rule finalized in 2006) did not include a probability distribution over the epistemic uncertainty on a potential end of the C-R relationship. Instead, following a recommendation by CASAC, that HREA assumed (deterministically) a threshold for long-term mortality risk specifically at 10 $\mu g/m^3$. This deterministic treatment of the most critical, decision-relevant uncertainty associated with the risk calculations had the effect of making 10 $\mu g/m^3$ the only reasonable level at which to set the long-term standard if one were to rely on that HREA's risk estimates. Observing this, the Administrator essentially dismissed that HREA as uninformative to the science-policy judgment that he faced.

In the third $PM_{2.5}$ HREA (for the current standard, promulgated in 2013), the HREA backed-off entirely from any quantitative treatment of scientific uncertainties in the risk calculations. It assumed the C-R functions, including that for long-term mortality, were of unchanging slope or uncertainty down to the the lowest measured level in each epidemiological dataset from which a C-R slope coefficient was obtained. (This was similar to assuming a deterministic threshold at 5.8 $\mu g/m^3$.) Anticipating the impending implications of that HREA's assumptions, CASAC advised the Agency to seek alternative rationales for characterizing the uncertainty in continued health risks at ever lower potential NAAQS levels. The final rule's preamble reflected that warning and used an extended discussion of evidence-based considerations as its sole justification for the final NAAQS decision. The HREA was rendered irrelevant to the decision.

Section 3.3 of the draft PA contains a risk analysis that is much the same as the last in the sequence of past $PM_{2.5}$ HREAs. It lacks any quantification of the still-important epistemic uncertainties associated with estimating health risks and health risk changes associated with alternative NAAQS. In fact, it now calculates risks at $PM_{2.5}$ levels down to zero exposure. The logical consequence of this should be obvious but is also clearly presaged by the history of $PM_{2.5}$ NAAQS decisions: the risk-based considerations presented in Section 3.3 of the draft PA are not informative to a decision maker faced with making a

[6] Section 3.2 of the draft PA contains information in the format associated with the "evidence-based" approach while Section 3.3, upon which I am specifically commenting, contains information in the format associated with the "risk-based" approach.

[7] These three alternative distributions were represented as hypothetical and not attributed to any specific person or group of people.

science-based judgment on a NAAQS level that is protective of the public health with an adequate margin of safety.

3. Themes in Special Issue of *Risk Analysis*

The reason the risk estimates in the past two prior $PM_{2.5}$ HREAs as well as those in the current draft PA are not informative for decision makers is *not* that they are poorly communicated or summarized. It is because they ignore virtually all forms of scientific uncertainty that are critical to the NAAQS policy judgment of what is requisite to protect the public health when zero-risk is not the required goal. When ignoring the impact of the most important factor in the relevant decision process (i.e., scientific uncertainty), a risk analysis renders itself irrelevant *by design*. Ignoring those scientific uncertainties also means that the draft PA's risk estimates are not reliable from any potential decision-making perspective.

The risk analysis profession has developed, over many decades, a structure or paradigm to guide the conduct of risk analyses in the direction of providing reliable decision-guidance. The risk analysis profession describes the process of risk analysis as having four components: (1) hazard identification, (2) exposure assessment, (3) dose-response assessment and (4) risk characterization. The first step, hazard identification, is to determine, through observational, clinical and laboratory investigation whether a health risk exists. If one is determined to exist, then estimates of exposures are combined with estimates of a dose-response relationship (i.e., the outputs of elements 2 and 3) to characterize the risk associated with the identified hazard. There are basic principles developed for each of the four steps.

Within this structure, an HREA reflects the second, third and fourth elements. They provide a synopsis of how the Agency has conducted its exposure assessments and what it assumes for its dose-response relationships, with the goal of reporting results of risk calculations that are based on the exposure and dose-response relationship inputs. However, these HREAs (including Section 3.3 of the draft PA) all start from a *presumption* of the existence of a hazard: that is, they *presume* a causal relationship between $PM_{2.5}$ and the health risk endpoints for which they make quantitative public health risk estimates.

Within this paradigm, epidemiological studies are, first and foremost, of relevance to the question of whether a hazard exists from exposure to ambient $PM_{2.5}$.[8] Unfortunately, the Agency conflates those studies of association with the different (and challenging) risk analysis element of dose-response assessment. That is, the Agency's analysts select a few specific epidemiological studies' coefficients from estimates of a health effects association and use those coefficients *at face value* as a rigid and deterministic prediction of how changes in concentration translate causally into changes in risk of the studied health endpoint. This egregiously short-changes the demands of a thorough dose-response assessment and represents the most fundamental flaw in the Agency's implementation of the risk analysis paradigm in its $PM_{2.5}$ HREAs and Section 3.3 of the draft PA.

In September 2016, the journal *Risk Analysis* presented a Special Issue on Air Pollution Health Risks that contained invited commentaries by six different professionals active in addressing risks associated with criteria pollutants such as $PM_{2.5}$. I was invited to read and respond to this set of commentaries, which resulted in publication in the same special issue of the article I have called **"Special Issue Synthesis"** in my list of papers above. That article points out that a consistent theme across those six sets of commentaries was that the primary challenges ahead for $PM_{2.5}$ risk assessment was associated with shifting from use of the epidemiological studies for the purpose to hazard identification into use of those studies for the dose-response element of the risk analysis paradigm. I do not see any evidence that

[8] Many concerns have been raised about the appropriateness of the Agency's causality framework. While I will not attempt to engage in that current line of discussion here, I did discuss my own concerns in technical comments that I wrote (and which were submitted into the docket) for the $PM_{2.5}$ NAAQS review that culminated in the 2013 rule. As I consider the issues that I raised at that time to remain relevant today, I provide a copy of those comments in an appendix to this document.

Agency staff have taken notice of the need to grapple better with the dose-response assessment challenges that was discussed by multiple different risk analysis professionals in that September 2016 Special Issue of *Risk Analysis*. The draft PA's discussion of risk-based considerations is thus out of touch with and unresponsive to concerns about how to more effectively address the dose-response element of the risk analysis paradigm – concerns that continued to be actively discussed among risk analysis professionals at the outset of the current $PM_{2.5}$ NAAQS review.

4. BenMAP's Capabilities and Limitations

The risk analysis calculations in the draft PA have been conducted using EPA's BenMAP model. BenMAP is a computational tool that calculates health risks when provided with a specific projection of ambient concentrations (i.e., the exposure input) and a specific assumed C-R function (i.e., the dose-response input). In my March 2015 paper that I have called the **"BenMAP Review"** in the list above, I provide a review of BenMAP's capabilities estimating the sensitivity of its risk estimates to uncertainties in the C-R relationships that users choose as inputs.

The actual formula for the risk calculation that BenMAP runs is simple and can be done on a hand-calculator. The reason the Agency has built this tool is so that it can make the application of the formula as disaggregated as one might desire. For example, exposure assumptions are often provided at the level of a 12 km by 12 km gridded surface of the U.S.[9] As there are nearly 50,000 such grid cells covering the 48 conterminous U.S. states, this means the simple risk formula needs to be applied about 50,000 times over. Furthermore, it requires 50,000 separate estimates of the cells' respective population and baseline incidence rates for each health effect. BenMAP's contribution is to estimate these highly disaggregated demographic data for whatever geographic scale is used for the ambient concentration input. However, BenMAP applies the same risk formula to the demographic and exposure estimates for every grid cell. My paper concludes that the main problem with BenMAP is that it suggests a great deal of sophistication in the risk analysis that is simply not there. It converts risk-relevant insights that can be obtained by a few back-of-the-envelope calculations into a vast and inscrutable array of model-generated numbers—a fact that is probably apparent to any reader of Section 3.3 of the draft PA.

BenMAP's design encourages its users to conduct risk calculations at a level of disaggregation that is far out of line with the degree of detail in risk relationships that the epidemiological studies provide. The **"BenMAP Review"** paper also concludes that its potential for computational complexity distracts users from assessing the sensitivity of its risk estimates to some of the most fundamental uncertainties associated with the C-R input assumptions. While a good risk analysis tool should encourage and enable users to conduct both sensitivity and uncertainty analysis, BenMAP cannot be used for integrated uncertainty analysis (IUA) and it discourages even sensitivity analysis.

My article goes on to illustrate three types of sensitivity analysis that would be helpful to add to BenMAP: (1) consideration of ranges of C-R slope estimates consistent with the full literature (rather than providing a "library" limited to the select studies that the Agency prefers to use); (2) a continuum of alternative assumptions about the $PM_{2.5}$ level where the assumed C-R relationship may cease to exist; and (3) alternative assumptions about the relative toxicity of $PM_{2.5}$ constituents.

In the third example (differential toxicity uncertainty), the article shows that the estimate of risk reduction due to a reduction in ambient $PM_{2.5}$ mass depends not just on what constituents are considered toxic, but also on what type of $PM_{2.5}$ reduction strategies would be taken. While the right assumptions are unknown, this sensitivity analysis shows that the risk estimate that the Agency produces under its

[9] This is the grid scale used in the draft PA's risk calculations for 47 urban study areas of the U.S.

standard assumption that all constituents are equally toxic is at the high end of the range of risk estimates that account for uncertainty in the relative toxicity of different $PM_{2.5}$ constituents.

In summary, the **"BenMAP Review"** paper focuses on the appropriateness of BenMAP's design for encouraging risk analyses that can provide useful and reliable guidance to decision makers. It finds that even sensitivity analyses to epistemic uncertainties are difficult to conduct given this tool's emphasis on enabling risk calculations to be conducted as a level of geographical detail that is far out of line with the current state of scientific knowledge about the underlying risk relationships.

In an invited comment on the **"BenMAP Review"** paper that appeared in the same journal issue, Agency staff cited an NAS report to dismiss the potential that there might be any discontinuation in the C-R relationship. In my **"Response to EPA"** (also in that issue), I pointed out that EPA was incorrectly characterizing the conclusions of that NAS report. I further noted that arguing against the existence of any particular scientific uncertainty was just another way of distracting rather than informing decision makers about the role of unresolved uncertainties in risk-based considerations. The commenters also expressed satisfaction that we were able to use BenMAP for our sensitivity analysis, but I noted that we had had to conduct our paper's sensitivity analyses in a spreadsheet outside of BenMAP because the spreadsheet was easier to work with than BenMAP itself, once BenMAP's estimates of the cell-by-cell demographic data had been extracted from it.

In brief, my advice to CASAC members regarding BenMAP is to have less concern about its ability to produce correct risk estimates for a given set of input assumptions and more concern about how its use is preventing the draft PA from exploring a reasonable range of variations in those input assumptions.[10]

5. Integrated Uncertainty Analysis

The discussion thus far has made the case that the risk analysis methods that EPA has historically used (and which are again used in the draft PA) fail to appropriately quantify the scientific uncertainties that are central to the $PM_{2.5}$ NAAQS decision. An important question is whether there are better methods to conduct such a quantification. It is exceedingly late in the current review process to revise the risk analysis, but with sufficient planning and a willingness to embrace the reality that these uncertainties are inherently subjective, there are alternative methodologies that could be taken. Often these approaches are called "integrated uncertainty analysis" (IUA) because they require that the effects of multiple different sources of uncertainty be accounted for in a single probabilistic analysis.

Among my own writings, I discuss several different ways that epistemic uncertainties can be incorporated into a NAAQS-related health risk analysis. I will summarize where these discussions can be found within the papers I have had published since the prior $PM_{2.5}$ NAAQS review:

- Section 2 of my **"HREA History"** paper ("Origins of Quantitative Risk Assessment for Criteria Pollutants") describes how IUA was incorporated into the 1997 HREA for $PM_{2.5}$. It was done by positing three alternative probability distributions (illustrative and implicitly subjective) over key input assumptions. The most important probabilistic input regarded how low the C-R function was believed to continue to exist. The HREA thus presented three alternative probability distributions over the resulting health risk estimates. No single distribution was endorsed, but it

[10] I do not, however, warrant that the particular estimates provided in the draft PA are correct. I have made no attempt to replicate them and am not aware of such effort by any other party. During the prior ozone NAAQS review, NERA identified significant errors in the BenMAP computations of risks for specific urban areas in that ozone HREA, and the Agency had to revise a large fraction of the HREA's risk tables. The errors were due to faulty mappings in BenMAP to access demographic data at different levels of geographic disaggregation—a BenMAP function that is especially "black box" in nature and thus difficult to check.

provided HREA readers and the Administrator with information about the net effect of the scientific uncertainties on the reliability of the risk estimates.

- Section 5 of my **"HREA History"** paper ("Rebuilding Relevance for the NAAQS Risk Assessment") makes a detailed recommendation for how future NAAQS reviews could alter the process by which HREAs could be conducted that would enable multiple alternative probability distributions over the continued existence of a causal C-R relationship to be identified and incorporated into the HREA. These alternative probability distributions would be proposed (and justified) by commenters rather than by EPA staff. Each proposed probability distribution would be run as a separate set of risk calculations, with resulting probabilistic risk estimates presented in collated fashion as the main output of the HREA. This would leave the Administrator free to examine the justifications associated with each submission, and ultimately give weight to the probabilistic risk estimates from the submission(s) that best fit the Administrator's own evaluation of the strengths and /or limitations of the scientific evidence.

- My January 2018 **"Ozone IUA"** paper provides a discussion of how epidemiological evidence of a potential threshold in the long-term ozone respiratory mortality could be incorporated into a probability distribution over that risk endpoint. The approach was to combine an initial (subjective) probability distribution on the potential level of such a threshold[11] with a statistically-based set of likelihoods on where such a threshold may lie. The latter set of likelihoods is based on the relative goodness-of-fit of the many alternative threshold model runs that were reported in the epidemiological paper that was being used as the source for a C-R function on this mortality risk. The result of this operation is a probability distribution over the potential threshold location that accounts for both subjective views and epidemiological evidence.[12] With such an approach, one can avoid having to choose between two deterministic assumptions: that there is no threshold (despite evidence that there may be one) or that there is a threshold at the level that produces the best-fitting model (despite that model having only marginal statistical significance over the no-threshold model). Using this combined distribution as the basis for an IUA-based risk analysis, this paper shows how even highly uncertain indications of a potential threshold, if incorporated probabilistically into the risk analysis, can result in a substantially different indication of the merits of alternative NAAQS levels compared to assuming (as the Agency did) that a linear-to-zero C-R function remained the best single deterministic assumption.

I recognize that the third of these examples was made possible by the availability of a set of epidemiological estimates for a wide range of alternative model specifications. However, the methodological benefits of having such alternative model estimates are apparent. They enabled a direct and quantitative characterization of the model uncertainty in the risk estimates without forcing any definitive (deterministic) judgment on the level of the true threshold. Even the possibility of no threshold continued to be included in the IUA. If it were to become a standard among air pollution epidemiologists to provide the statistical fits of alternative model specifications applied to their dataset, such hybrid statistical-and-subjective judgments could be used more often. This could provide a middle ground between the purely deterministic approach of the current draft PA and the purely subjective approach suggested in the **"HREA Review"** paper.

6. Inconsistencies Between NAAQS and RIA Risk Analysis

All of the above discussion concerns incorporation of epistemic uncertainty into the risk analyses that are used to inform a decision on whether to revise or keep a NAAQS standard, i.e., in the HREAs. Those risk

[11] In technical terms, this initial probability distribution is called a "prior" distribution.

[12] In technical terms, this is called a "posterior" distribution.

analyses are the relevant ones for CASAC to evaluate. However, many of the same issues about epistemic uncertainty pertain to the risk and benefit calculations of the RIAs that accompany a NAAQS rulemaking. The risk calculations in RIAs for criteria pollutants may seem to mirror the risk calculations in the HREAs that occur before a NAAQS decision is made. However, in my recent research I have noticed a significant inconsistency between the rationales that are used to justify a final NAAQS decision and the risk input assumptions that are then used in the RIA for that same NAAQS. This inconsistency is documented in my September 2016 **"Inconsistencies"** paper.

Briefly, the rationale for setting the annual $PM_{2.5}$ NAAQS at 12 μg/m^3 was largely an evidence-based judgment by the Administrator that the available epidemiological evidence was too limited to give her confidence that the key C-R relationships continued to exist below about 12 to 13 μg/m^3. Thus, while health risks could not be said to be zero below 12 μg/m^3, protection against exposure to lower levels of $PM_{2.5}$ was considered not to be requisite. Nevertheless, the accompanying RIA assigned as much confidence to the risks it estimated for $PM_{2.5}$ exposures well below the standard (including those near zero) as it did to those estimated for exposures above or near the standard. My paper demonstrated that 70% to 96% of the benefits attributed to that rulemaking were due to this inconsistent assumption in the RIA.

In my most recently accepted paper, which I have labelled **"Improving Consistency,"** I revisit this inconsistency issue. I make a recommendation for how an Administrator's expressions of uncertainty about the C-R relationship might be incorporated into NAAQS-related RIAs' risk calculations to generate consistency between the reasoning behind a NAAQS decision and its associated RIA. The concept is that the Administrator's subjective probability distribution on the continued existence of the mortality C-R relationship could be formally elicited and directly reported as part of the formal rationale for a NAAQS decision. This elicited distribution would be presented as a quantitative supplement to the usual evidence-based reasoning already found in NAAQS rationales. (It could be provided without any risk-based calculations, or it could also be incorporated into HREA estimates to be summarized in the preamble.) Importantly, however, it would become a direct input to a probabilistic (confidence-weighted) estimate of the rule's benefits that would be presented in the accompanying RIA. In my **"Improving Consistency"** paper, I provide illustrative examples of what the resulting probabilistic risk estimates might be and contrast them to the inconsistent risk estimates that result from the deterministic (no threshold) C-R assumptions presently used in RIAs. (The illustrative example is based on the RIA for the $PM_{2.5}$ NAAQS promulgated in 2013.)

Instituting such a process for preparing RIAs that are logically-consistent with the basis for the NAAQS decision would have no effect on the NAAQS decision itself, as the RIA is not an input to that decision. It would, however, create greater acceptance of the RIA results. Also, the impact on RIA estimates would carry over into RIAs for rules that do not target $PM_{2.5}$, but which treat coincidental reductions in $PM_{2.5}$ as *co-benefits*. Adoption of this confidence-weighted risk calculation for all RIAs could significantly reduce the current controversies over the appropriateness of basing non-NAAQS rulemakings on co-benefits.

APPENDIX.
COPY OF 2009 COMMENTS OF DR. SMITH ON EPA's CAUSALITY FRAMEWORK

CRA Charles River Associates

Comments on the External Review Draft of EPA's "Risk Assessment to Support the Review of the PM Primary National Ambient Air Quality Standards"

Anne E. Smith, Ph. D.
Charles River Associates
1201 F Street NW, Suite 700
Washington, DC 20004

Prepared at the request of the Utility Air Regulatory Group, American Petroleum Institute and American Chemistry Council

November 8, 2009

In September 2009, the U.S. Environmental Protection Agency (EPA) released its external review draft of a document titled "Risk Assessment to Support the Review of the PM Primary National Ambient Air Quality Standards" (hereafter, the "Draft PMRA"). Following are my comments on issues in conducting a quantitative risk assessment for $PM_{2.5}$.

I. Introduction and Summary of Main Points

EPA's Draft PMRA is intended to provide quantitative estimates of the levels of risk from $PM_{2.5}$ actually being experienced by the U.S. population today, and how those risks will change if current ambient $PM_{2.5}$ is reduced by the application of more stringent National Ambient Air Quality Standards (NAAQS). The Draft PMRA only quantifies those risks that have been determined to be "causal" or "likely causal" in the second draft of the Integrated Science Assessment (hereafter, the "Draft ISA"). Once such a determination is made in the ISA, however, the Draft PMRA not only presumes that statistical estimates of $PM_{2.5}$'s relative risks are causal, but also that they can be interpreted quite literally as the quantitative concentration-response functions that determine actual risks. Whatever merit the observed epidemiological associations may have as indicators of a causal relationship, the unquestioning numerical credence that EPA assigns to the epidemiological estimates undermines the credibility and reliability of the Draft PMRA results. A common adage is "association does not imply causation;" to this can be added that even if a statistical association *does* reflect causation, it does not define the actual quantitative response function.

There are several layers of problems with the quantitative risk calculations in the Draft PMRA:

- At the most fundamental level, the Draft PMRA *presumes* that there is a causal association in the epidemiological evidence. That presumption is less than settled, as my comments will explain, because all of the studies may be wrong for the same systematic reason.

- At the next level, the estimates of relative risk from the epidemiological studies are almost certainly highly biased and, as my comments will explain, the bias is likely in the upward direction. Because the Draft PMRA uses those estimates directly for its quantifications of risk, the Draft PMRA's estimates of current premature mortality from $PM_{2.5}$ exposures are probably overstated.
- Topping it off, the epidemiological studies are incapable of defining how the relative risk would tend to vary at increasingly lower levels of $PM_{2.5}$. This creates an increasingly large error as risk reductions are estimated for tighter and tighter alternative ambient $PM_{2.5}$ standards.

All of the above problems in using the epidemiological relative risk estimates for quantitative purposes stem from fundamental data limitations that face every single one of the epidemiological research teams. This situation does not imply any fault on the part of the research teams or that the quality of their work is not of high quality. Unfortunately, however, EPA understates the remaining uncertainties that result from these studies' inherent data limitations when it engages in the quantitative risk assessment in its Draft PMRA.

Given that it does not address this array of problems in using epidemiological studies to attempt to quantify risk, the quantitative risk assessment in the Draft PMRA is highly misleading as an input to policy decisions. EPA could mitigate this situation by finding ways to quantitatively incorporate corrections for the systematic biases. This would produce a larger range of uncertainty in its estimates of risk, but one that reflects the true current state of knowledge. If this is not done, however, then the Draft PMRA should not be used in the consideration of alternative $PM_{2.5}$ NAAQS.

The rest of my comments are organized in the following way:

- Section II explains the problem in making a presumption of causality in the Draft PMRA, even though this is the determination in the Draft ISA.
- Section III discusses how the data limitations of the available $PM_{2.5}$ epidemiology literature make it inappropriate to use the estimated relative risks from those studies directly in a quantitative risk assessment as the Draft PMRA does.
- Section IV points out that even the epidemiological studies indicate a non-negligible chance that $PM_{2.5}$ imposes no long-term risk to all-cause mortality at all, once they are reviewed in a less selective manner than in the Draft PMRA.
- Section V summarizes and concludes that until these uncertainties are addressed quantitatively in the PMRA, it will be unreliable, and should not be used in the consideration of alternative $PM_{2.5}$ NAAQS.

II. EPA's Causality Criteria Are Inappropriate and Promote False Confidence in Quantified Risk Estimates in the Draft PMRA.

The Draft ISA provides a review of the weight of evidence in favor of alternative degrees of causal inference for various effects, such as between long-term exposures to $PM_{2.5}$ and cardiovascular mortality. The Draft ISA's determinations of causality for long-term $PM_{2.5}$ associations with CVD mortality and likely-causal for all-cause mortality associations are heavily driven by evidence of statistical associations in observational epidemiology studies. As explained below, this degree of reliance on the epidemiological evidence is excessive, highlighting a weakness in the criteria for causality that EPA establishes in that document. Given that uncertainties in the causality determination are never questioned again in the Draft PMRA, they are important to discuss in these comments on the Draft PMRA.

(II.A.) EPA's criteria allow excessive reliance on epidemiological findings in making a determination whether pollutants are causally associated with health effects.

The Draft ISA provides a set of criteria that must be met to determine that a particular health effect is causally related to $PM_{2.5}$ exposure. I quote the criteria below, which I have broken into two parts for purposes of the discussion that follows:[1]

Part 1: "The pollutant has been shown to result in health effects in studies in which chance, bias, and confounding could be ruled out with reasonable confidence. For example: a) controlled human exposure studies that demonstrate consistent effects; or b) observational studies that cannot be explained by plausible alternatives or are supported by other lines of evidence (*e.g.*, animal studies or mode of action information)."

Part 2: "Evidence includes replicated and consistent high-quality studies by multiple investigators."

On its own, Part 1 provides what would seem to be a perfectly appropriate set of criteria for making a causal determination. However, the addendum of Part 2 greatly weakens the requirements, because as a logical "or" statement, it provides a way for EPA to conclude in favor of causality even if controlled human exposures studies do *not* demonstrate consistent effects, *and* observational studies *can* be explained by plausible alternatives *and* are not supported by animal studies or mechanistic actions; Part 2 allows EPA to conclude in favor of causality even in the face of all of the foregoing findings as long as multiple authors have published quality epidemiological studies that replicate each other. Thus, the single sentence of Part 2, treated as a logical "or" rather than as a logical "and" to the requirements of Part 1, serves to absolve EPA from having to demonstrate that the associations in chronic studies "cannot be explained by plausible alternatives" before it can make its causal determination.

[1] Draft ISA, Table 1-3, p. 1-29.

Unfortunately, EPA relies almost entirely on the type of evidence allowed by Part 2 to make its causal determination regarding long-term mortality risks of $PM_{2.5}$. Take the case of long-term cardiovascular (CVD) mortality risk, for example. The Draft ISA relies almost entirely on the existence of a multiplicity of separate chronic studies that all find a $PM_{2.5}$-mortality association. It begins and ends that causality discussion with the following respective quotes:

> "A number of large, multicity U.S. studies (the ACS, Six Cities Study, WHI, and AHSMOG) provide consistent evidence of an effect between long-term exposure to $PM_{2.5}$ and cardiovascular mortality."[2]

and:

> "In summary, a number of large U.S. cohort studies report associations of long-term $PM_{2.5}$ concentration with cardiovascular mortality. These studies provide the strongest evidence for an effect of long-term $PM_{2.5}$ exposure on CVD effects."[3]

EPA determines that this evidence from multiple chronic studies suffices to identify a causal relationship, even though the remainder of the supporting evidence that the Draft ISA musters in support of a long-term CVD mortality risk are some inconsistent morbidity studies, inconsistent clinical studies, and a few toxicological studies that are suggestive of some possible relevant responses. The Draft ISA does not offer any reasons to believe that any of the many observational studies have met EPA's conditions for making a causal determination under Part 1. Nowhere does EPA make a case that the associations in the chronic mortality studies "cannot be explained by plausible alternatives." In fact, it could not possibly make such a claim given that epidemiologists continue actively to try to rule out various plausible alternatives. The plausible alternatives that researchers have not yet been able to rule out include confounding (residual or otherwise) by co-pollutants, noise, stress, and socioeconomic factors.[4] For example, the most recent ACS cohort analysis (Krewski *et al.*, 2009) focused vigorously on more effectively controlling for socioeconomic factors than in past studies, but it made no attempt to rule out possible confounding by co-pollutants.

Thus, the available evidence for long-term mortality risk does not meet the causality criteria contained within Part 1, and EPA is relying very heavily on Part 2 to defend its "causal" determination. <u>Part 2, however, is not a valid basis for a causal determination if observational studies are not also supported by consistent effects in controlled studies</u>. It would be a reasonable addendum if it were required *in addition* to Part 1 conditions, but not when used *instead of* meeting Part 1 conditions. The reason it is insufficient for

[2] Draft ISA, p. 7-25.

[3] Draft ISA, p. 7-26.

[4] The effect modification by educational status in the chronic studies remains as an indication of some residual confounding by socioeconomic factors. Although this pattern was reduced in the Krewski *et al* (2009) study, it was not eliminated.

identifying whether a statistical association is causal is because it fails to address the possibility of *systematic biases*, which cannot be ruled out except with evidence from controlled studies.

Systematic biases will occur if the studies in question have relied on similar methodologies and similar data sources. In this case, if a bias (*e.g.* due to residual confounding) exists in one study, then it is likely to exist in all the studies. All of the epidemiological results can be wrong for the same reason. A multiplicity of studies finding a statistical association do not provide independent confirmation supporting a causal inference unless one can demonstrate that there is no potential for such systematic bias among those multiple studies. Thus, Part 2 of EPA's causality criteria enables causality to be declared even if there remains a very large likelihood of no causality.

There is substantial potential for systematic bias in the case of chronic $PM_{2.5}$-mortality epidemiology.[5] Table 7-8 of the Draft ISA lists 14 recent U.S. cohort studies that find an association between $PM_{2.5}$ and mortality.[6] All of these studies draw from the same fundamental data set, however, because they all sample individuals across the U.S. and assess the correlation between their local monitors' $PM_{2.5}$ levels and their mortality risks after attempting to control as best possible for the very broad swath of much stronger determinants of risk (*e.g.*, age, sex, diet, smoking habits, and socioeconomic factors). Controlling effectively for these other factors is the key to getting a sound answer, y*et al*l of the studies are reduced to using approximately the same approximate data, all of them facing enormous amounts of error in how those variables are assigned to individual cohort members. In any single study, there is a good chance that the controls for the primary determinants of mortality risk are incomplete, and some confounding remains to bias the association estimated for $PM_{2.5}$. Unfortunately, all of these studies face the same problem, in a systematic way, because they all rely on the same types of data, and face the same fundamental data limitations.[7]

The fact that these studies rely on several different cohorts does not make them independent of each other with respect to confounding and effect modification. If ambient $PM_{2.5}$ is correlated with the missing or poorly measured non-PM explanatory variables across the U.S., then almost any reasonably diverse subset of the U.S population are likely to embody that same underlying correlation. For example, a sample of mostly white volunteers for a cancer study and a sample of veterans may have quite different socioeconomic profiles, but both will reflect the general correlations that exist across the U.S. between $PM_{2.5}$ measured at central monitors and key socioeconomic or other non-$PM_{2.5}$ causal factors. The same biases can apply to every single cohort of people drawn from the U.S. population. This is a particular systematic concern for

[5] Goodman (2009), pp. 8-9, makes a similar point.

[6] Draft ISA, pp. 7-119 to 120.

[7] There is also uniformity in the methodologies being used, in that almost all of the researchers use the same statistical model, the Cox proportional-hazards (PH) model, and thus systematically share any biases that may derive from the limitations of this statistical model. For example, the Cox PH model assumes the effects of pollution levels and of potential confounders on the logarithm of hazard are all linear. The assumption of proportional hazards has received limited testing, but that which has been done raises serious questions about this key assumption (see, for example, Abrahamowicz *et al.*, 2003).

variables that must be controlled at the ecological level. Thus, the biases from confounding and effect modification that are most difficult to control are also likely to be systematic across multiple U.S. cohorts.

This is not a criticism of the quality of the research teams' efforts; it is just an unfortunate reality of the limitations of the available data and tools to study such a subtle possible risk without the ability to perform controlled experiments. Nevertheless, the potential for systematic bias should not be ignored.

(II.B.) Differential measurement error can systematically create bias from confounding, even if confounders are included in the epidemiological study.

The potential for systematic confounding and effect modification cannot be eliminated simply by including the relevant explanatory variable in the data analysis if there is measurement error. A confounder is a variable that has a direct effect of its own on risk, *and* which is correlated with the pollutant being studied. When this is the case, if the confounder is left out of the analysis, the pollutant will be attributed some of the explanatory power that is actually due to the missing confounder, and thus the relative risk estimate will be biased. If personal exposures to all the confounders and effect modifiers can be measured accurately, the bias in the $PM_{2.5}$ exposure-risk association can be eliminated by identifying and including the confounder in the data analysis. However, if the confounder cannot be measured with accuracy, even if it is included in the data analysis, *residual* confounding will remain in the estimated association between the exposure variable and risk.

The situation is complex, but simulation studies can help understand the potential for biased effects estimates in situations that have confounders with differential measurement error. One such study (Fewell *et al.*, 2007) demonstrates that typical amounts of confounding, combined with typical amounts of measurement error, can cause quite large relative risks to be assigned to an exposure variable that has no effect at all, even when measures of the confounder are included among the controlling covariates in the analysis. The size of the potential erroneous relative risk reported by Fewell *et al.* exceeds the magnitude of the $PM_{2.5}$ relative risk in the $PM_{2.5}$ epidemiological literature. The implication is that the estimated chronic $PM_{2.5}$ relative risks – given their fairly small magnitude – could be the result of residual and unmeasured confounding by either socioeconomic factors or other environmental factors. Because these potential confounding relationships are likely to exist for every cohort if they exist for one, all of the multiple, independent long-term $PM_{2.5}$-risk associations could be reflecting the same systematic biases (Boffeta *et al.*, 2008).

Cohort studies in other countries might not face all of the same systematic errors that would apply to cohort studies all from the U.S. However, if the source of the confounder is physically related to PM sources, then one would see the same systematic bias even for cohort studies outside of the U.S. Thus, positive findings from cohort studies in other countries might reduce some of the concern with a socioeconomic-pollutant correlation, but cannot eliminate concerns that PM is a proxy for another co-emitted pollutant, some non-chemical effect coinciding with chemical emissions (*e.g.*, noise), or for a single constituent of PM. Table 7-8 of the Draft ISA identifies one $PM_{2.5}$-mortality cohort study from the Netherlands (Brunekreef *et al.*, 2009). The abstract for this study states that it "differs from cohort studies based on city-level differences in exposure" because it considered exposures to pollution sources that exist mainly *within* cities. The U.S. cohort studies, which are based on city-level differences in pollution, are reporting an association between health and the total, undifferentiated mass of *all* forms of $PM_{2.5}$. It is those total mass associations that are then being used for risk calculations in the Draft PMRA. Thus, this non-U.S. study does not provide corroborating evidence of the same kinds of relationships being estimated by the U.S. cohort studies, and so it does not help eliminate concerns that potential systematic biases may pervade the latter.

(II.C.) Other pollutants are a likely source of residual and unmeasured confounding bias in the long-term associations.

The potential for the long-term $PM_{2.5}$ estimates to all share bias due to confounding from other pollutants is clear, regardless of what one might think about residual socioeconomic variable confounding in this body of literature. Rarely are other pollutants or pollutant-sources included in $PM_{2.5}$ epidemiological regressions; however, when they are included, there is often a marked reduction in the size and statistical significance of the $PM_{2.5}$ effect. Such a sensitivity upon the inclusion of SO_2 was a major finding of the reanalysis of the ACS data by Krewski *et al.* (2000); yet, the subsequent papers of Pope *et al.* (2002) and Krewski *et al.* (2009) that extended the ACS cohort analysis did not report any $PM_{2.5}$ relative risks that had also been controlled for SO_2. This omission in recent ACS-based studies leaves an important question regarding the quantitative validity of the $PM_{2.5}$ associations taken from Krewski *et al.* (2009) that the Draft PMRA uses.

The other study that EPA places high reliance on is the update of the Harvard Six Cities study by Laden *et al.* (2006). EPA cites this study as providing confirmatory evidence that long-term reductions in $PM_{2.5}$ produce a corresponding reduction in mortality. However, this study does not consider any pollutants other than $PM_{2.5}$, even though the levels of various gaseous pollutants have fallen concomitantly with $PM_{2.5}$ in the six cities. Rather, Laden *et al.* attribute the entire change in health risk associated with air pollution reduction to $PM_{2.5}$ without any apparent attempt to test whether any other pollutant might have equivalent explanatory value.[8] The paucity of data points available with this cohort make it impossible for statisticians to even attempt to control for more than one pollutant at a time. That is, with the Harvard Six Cities cohort, the estimate of the effect of pollution on inter-city mortality risk differences must be based on only 6 cities/data points. In contrast, as many as 150 cities/data points are available for inferring relative risk estimates with the ACS data set. Nevertheless, the researchers still could have used a series of one-pollutant models to explore whether pollutants other than $PM_{2.5}$ might also be associated with the observed changes in inter-city mortality risks.

[8] Furthermore, in the critical second period of this study, the results were not based on actual fine PM measurements. Rather, they are rather based on measurement of another NAAQS pollutant (PM_{10}) and extinction coefficients.

Even if other pollutants were to be included in these long-term risk studies, it is quite likely that they would fail to control for confounding because of differential measurement error. $PM_{2.5}$ is generally believed to have a much more uniform distribution in space than other pollutants with which it is correlated, including NO_2, CO, coarse PM, and even ozone. Thus, the standard practice of using data from a central monitor to estimate individuals' exposures probably results in greater exposure misclassification for these other pollutants than for $PM_{2.5}$. The result could be that $PM_{2.5}$ will persistently appear to carry the best explanatory power, yet just be serving as a proxy for the health effects of the other, more erroneously measured pollutant exposures. This possibility was demonstrated in a simulation study that contained considered two hypothetical correlated pollutants, one a "True Culprit" measured with relatively large error, and the other an "Innocent", but measured with relatively small error. The simulation results showed that:

> "in circumstances like this, *which* pollutant would appear to have the most significant and consistent relationship with health may be determined more by its relative observation error than by its actual contribution to the health effects in question. The greatest problem with this spurious association of Innocent with health is that it remains stable whether or not True Culprit is added into the regression. Further, if only True Culprit is included in the regression, the R^2 falls to zero. True Culprit is the pollutant that seems to have a highly unstable association and very little explanatory power on its own. Thus, unlike in the simple confounding case, the usual methods for checking for confounding no longer function well when there are observation errors as well as strong correlation among pollutants."[9]

(II.D.) Evidence of proxy effect exists in changes in estimated $PM_{2.5}$ relative risk coefficients over time.

One signal that a non-causal proxy effect might account for the $PM_{2.5}$-mortality associations would occur if the estimated relative risk for $PM_{2.5}$ mass were to increase over time as $PM_{2.5}$ levels decline. That is, if there is a given amount of risk associated with a certain non-$PM_{2.5}$ causal factor that is correlated with $PM_{2.5}$ mass, if the unidentified causal factor was not reduced while $PM_{2.5}$ was reduced, then the remaining lower levels of $PM_{2.5}$ would account for the same total level of risk from the unidentified factor. The result would be a greater relative risk associated with a given amount of PM difference.[10]

[9] Smith and Chan (1997), p. 21. (The observed correlation between the two pollutants in this analysis was 0.56, which is in the range often observed in the US.)

[10] This signal will not necessarily occur even if $PM_{2.5}$ is serving solely as a proxy for some unnamed causal factor, if the causal factor were to be reduced in roughly the same degree as $PM_{2.5}$ over that same period of time. That could be the case if $PM_{2.5}$ is serving as proxy for a gaseous pollutant, since most of the pollutants have been declining simultaneously due to parallel environmental regulations.

We do observe this pattern in the extended analyses of the ACS cohort. For example, in Pope *et al.* (2002), the estimated relative risk per 10 µg/m^3 of $PM_{2.5}$ for all-cause mortality rose from 1.04 when using 1979-1983 $PM_{2.5}$ data (which averaged 21.2 µg/m^3) to 1.06 when using 1999-2000 data (which averaged 14.0 µg/m^3). The estimated relative risk for cardiopulmonary risks rose from 1.06 to 1.08.[11] Similarly, in Krewski *et al.* (2009), the all-cause relative risk rises from 1.043 to 1.056 and the cardiopulmonary relative risk rises from 1.089 to 1.129 when estimated with the earlier or later $PM_{2.5}$ data.[12] (Laden *et al.* (2006) report a decline in the relative risks from an earlier period to a later period of exposure, but this is not the same comparison. In the ACS examples, the risk over the same follow up period is estimated using $PM_{2.5}$ from two different parts of the time period. Laden *et al.* do not report estimates or relative risk for the entire time period using first the earlier, then the later $PM_{2.5}$ measures. Since they do not report a comparable set of relative risks, their finding cannot be said to conflict with the ACS finding just described.)

The increase in the estimated relative risk that occurs in the ACS data set when more recent $PM_{2.5}$ data are used might also occur if there is a real effect from $PM_{2.5}$ mass that is truly long-term in nature. In that case, on-going mortality outcomes might be a function of earlier exposures to $PM_{2.5}$, when it was at higher levels, while more recent $PM_{2.5}$ measures might be serving as a proxy for the historically higher $PM_{2.5}$ exposures. Even if this does explain the upward trend in estimates of $PM_{2.5}$ relative risks, it implies that any quantitative estimate of benefits from reducing current $PM_{2.5}$ based on the numerical results of recent epidemiological associations will be biased upwards. That is, adoption of a relative risk estimated using the more recent $PM_{2.5}$ "at face value" as the quantitative $PM_{2.5}$ concentration-response function would be erroneously assuming that the entire increase in risk due to higher historical $PM_{2.5}$ exposure is caused by a much smaller amount of $PM_{2.5}$ exposure. This concentration-response function would overstate the risk from as-is $PM_{2.5}$, and it would overstate reductions in risk that could be expected by reducing today's lower $PM_{2.5}$ to yet lower levels.

In summary, proxy effects can be at play in chronic studies, even if there is a causal relationship for $PM_{2.5}$ mass generally, and this proxy effect would create erroneous (overstated) risk and risk-reduction estimates in the PMRA. The bias would then be exacerbated even further when considering the benefits of further reductions in $PM_{2.5}$ due to rollbacks to alternative, more stringent $PM_{2.5}$ standards. The latter possibility is discussed further in Section III.B.

(II.E.) Epidemiological findings on short-term mortality are far more heterogeneous, and do not provide strong back-up to long-term studies.

Some people prefer to rely on short-term, time-series studies for evidence of an effect from $PM_{2.5}$ because effects observed within each city provide more inherent control for

[11] Pope *et al.* (2002), Table 2, p. 1136.

[12] Krewski *et al.* (2009), Table 6, p. 23. Values reported are for regressions with MSA & DIFF ecological controls, but the pattern also appears in other regression in the table.

socioeconomic factors that are otherwise difficult to measure accurately. While this may be true to some extent for the socioeconomic factors, short-term studies are still subject to potentially uncontrollable confounding from other pollutants. Nevertheless, existing short-term risk studies offer very little support for a causal interpretation of the observed long-term $PM_{2.5}$ associations. In particular, the quantitative level of the risk in the long-term studies is roughly an order of magnitude higher than associations found in short-term studies. The difference could be entirely due to confounding bias, or – as EPA prefers to explain it – the difference could be that cumulative, long-term effects are much larger than acute effects. Neither explanation can be held up as more correct, but even EPA's preferred explanation implies that the short-term studies cannot be viewed as corroborating the long-term study findings, because it implies that the long-term associations would have to be for an effect that acute studies cannot even detect.

Short-term studies also produce results that vary enormously from city to city and regionally, often finding no effect at all, even in cities and regions with relatively high $PM_{2.5}$ levels. This heterogeneity may indicate that the smaller, short-term $PM_{2.5}$ associations are not necessarily causal either.

III. Even If the $PM_{2.5}$ Association Is Causal, Statistical Estimates of $PM_{2.5}$ Relative Risk Remain Subject to Biases that Make Them Unreliable for Quantifying Risks.

(III.A.) Biases in estimates of the average magnitude of the $PM_{2.5}$ association are likely due to four types of data problems.

As explained in Section II, the Draft ISA's conclusion that $PM_{2.5}$ is causally related to cardiovascular mortality risks (and likely causally related to mortality risks in general) remains open to reasoned debate. However, a variety of uncertainties also exist that directly undermine the *quantitative* interpretation of the epidemiological findings for determining what numbers of deaths are premature at present, and especially for predicting how mortality risks would change if $PM_{2.5}$ mass were reduced. There are at least four ways in which quantitative biases can be present in the epidemiologically-estimated associations that would undercut their reliability for quantification of risks, as discussed below.

<u>(1) Differences in potency of various $PM_{2.5}$ constituents.</u> There are uncertainties about the relative potency of different constituents within the $PM_{2.5}$ mass.[13] Thus, even if a relative risk estimate is quantitatively valid as an *average* effect of the current mix of $PM_{2.5}$, if some constituents would not be reduced as much as others when an alternative $PM_{2.5}$ standard is imposed, then the reduction in risk from that standard would not be what one would predict using the *average* relative risk. In fact, if some small subset of the mass is highly potent and accounts for most or all of the observed association, it is

[13] The Draft ISA states (at p. 7-129) that "only a very limited number of the chronic exposure cohort studies have included direct measurements of chemical-specific PM constituents other than sulfates, or assessments of source-oriented effects, [in] their analyses." Also (at p. 2-25): "It remains a challenge to determine relationships between specific constituents, combination of constituents, or sources of $PM_{2.5}$ and the various health effects observed."

quite likely that this culprit would escape implementation plans, which will naturally be focused on reducing the constituents that account for the largest portions of the mass. The result could be no risk reduction at all, despite reductions in $PM_{2.5}$ mass; yet the Draft PMRA's methodology would predict substantial benefits from the tighter standard.

(2) Missing or inaccurate socio-economic variables correlated with regional $PM_{2.5}$ levels. All of the epidemiological studies have taken steps to provide controls for socioeconomic variables that affect mortality risks, but information is insufficient to ensure that these have been fully specified; also, many of the potential socioeconomic confounders and effect modifiers can only be measured with substantial amounts of error. Sometimes these errors can be found and eliminated through careful data quality work.[14] However, the errors of concern here are for socioeconomic data that are simply not possible to obtain. For example, although various updates of the key ACS study have been published up through 2009, they rely on the same individual and socioeconomic data collected in 1982, almost 30 years ago. Thus, it is not possible to assess changes in key confounding factors such as smoking cessation rates that are well known to fall along socioeconomic lines.[15] Other socioeconomic variables, such as data on the degree of stress in family life, are simply not possible to obtain and will never be possible to control for in the chronic risk studies.

Thus, despite extensive socioeconomic controls in all of the chronic risk studies, there remains the possibility that $PM_{2.5}$ mass is at least partially serving as a proxy for unidentified, or poorly measured, socioeconomic variables. If so, the $PM_{2.5}$ risk coefficient is biased. In this case, while it is not certain what the direction of bias would be, it is likely to be in the upward direction because lower socioeconomic status tends to be positively correlated with mortality risk and also with living in areas with higher pollution.[16] Regardless of direction of bias, quantified risk estimates that use the estimated $PM_{2.5}$ relative risk "at face value" will be incorrect.

(3) Other pollutants, even if included in the analysis. Much has been said about the possibility that $PM_{2.5}$ is serving as a proxy for another pollutant that has the true causal role. Studies have, at times, considered the role of other pollutants but this practice has been inconsistent.[17] When multi-pollutant results are not reported, one never knows if

[14] For example, the reanalysis by Krewski *et al*. (2000) of the Harvard Six Cities Study found that the coding protocol allowed cigar and pipe smokers to be classified as "non smokers;" the calculation of pack-years of smoking cigarettes was inconsistent, resulting in an underestimate of smoking pack-years of about 3% in some cities; and the error rate for the education variable on the earliest form used was 18%; etc. These kinds of errors can be avoided through careful review, and fixed, if detected.

[15] Some hypothesize that the errors in data on smoking cessation might explain the education gradient in $PM_{2.5}$ mortality observed in this study as well as the Harvard Six Cities Study; if so, the $PM_{2.5}$ relative risk estimate is probably a biased estimate.

[16] In the one known example where those with higher socioeconomic status happen to live in an area where the $PM_{2.5}$ is higher (*i.e.*, New York City), the lack of any increased mortality risk attributed to exposure to $PM_{2.5}$ in this group, versus those with lower socioeconomic status and lower $PM_{2.5}$ exposure, may illustrate the importance of socioeconomic confounding in air pollution epidemiology studies (Krewski *et al*. 2009).

[17] The Draft ISA states (at p. 7-82): "Given similar sources for multiple pollutants (e.g., traffic), disentangling the health responses of co-pollutants is a challenge in the study of ambient air pollution."

they may have been performed and found to have had the effect of attenuating the reported association for $PM_{2.5}$. Even if those studies simply did not perform any multi-pollutant regressions, one must wonder, why not? For example, the most recent papers based on the ACS cohort (Pope *et al.*, 2002; Krewski *et al.*, 2009) did not attempt to explore whether SO_2 had greater explanatory power than $PM_{2.5}$ mass, even though this was a widely discussed source of sensitivity reported in the preceding ACS paper (Krewski *et al.*, 2000).[18] Thus, the most recent epidemiological studies are not necessarily the most thorough in their efforts to explore confounding by other co-pollutants; as a result, their quantitative estimates cannot be viewed as more reliable for use in a quantitative assessment of $PM_{2.5}$-related risks. Relative risk estimates from the earlier studies that *did* account for co-pollutants could be less biased than the more recent relative risk estimates, even though the earlier ones have less statistical power due to their reliance on shorter cohort follow-up periods.

(4) Other environmental factors. Gaseous pollutants are not necessarily the only other non-$PM_{2.5}$ environmental factor for which $PM_{2.5}$ might be serving as a proxy. Measures such as proximity to traffic and intensity of local traffic have been the subject of much recent exploration of the basis for the $PM_{2.5}$ associations (*e.g.*, Lipfert, Wyzga *et al.*, 2006; Lipfert, Baty, *et al.*, 2006; Jerrett *et al.*, 2005; Beelen *et al.*, 2008a). In most cases where they have been accounted for in the data analysis, traffic-related variables appear to have the stronger associations. This could point to certain $PM_{2.5}$ constituents, or to some gaseous pollutants, and it could point to other factors such as noise or stress (Lipfert, Baty *et al.*, 2006; Beelen *et al.* 2008b).[19] Rolling back $PM_{2.5}$ mass would not have any effect on these other possible causal factors; again, the quantitative interpretation of a $PM_{2.5}$ relative risk would produce completely erroneous estimates of risk reductions from alternative $PM_{2.5}$ standards.

The point is obvious: even if the associations with $PM_{2.5}$ have a causal element, the many limitations of the epidemiological data mean that the relative risks estimated by epidemiological studies do not offer a direct quantitative relationship for how $PM_{2.5}$ mass alone affects either current health risks, or changes in risks if $PM_{2.5}$ mass is reduced. In all of the situations described above, the observed association could be "statistically significant," yet the health benefits from rolling back $PM_{2.5}$ could be as low as zero, because statistical significance calculations cannot detect the presence of bias; in fact, they presume it does not exist. Also, if differential measurement errors are at work, then

[18] HEI's commenters on the 2009 Krewski *et al.* study lament the lack of further study of confounding by copollutants, but offer their own excuse for this omission: "Given that the Reanalysis (Krewski *et al.*, 2000) had extensively tested the potential for the gaseous pollutants to confound the relationship between exposure to $PM_{2.5}$ and mortality and had not found any significant confounding (other than by SO_2), it is understandable that the current investigators chose to focus their limited resources on the extensive exploration of spatial autocorrelation in a series of one-pollutant models." (Krewski *et al.*, 2009, p. 130, emphasis added). This is a rather weak reason if their goal is to explore the strength of the $PM_{2.5}$ mass association in greater detail, given that their previous paper's findings on that association were the most sensitive to the inclusion of SO_2 as a co-pollutant.

[19] Bukowski (2008) has suggested that noise and stress could be an uncontrolled factor also affecting short-term $PM_{2.5}$-exposure risk studies.

one cannot have confidence that the bias would be eliminated, or even mitigated, by having included these other factors in the epidemiological regressions.

(III.B.) The magnitude of the $PM_{2.5}$ association at varying $PM_{2.5}$ exposure levels (i.e., the "shape" of the relationship) is estimated with error.

Brauer *et al.* (2002), among others, have demonstrated that measurement error for personal exposures when using central monitor data can hide a threshold from the statistical methods, even when one exists in reality.[20] Thus EPA's assumption of a linear relationship is not valid for quantitative risk analysis, even if the estimated constant relative risk were quantitatively valid as an average over the range of observed exposure levels.

EPA does not consider concerns with measurement error at all when it concludes that a linear relationship "most adequately" represents the association for purposes of statistical fit. When the Draft ISA states that "the C-R curve was found to be indistinguishable from linear,"[21] it is only making a statement about statistical goodness of fit to the available data. These statistical tests offer no information about whether the shape of the underlying true relationship has been obscured by measurement error, yet this is the critical question when trying to develop a quantitative estimate of risk from statistical associations using messy data.[22] The risk quantification step of the PMRA requires a functional relationship that properly reflects the true shape of the underlying relationship in order to reliably predict the changes in risk that would result from changes in $PM_{2.5}$. If a risk relationship has a non-linear shape (as one would logically expect for a true concentration-response situation, given a normally distributed degree of sensitivity across a population), then it is not appropriate to simply use a linear relationship just because the available epidemiological data do not offer the sensitivity necessary to detect that shape.

Errors due to an incorrect shape of the concentration-response function will become more and more pronounced with rollbacks to increasingly lower levels of $PM_{2.5}$ because the amount of change in risk associated with increasingly lower $PM_{2.5}$ exposures may

[20] Smith and Chan (1997) also demonstrated the impossibility of statistically detecting a real threshold in the presence of exposure measurement error. For simulated data with a pronounced ("hockey-stick") threshold at 20, the best fit for alternative thresholds was no threshold at all. They also fit a nonparametrically smoothed curve to the simulated data, with the resulting estimated relationship appearing to have the opposite of a threshold, that is, that the estimated concentration-response curve became *steeper* at concentrations closer to zero (and well below the point of the actual threshold). See Figure 9 and associated discussion in Smith and Chan (1997), p. 14, for the nonparametrically smoothed estimate of the concentration-response curve.

[21] Draft ISA, p. 2-37.

[22] Even the conclusion that linearity is the best *statistical fit* remains a debatable conclusion: see, for example, Abrahamowicz *et al.* (2003), and Goodman (2009), pp. 21-22. Without even considering debates about statistical tests of nonlinearity, potential evidence of nonlinearity can be found in the extended cohort analyses. For example, the finding reported by Laden *et al.* (2006) that estimated relative risks were lowered as $PM_{2.5}$ levels fell over time implies a non-linear relationship, while Gamble and Nicolich (2006) argue that a non-linear relationship may be observable even in the relative risks for a single time period. Also, the ACS evidence discussed in II.D that recent $PM_{2.5}$ levels used to estimate long-term associations could be serving as a proxy for earlier, higher $PM_{2.5}$ exposures also implies a non-linear actual relationship.

become vanishingly small, while the presumed linear statistical association assumes equally large amounts of risk reduction for a unit of improvement in $PM_{2.5}$, whether that change occurs from high levels of as-is exposure, or from near-background levels. The practice in the Draft PMRA of not counting risks below the lowest measured level (LML) of $PM_{2.5}$ does not eliminate this quantitative error. In fact, this practice exacerbates the misleading nature of the PMRA because it produces the bizarre effect of suggesting that there is no threshold, y*et al*so that 100% of the currently existing risk attributed to $PM_{2.5}$ would be eliminated if $PM_{2.5}$ is rolled-back only as far as the LML.

(Setting aside the logical inconsistency, if one truly believes that a linear relationship can be assumed, then the Draft PMRA is misleading when it reports that nearly 100% the long-term $PM_{2.5}$ risk can be eliminated by tightening the standard for $PM_{2.5}$ to 12 µg/m^3 annual average. All EPA can really say is that the latter standard would reduce $PM_{2.5}$ to the lowest levels that were observed in the most recent years among the cities included in the ACS cohort studies.)

IV. The Epidemiological Evidence Itself Indicates a Meaningful Chance that Long-Term $PM_{2.5}$ Exposures Do Not Elevate Public Health Risk.

Sections II and III have provided multiple reasons why quantitative estimates of risk in the Draft PMRA are unreliable. The Draft PMRA uses estimates of relative risks from the epidemiological studies at "face value" for its quantitative concentration-response functions, ignoring the many likely sources of bias those estimates. At a minimum, this results in large errors in its estimates of risk and risk reductions from reducing ambient $PM_{2.5}$ exposures that are not captured in statistical confidence intervals. However, EPA also commits an error of omission in its Draft PMRA by relying selectively on a few relative risk estimates from a few studies. This hides the degree of uncertainty that is detectable in the full body of epidemiological evidence, even if taken "at face value."

Even if one accepts the existing body of long-term epidemiological relative risk estimates at face value, one can observe a substantial chance that no effect exists at all if the full set of available relative risk estimates is given fair consideration. Figure 7-6 of the Draft ISA (p. 7-123) shows a very selective set of results. However, the entire body of evidence includes past as well as current studies. Even earlier estimates from the ACS studies remain relevant to the extent that certain regressions in them have not been reproduced in more recent ACS analyses. (The results that include SO_2 as well as $PM_{2.5}$ in the regressions from Krewski *et al.*, 2000, are a case in point.) Other studies also not shown in Figure 7-6 have produced non-positive and/or no statistically significant association between $PM_{2.5}$ and chronic mortality.[23] Nothing in the more recent literature

[23] Lipfert *et al.* (2000) is an example that found a negative association. More recent papers for the Veterans cohort (*e.g.*, Lipfert, Wyzga, *et al.*, 2006) have reported positive $PM_{2.5}$-risk associations (although significant only in 1-P formulations) but these newer findings do not make the earlier findings irrelevant. The earlier findings used a different subset of the Veteran's Cohort than the later findings, where the subset was driven by the locations of the available pollution data being used. The earlier studies also considered mortality risk in a different (earlier) time period. Thus, one can find both negative and positive associations within this single cohort, depending solely on the time period and air pollution data used.

necessarily supersedes those other studies. When the entire set of regressions are considered, giving equal weights to single and multipollutant models, and giving equal weights to the various cohorts that have been studied, one finds that this literature *as a whole* suggests that there is about a 15-20% chance that there is zero risk from $PM_{2.5}$.[24] This reflects the fact that the literature does contain quite a few findings of insignificance, which imply, based on statistical error alone, that the effect is zero. (In fact, it implies a possibility of a negative effect, but for this discussion, those are considered zero, not beneficial, effects.)

V. Conclusion

EPA's criteria for determining whether a causal relationship exists between exposures to $PM_{2.5}$ and health endpoints are fundamentally flawed because they allow an incorrect determination of causality to be made in circumstances that are marked by systematic biases. The epidemiological literature for long-term $PM_{2.5}$ exposure risks is clearly open to the possibility, and even likelihood, of systematic bias. Therefore, EPA's "causal" determination for long-term cardiovascular mortality risk and its "likely-causal" determination for long-term all-cause mortality risk in the Draft ISA are subject to a much greater chance of being wrong than the words themselves suggest. This fact alone places significant limitations on the usefulness and reliability of EPA's quantitative estimates of long-term mortality risks in its Draft PMRA, because the entire risk assessment is predicated on an unquestioned *presumption* of causality.

The quantitative estimates of mortality risk in the Draft PMRA remain unreliable, however, even if it is correct that there is some causal relationship between $PM_{2.5}$ and risk of dying. This is because the Draft PMRA also presumes that the statistical associations in the epidemiological literature can be interpreted *literally* as the actual concentration-response relationships for quantifying current levels of risk, and changes in risks for altered ambient $PM_{2.5}$ conditions.

The translation from an epidemiologically-derived association to a real "concentration-response function" that quantifies how much risk would change if PM mass were changed is highly problematic, regardless of the quality of the epidemiological studies that are being relied on. Even if one has great confidence that an association between $PM_{2.5}$ and health risk detected in an epidemiological study is reflecting a true causal relationship, the statistical model and its parameter estimates (*e.g.*, the "relative risk") cannot be assumed to be a precise numerical estimate of the true causal relationship, given the many limitations of the available data. As explained above, there remain good reasons to suspect that some or all of the estimated association bears no causal implication for $PM_{2.5}$ mass itself, or even of one of its constituents. If there are any missing explanatory variables – which is almost certainly true – then the statistical estimate of relative risk is not quantitatively reliable to assess either "as-is" risks, or

[24] This includes consideration of relative risks in Eftim *et al.* (2008), Enstrom (2005), Jerrett *et al.* (2005), Krewski *et al.* (2000, 2009), Laden *et al.* (2006) Lipfert *et al.* (2002), Lipfert, Baty *et al.* (2006), Lipfert, Wyzga *et al.* (2006), McDonnell *et al.* (2000), Pope *et al.* (2002), and Villeneuve *et al.* (2002).

changes in risk as $PM_{2.5}$ is reduced. The kinds of measurement errors and confounding that are present in the $PM_{2.5}$ epidemiological data also mean that the shape of the true relationship cannot be identified. The inability to define the shape of the true relationship means that one can have no confidence in statements of how risk will change as $PM_{2.5}$ is reduced.

The statistical confidence intervals that the Draft PMRA offers up as "uncertainty" do not measure biases due to missing variables or measurement error, and thus do not offer a way of characterizing the numerical uncertainty in actual risk levels at current or rolled-back $PM_{2.5}$ levels. The Draft PMRA's sensitivity analyses, which simply substitute one statistical estimate of relative risk for another, also cannot begin to characterize the quantitative uncertainty, given that all the available epidemiological estimates suffer from the same limitations in data and methods, and are thus subject to a systematic bias. As a result, the numerical estimates provided in the Draft PMRA have no reliable relationship to reality, even if one accepts the presumption that the epidemiological studies are detecting a causal relationship between one or more constituents of $PM_{2.5}$ and health risk. With all of these unstated and unanalyzed presumptions, one can have no confidence in the estimates of how much health risk is being created by current $PM_{2.5}$ exposures, nor whether any of that estimated risk would be reduced by reducing an undifferentiated measure of total $PM_{2.5}$ mass. The Draft PMRA's quantitative estimates of risk from as-is $PM_{2.5}$, and quantitative estimates of reductions in due to lowered $PM_{2.5}$, are unreliable.

Until it contains a more explicit analysis of the quantitative implications of these inherent challenges for estimating risks under current and alternative ambient $PM_{2.5}$ standards, the quantitative risk assessment of the Draft PMRA is, at best, not useful; at worst, its results are highly misleading as an input to policy decisions for setting a NAAQS. The only way to obtain reliable estimates would be to find ways to quantitatively incorporate corrections for systematic biases due to differential measurement errors, and potentially unmeasured causal confounders of a non-pollutant nature. This would produce a larger range of uncertainty, but one that reflects the true current state of knowledge. If this is not done, however, then the Draft PMRA should not be used in the consideration of alternative $PM_{2.5}$ NAAQS.

References

Abrahamowicz M, Schopflocher T, Leffondre K, du Berger R, and Krewski D. 2003. "Flexible Modeling of Exposure-Response Relationship between Long-Term Average Levels of Particulate Air Pollution and Mortality in the American Cancer Society Study." *J. Toxicology and Environmental Health*, Part A, 66:1625-1653.

Beelen R, Hoek G, van den Brand P, Goldbohm R, Fischer P, Schouten L, Jerrett M, Hughes E, Armstrong B, Brunekreef B. 2008a. "Long-term effects of traffic-related air pollution on mortality in a Dutch Cohort (NLCS-Air Study)," *Environmental Health Perspectives* 166(2):196-202.

Beelen R, Hoek G, Houthuijs D, van den Brandt PA, Goldbohm RA, Fischer P, Schouten LJ, Armstrong B, Brunekreef B. 2008b. “The Joint Association of Air Pollution and Noise from Road Traffic with Cardiovascular Mortality in a Cohort Study,” *Occup. Environ. Med.* (published online, November 18).

Boffeta P, McLaughlin, JK, LaVecchia C, Tarone RE, Lipworth L, Blot WJ. 2008. “False Positive Results in Cancer Epidemiology: A Plea for Epistemological Modesty,” *J. of the National Cancer Institute* 100:988-995.

Brauer M, Brumm J, Vedal S, Petkau AJ. 2002. “Exposure Misclassification and Threshold Concentrations in Time Series Analyses of Air Pollution Health Effects,” *Risk Analysis* 22(6):1183-1193.

Brunekreef B, Beelen R, Hoek G, Schouten L, Bausch-Goldbohm S, Fischer P, Armstrong B, Hughes E, Jerrett M, van den Brandt P. 2009. *Effects of Long-term Exposure to Traffic-related Air Pollution on Respiratory and Cardiovascular Mortality in the Netherlands: The NLCS-AIR Study*, Research Report #139, Health Effects Institute, Boston, Massachusetts.

Bukowski J. 2008. “Do Pollution Time-Series Studies Contain Uncontrolled or Residual Confounding by Risk Factors for Acute Health Events?” *Regulatory Toxicology and Pharmacology* 51:135-140.

Eftim SE, Samet JM, Janes H, McDermott A, Dominici F. 2008. “Fine Particulate Matter and Mortality: A Comparison of the Six Cities and American Cancer Society Cohorts with a Medicare Cohort,” *Epidemiology* 19(2):209-216.

Enstrom J. 2005. “Fine Particulate Air Pollution And Total Mortality among Elderly Californians, 1973-2002,” *Inhalation Toxicology* 17:803-816.

Fewell Z, Davey Smith G, Stern JA. 2007. “The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study,” *Am. J. Epidemiol*. 166(6): 646-655.

Gamble J, Nicolich M. 2006. “Comments on the updated Harvard Six Cites Study,” Correspondence, *American J. Respiratory and Critical Care Medicine* 174:722-724.

Goodman J. 2009. “Comments on the Epidemiology Evaluation in the Integrated Science Assessment for Particulate Matter Second External Review Draft,” Gradient Corporation, on behalf of American Petroleum Institute (October 13).

Jerrett M, Burnett RT, Ma R, Pope CA, Krewski D, Newbold KB, Thurston G, Shi Y, Finkelstein N, Calle EE, Thun, MJ. 2005. “Spatial Analysis of Air Pollution and Mortality in Los Angeles,” *Epidemiology* 16(6):1-10.

Krewski D, Burnett RT, Goldberg M, Hoover K, Siemiatycki J, Jerrett M, Abrahamowicz M, White WH. 2000. *Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality*. Special Report, Health Effects Institute, Cambridge, Massachusetts (July).

Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi L, Turner MC, Pope CA, Thurston G, Calle EE, Thun MJ. 2009. *Extended Follow-up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality*. Research Report #140, Health Effects Institute, Boston, Massachusetts (May).

Laden F, Schwartz J, Speizer FE, Dockery DW. 2006. "Reduction in Fine Particulate Air Pollution and Mortality, Extended Follow-up of the Harvard Six Cities Study," *Am J Respir Crit Care Med*. 173:667-672.

Lipfert FW, Baty JD, Miller JP, Wyzga RE. 2006. "$PM_{2.5}$ Constituents and Related Air Quality Variables as Predictors of Survival in a Cohort of U.S. Military Veterans," *Inhalation Toxicology* 18:645-657.

Lipfert F, Perry HM Jr., Miller JP, Baty JD, Wyzga RE, Carmody SE. 2000. "The Washington University-EPRI Veterans' Cohort Mortality Study: Preliminary Results," *Inhalation Toxicology* 12(Supplement 4):41-73.

Lipfert FW, Wyzga RE, Baty JD, Miller JP. 2006. "Traffic Density as a Surrogate Measure of Environmental Exposures in Studies of Air Pollution Health Effects: Long-term Mortality in a Cohort of U.S. Veterans," *Atmospheric Environment* 40:154-169.

McConnell WF, Nishino-Ishikawa N, Petersen FF, Chen LH, Abbey DE. 2000. "Relationships of Mortality with the Fine and Coarse Fractions of Long-Term Ambient PM_{10} Concentrations in Nonsmokers," *Journal of Exposure Analysis and Environmental Epidemiology* 10:427-436.

Pope CA, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. 2002. "Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution," *JAMA* 287(9):11332-1141.

Smith AE, Chan NY. 1997. "How Statistics Can Mislead PM Policy: A Case of Smoke and Mirrors?" Attachment 4 to Comments of the Utility Air Regulatory Group on the National Ambient Air Quality Standards for Particulate Matter: Proposed Decision, 61 Fed. Reg. 65,638 (December 13, 1996) (Docket A-95-54) and Proposed Requirements for Designation of Reference and Equivalent Methods for $PM_{2.5}$ and Ambient Air Quality Surveillance for Particulate Matter: Proposed Rule, 61 Fed. Reg. 65,780 (December 13).

Villeneuve PJ, Goldberg MS, Krewski D, Burnett RT, Chen Y. 2002. "Fine Particulate Air Pollution and All-Cause Mortality within the Harvard Six-Cities Study: Variations in Risk by Period of Exposure," *Ann. Epidemiol.* 12(8):568-576.