Comments on EPA's "Integrated Review Plan for the Review of the Ozone National Ambient Air Quality Standards, External Review Draft"

Docket ID: EPA-HQ-OAR-2018-0279

Julie E. Goodman, Ph.D., DABT, FACE, ATS
Gradient

Thank you for the opportunity to speak today regarding the draft Ozone Integrated Review Plan (IRP). I am an epidemiologist and board-certified toxicologist at Gradient, an environmental consulting firm. I am speaking on behalf of Gradient, but my time spent preparing these comments and calling into this teleconference has been funded by the American Petroleum Institute.

The IRP presents the Integrated Science Assessment (ISA) development process and the overarching plan for the ISA analysis; this is described in greater detail in the Preamble to the ISAs, which is intended to serve as a companion document to ISAs. The Preamble does a good job providing an outline of the general process used by EPA for developing the ISAs, including the overall framework for evaluating evidence and making causal conclusions.

My major concern with the draft IRP, as well as the Preamble, is that neither provide specific instructions for a thorough, systematic, and reproducible analysis. The Preamble should be thought of as a standard operating procedure (SOP) for ISAs, and the ozone IRP should be thought of as the pollutant-specific protocol for the ISA and other assessments. In both cases, step-by-step instructions are necessary. As they currently stand, these documents do not have sufficient detail to ensure that studies are identified and reviewed in a systematic and consistent manner or integrated in a way that considers study quality and the coherence of results across studies, both within and across disciplines. For comparison, SOPs and protocols for experimental studies require a high level of detail to be sure the studies are conducted in an unbiased manner and can be replicated by others. SOPs and protocols for systematic reviews should be thought of the same way.

In the same vein, the IRP should list quality metrics that are specific to the realm of evidence (*e.g.*, epidemiology, controlled exposure studies, animal toxicity studies), study design (*e.g.*, case- control, cohort), exposure metrics (*e.g.*, hourly, daily, annual), and outcome (*e.g.*, asthma exacerbation, cancer). Criteria should not—and cannot—be the same for all study designs. The IRP should explicitly describe the criteria for evaluating the studies that will be included in the ISA, so that the public and CASAC can review the criteria and make recommendations.

EPA has tables with lists of study quality criteria in the NO_x ISA and draft PM ISA (I have attached the table from the draft PM ISA to these comments). These provide a good overview of quality issues that should be addressed but are not complete and do not address *in vitro* studies, and the level of detail varies. Still, these lists could serve as an excellent start for one that could be used in the ozone IRP, and I have described ways these quality criteria can be improved in my written comments. Having detailed study quality criteria in the ozone IRP will allow for a more comprehensive and objective assessment of study quality and will ensure

that key studies represent the most reliable available evidence.

My final, yet very important, comment on the plan for the ISA in the IRP is that the causal framework in the IRP, which has been used in several previous ISAs, is biased towards casual conclusions. This framework is a modified version of the Institute of Medicine (IOM) Framework. The original IOM framework should be used to ensure evidence is reviewed in a robust, systematic, consistent manner, using well-specified criteria, so that causality can be objectively assessed.

With respect to a potential REA, the IRP provides very little detail. The IRP should explain how it will determine whether there is a need for an REA in this review cycle and whether new information, tools, or methods address uncertainties discussed in the last review. It should also properly characterize the uncertainty surrounding controlled exposure studies, and all aspects of uncertainty should be quantitatively addressed in risk estimates, if possible. Finally, the IRP should indicate that the REA will consider a threshold concentration-response function when modeling the risk of ozone-induced health effects, if data allow.

In closing, CASAC should recommend the use of a standard causality framework. Further, CASAC should require EPA to provide a more detailed, objective, and transparent methodology in the ozone IRP, building on what is found in the NO_x ISA, highlighting any information that informs the need for an updated REA, and listing new approaches the could be employed. This will limit potential biases in the ISA and REA, help ensure that the analyses can be reproduced by others, and increase confidence in the causal determinations regarding ozone-associated effects.

Scientific Considerations for Evaluating the Strength of Inference from Studies on the Health Effects of Particulate Matter

Study Design
Controlled Human Exposure
Studies should clearly describe the primary and any secondary objectives of the study, or specific hypotheses being tested. Study subjects should be randomly exposed without knowledge of the exposure condition. Preference is given to balanced crossover (repeated measures) or parallel design studies which include control exposures (e.g., to clean filtered air). In crossover studies, a sufficient and specified time between exposure days should be provided to avoid carry over effects from prior exposure days. In parallel design studies, all arms should be matched for individual characteristics such as age, sex, race, anthropometric properties, and health status. In studies evaluating effects of disease, appropriately matched healthy controls are desired for interpretative purposes.
Animal Toxicology
Studies should clearly describe the primary and any secondary objectives of the study, or specific hypotheses being tested. Studies should include appropriately matched control exposures (e.g., to clean filtered air, time matched). Studies should use methods to limit differences in baseline characteristics of control and exposure groups. Studies should randomize assignment to exposure groups and where possible conceal allocation to research personnel. Groups should be subjected to identical experimental procedures and conditions; animal care including housing, husbandry, etc. should be identical between groups. Blinding of research personnel to study group may not be possible due to animal welfare and experimental considerations; however, differences in the monitoring or handling of animals in all groups by research personnel should be minimized.
Epidemiology
Inference is stronger for studies that clearly describe the primary and any secondary aims of the study, or specific hypotheses being tested. For short-term exposure, time-series, case crossover, and panel studies are emphasized over cross-sectional studies because they examine temporal correlations and are less prone to confounding by factors that differ between individuals (e.g., SES, age). Panel studies with scripted exposures, in particular, can contribute to inference because they have consistent, well-defined exposure durations across subjects, measure personal ambient pollutant exposures, and measure outcomes at consistent, well-defined lags after exposures. Studies with large sample sizes and conducted over multiple years are considered to produce more reliable results. Additionally, multi-city studies are preferred over single-city studies because they examine associations large diverse geographic areas using a consistent statistical methodology, avoiding the publication bias often associated with single-city studies[a]. If other quality parameters are equal, multicity studies carry more weight than single-city studies because they tend to have larger sample sizes and lower potential for publication bias. For long-term exposure, inference is considered to be stronger for prospective cohort studies and case-control studies nested within a cohort (e.g., for rare diseases) than cross-sectional, other case-control, or ecologic studies. Cohort studies can better inform the temporality of exposure and effect. Other designs can have uncertainty related to the appropriateness of the control group or validity of inference about individuals from group-level data. Study design limitations can bias health effect associations in either direction.

Study Population/Test Model
Controlled Human Exposure
In general, the subjects recruited into study groups should be similarly matched for age, sex, race, anthropometric properties, and health status. In studies evaluating effects of specific subject characteristics (e.g., disease, genetic polymorphism, etc.), appropriately matched healthy controls are preferred. Relevant characteristics and health status should be reported for each experimental group. Criteria for including and excluding subjects should be clearly indicated. For the examination of populations with an underlying health condition (e.g., asthma), independent, clinical assessment of the health condition is ideal, but self-report of physician diagnosis generally is considered to be reliable for respiratory and cardiovascular disease outcomes[b]. The loss or withdrawal of recruited subjects during the course of a study should be reported. Specific rationale for excluding subject(s) from any portion of a protocol should be explained.
Animal Toxicology
Ideally, studies should report species, strain, substrain, genetic background, age, sex, and weight. Unless data indicate otherwise, all animal species and strains are considered appropriate for evaluating effects of PM exposure. It is preferred that the authors test for effects in both sexes and multiple lifestages, and report the result for each group separately. All animals used in a study should be accounted for, and rationale for exclusion of animals or data should be specified.
Epidemiology
There is greater confidence in results for study populations that are recruited from and representative of the target population. Studies with high participation and low drop-out over time that is not dependent on exposure or health status are considered to have low potential for selection bias. Clearly specified criteria for including and excluding subjects can aid assessment of selection bias. For populations with an underlying health condition, independent, clinical assessment of the health condition is valuable, but self-report of physician diagnosis generally is considered to be reliable for respiratory and cardiovascular diseases[b]. Comparisons of groups with and without an underlying health condition are more informative if groups are from the same source population. Selection bias can influence results in either direction or may not affect the validity of results but rather reduce the generalizability of findings to the target population.
Pollutant
Controlled Human Exposure
Studies should: (1) include a composite measure of PM (i.e., $PM_{2.5}$, $PM_{10-2.5}$, or ultrafine particles [UFP][c]) or (2) apply some approach (e.g., particle trap or filter) to assess the effects of PM in a complex air pollution mixture (i.e., diesel exhaust, gasoline exhaust, wood smoke).
Animal Toxicology
Studies should: (1) include a composite measure of PM (i.e., $PM_{2.5}$, $PM_{10-2.5}$, or ultrafine particles [UFP][c]) or (2) apply some approach (e.g., particle trap or filter) to assess the effects of PM in a complex air pollution mixture (i.e., diesel exhaust, gasoline exhaust, wood smoke).
Epidemiology
Health effects are evaluated primarily using a composite measure of PM (i.e., $PM_{2.5}$, $PM_{10-2.5}$, or ultrafine particles [UFP][c]) from studies using ambient measurements, model predictions, or a combination of measured and modeled data. Studies of PM components must also include a composite measure of PM. Studies of source-related indicators are also evaluated where the indicator is derived using ambient PM concentrations.

Exposure Assessment or Assignment
Controlled Human Exposure
For this assessment, the focus is on studies that utilize PM concentrations <2 mg/m^3. Studies that use higher exposure concentrations may provide information relevant to biological plausibility, dosimetry, or inter-species variation. Studies should have well-characterized pollutant concentration, temperature, and relative humidity and/or have measures in place to adequately control the exposure conditions. Preference is given to balanced crossover or parallel design studies which include control exposures (e.g., to clean filtered air). Study subjects should be randomly exposed without knowledge of the exposure condition. Method of exposure (e.g., chamber, facemask, etc.) should be specified and activity level of subjects during exposures should be well characterized.
Animal Toxicology
For this assessment, the focus is on studies that utilize PM concentrations <2 mg/m^3. Studies that use higher exposure concentrations may provide information relevant to biological plausibility, dosimetry, or inter-species variation. Studies should characterize pollutant concentration, temperature, and relative humidity and/or have measures in place to adequately control the exposure conditions. The focus is on inhalation exposure. Non-inhalation exposure experiments (i.e., intratracheal instillation [IT]) are informative for size fractions (e.g., $PM_{10-2.5}$) that cannot penetrate the airway of a study animal and may provide information relevant to biological plausibility and dosimetry. In vitro studies may be included if they provide mechanistic insight or examine similar effects as in vivo studies, but are generally not included. All studies should include exposure control groups (e.g., clean filtered air).
Epidemiology
Of primary relevance are relationships of health effects with the ambient component of PM exposure. However, information about ambient exposure rarely is available for individual subjects; most often, inference is based on ambient concentrations. Studies that compare exposure assessment methods are considered to be particularly informative. Inference is stronger when the duration or lag of the exposure metric corresponds with the time course for physiological changes in the outcome (e.g., up to a few days for symptoms) or latency of disease (e.g., several years for cancer). Given that the spatial variability of PM composite measures varies among size fractions, with more homogeneity for $PM_{2.5}$ than either $PM_{10-2.5}$ or UFP, the need for capturing spatial contrasts is stronger for $PM_{10-2.5}$ or UFP compared with $PM_{2.5}$. Validated measurements, whether averaged across multiple monitors or assigned from the nearest or single available monitor, adequately capture temporal or spatial variation in exposure to $PM_{2.5}$ due to the high correlation between personal exposure and ambient concentration. However, for more spatially heterogeneous $PM_{10-2.5}$ and UFP, the spatial correlation between personal exposure and ambient concentrations is lower. Similarly, PM components show increased spatial variability relative to $PM_{2.5}$. In this case, validated methods that capture the extent of variability for the particular study design (temporal vs. spatial contrasts) and location carry greater weight. Inference based on central site measurements can be adequate if correlated with personal exposures, closely located to study subjects, highly correlated across monitors within a location, used in locations with well-distributed sources, or combined with time-activity information. In studies of short-term exposure, temporal variability of the exposure metric is of primary interest. For all PM size fractions, studies that incorporate time-activity data with personal or microenvironmental monitoring or modeling data may carry greater weight because residential, in-vehicle, and workplace PM exposures may differ in their temporal variability. Results for total personal and indoor PM exposure are other lines of evidence that may inform judgments about causality of PM because inference is based on an individual's microenvironmental exposures and the potential for copollutant confounding may be reduced compared to ambient exposures. Results for total personal exposure can inform understanding of the effects of ambient exposure when well correlated with ambient concentrations. For long-term exposures, methods that well represent within-community spatial variation in individual exposure may be given more weight for spatially-variable ambient $PM_{10-2.5}$ or ultrafine particles. For $PM_{2.5}$, within-community variation in exposure is less important given that $PM_{2.5}$ tends to be more homogeneous. Exposure measurement error often attenuates health effect estimates or increases the imprecision of the association (i.e., width of 95% CIs), particularly associations based on temporal variation in short-term exposure. However, exposure measurement error can bias estimates away from the null in some epidemiologic studies of long-term exposures where the PM size fraction is more spatially heterogeneous (i.e., $PM_{10-2.5}$ or UFP), depending on the locations of the monitor and sources with respect to the study population. To streamline the health effects discussion on studies that are most policy-relevant, for those health categories where the 2009 PM ISA concluded a "causal relationship" the focus is on studies with mean $PM_{2.5}$ concentrations <20 µg/m^3. However, studies that examine a previously identified uncertainty or limitation in the evidence are evaluated even if mean $PM_{2.5}$ concentrations are >20 µg/m^3.

Outcome Assessment/Evaluation

Controlled Human Exposure

Endpoints should be assessed in the same manner for control and exposure groups (e.g., time after exposure, methods, endpoint evaluator) using valid, reliable methods. Blinding of endpoint evaluators is ideal, especially for qualitative endpoints (e.g., histopathology). For each experiment and each experimental group, including controls, precise details of all procedures carried out should be provided including how, when, and where. Time of the endpoint evaluations is a key consideration that will vary depending on endpoint evaluated. Endpoints should be assessed at time points that are appropriate for the research questions.

Animal Toxicology

Endpoints should be assessed in the same manner for control and exposure groups (e.g., time after exposure, methods, endpoint evaluator) using valid, reliable methods. Blinding of endpoint evaluators is ideal, especially for qualitative endpoints (e.g., histopathology). For each experiment and each experimental group, including controls, precise details of all procedures carried out should be provided including how, when, and where. Time of the endpoint evaluations is a key consideration that will vary depending on endpoint evaluated. Endpoints should be assessed at time points that are appropriate for the research questions.

Epidemiology

Inference is stronger when outcomes are assessed or reported without knowledge of exposure status. Knowledge of exposure status could produce artefactual associations. Confidence is greater when outcomes assessed by interview, self-report, clinical examination, or analysis of biological indicators are defined by consistent criteria and collected by validated, reliable methods. Independent, clinical assessment is valuable for outcomes such as lung function or incidence of disease, but report of physician diagnosis has shown good reliability[b]. When examining short-term exposures, evaluation of the evidence focuses on specific lags based on the evidence presented in individual studies. Specifically, the following hierarchy is used in the process of selecting results from individual studies to assess in the context of results across all studies for a specific health effect or outcome:

- Distributed lag models;
- Average of multiple days (e.g., 0−2);
- If a priori lag days were used by the study authors these are the effect estimates presented; or
- If a study focuses on only a series of individual lag days, expert judgment is applied to select the appropriate result to focus on considering the time course for physiologic changes for the health effect or outcome being evaluated.

When health effects of long-term exposure are assessed by acute events such as symptoms or hospital admissions, inference is strengthened when results are adjusted for short-term exposure. Validated questionnaires for subjective outcomes such as symptoms are regarded to be reliable[c], particularly when collected frequently and not subject to long recall. For biological samples, the stability of the compound of interest and the sensitivity and precision of the analytical method is considered. If not based on knowledge of exposure status, errors in outcome assessment tend to bias results toward the null.

Potential Copollutant Confounding

Controlled Human Exposure

Exposure should be well characterized to evaluate independent effects of PM of various size fractions. Studies should apply some approach (e.g., particle trap or filter) to assess the effects of PM when examining exposures to complex air pollution mixtures (i.e., diesel exhaust, gasoline exhaust, wood smoke).

Animal Toxicology

Exposure should be well characterized to evaluate independent effects of PM of various size fractions. Studies should apply some approach (e.g., particle trap or filter) to assess the effects of PM when examining exposures to complex air pollution mixtures (i.e., diesel exhaust, gasoline exhaust, wood smoke).

Epidemiology

Not accounting for potential copollutant confounding can produce artefactual associations; thus, studies that examine copollutant confounding carry greater weight. The predominant method is copollutant modeling (i.e., two-pollutant models), which is especially informative when correlations are not high. However, when correlations are high ($r > 0.7$), such as those often encountered for UFP and other traffic-related copollutants, copollutant modeling is less informative. Although the use of single-pollutant models to examine the association between PM and a health effect or outcome are informative, ideally studies should also include copollutant analyses. Copollutant confounding is evaluated on an individual study basis considering the extent of correlations observed between the copollutant and PM, and relationships observed with PM and health effects in copollutant models.

Other Potential Confounding Factors[d]
Controlled Human Exposure
Preference is given to studies utilizing experimental and control groups that are matched for individual level characteristics (e.g., race/ethnicity, sex, body weight, smoking history, age) and time varying factors (e.g., seasonal and diurnal patterns).
Animal Toxicology
Preference is given to studies utilizing experimental and control groups that are matched for individual level characteristics (e.g., strain, sex, body weight, litter size, food and water consumption) and time varying factors (e.g., seasonal and diurnal patterns).
Epidemiology
Factors are considered to be potential confounders if demonstrated in the scientific literature to be related to health effects and correlated with PM. Not accounting for confounders can produce artefactual associations; thus, studies that statistically adjust for multiple factors or control for them in the study design are emphasized. Less weight is placed on studies that adjust for factors that mediate the relationship between PM and health effects, which can bias results toward the null. Confounders vary according to study design, exposure duration, and health effect and may include, but are not limited to the following: Short-term exposure studies: Meteorology, day of week, season, medication use, allergen exposure, and long-term temporal trends. Long-term exposure studies: Socioeconomic status, race, age, medication use, smoking status, stress, noise, and occupational exposures.
Statistical Methodology
Controlled Human Exposure
Statistical methods should be clearly described and appropriate for the study design and research question (e.g., correction for multiple comparisons). Generally, statistical significance is used to evaluate the findings of controlled human exposure studies. However, consistent trends are also informative. Detection of statistical significance is influenced by a variety of factors including, but not limited to, the size of the study, exposure and outcome measurement error, and statistical model specifications. Sample size is not a criterion for exclusion; ideally, the sample size should provide adequate power to detect hypothesized effects (e.g., sample sizes less than 3 are considered less informative). Because statistical tests have limitations, consideration is given to both trends in data and reproducibility of results.
Animal Toxicology
Statistical methods should be clearly described and appropriate for the study design and research question (e.g., correction for multiple comparisons). Generally, statistical significance is used to evaluate the findings of animal toxicology studies. However, consistent trends are also informative. Detection of statistical significance is influenced by a variety of factors including, but not limited to, the size of the study, exposure and outcome measurement error, and statistical model specifications. Sample size is not a criterion for exclusion; ideally, the sample size should provide adequate power to detect hypothesized effects (e.g., sample sizes less than 3 are considered less informative). Because statistical tests have limitations, consideration is given to both trends in data and reproducibility of results.
Epidemiology
Multivariable regression models that include potential confounding factors are emphasized. However, multipollutant models (more than two pollutants) are considered to produce too much uncertainty due to copollutant collinearity to be informative. Models with interaction terms aid in the evaluation of potential confounding as well as effect modification. Sensitivity analyses with alternate specifications for potential confounding inform the stability of findings and aid in judgments of the strength of inference from results. In the case of multiple comparisons, consistency in the pattern of association can increase confidence that associations were not found by chance alone. Statistical methods that are appropriate for the power of the study carry greater weight. For example, categorical analyses with small sample sizes can be prone to bias results toward or away from the null. Statistical tests such as *t*-tests and Chi-squared tests are not considered sensitive enough for adequate inferences regarding PM-health effect associations. For all methods, the effect estimate and precision of the estimate (i.e., width of 95% CI) are important considerations rather than statistical significance.

Source: US EPA, National Center for Environmental Assessment (NCEA) - RTP Division. October 2018. "Integrated Science Assessment for Particulate Matter (External Review Draft)." EPA/600/R-18/179. 1881p., Table A-1