Professional Statistician response to the CSAC Review Panel

February 18, 2022

I am Robert Obenchain of Clayton, CA 94517. I hold a Ph.D. in Statistics (1969) and was elected a Fellow of the American Statistical Association in 1997. I have designed and analyzed statistical studies at AT&T Bell Labs/Bell Communications Research (1970-1986) and at Eli Lilly and Company, Health Outcomes Research (1990-2007). At Lilly, one of my specialties was development of analysis strategies for "observational" data to provide valid Real-World Evidence.

I think that environmental legislation should be based upon <u>unquestionably</u> <u>valid and unbiased scientific studies</u>. That is, studies where appropriately de-identified data have been made available for re-analysis by qualified analysts representing dissenting perspectives.

Three of my concerns with the current Clean Air Scientific Advisory Committee (CASAC) Particulate Matter Review Panel are outlined here:

- 1. Recently, I have been concerned about whether CASAC will embrace the admirable scientific approach endorsed by the Whitehouse OSTP in January 2022. The EPA and the academic researchers they fund rarely cite or comment upon published studies that disagree with their findings and/or with EPA policy. Good science would be much more "open for appropriate discussion" than this.
- 2. In the long term, I think that the CASAC PM-panel would greatly benefit from more "public" oversight of EPA sponsored studies. Obviously, this would require (de-identified) data used in individual studies to be make "public." Without data that are "public," how can scientific discussion be truly "open"?
 - a. Effective de-identification of data can be as simple as reporting only the average value from "more than 10 similar subjects"; this has been a <u>CDC rule</u> for **publication** of "outcomes" for many years.
 - b. My free R-package of functions for de-identification of cross-sectional data using "micro-aggregation" is available from: <u>https://CRAN.R-project.org/package=LocalControlStrategy</u>.

3. Finally, I am currently having an unpleasant experience requesting deidentified (i.e. **publishable**) data from a paper on Secondary Organic Aerosols (SOAs) published in *Nature Communications* and authored by **EPA employees**. Thus, this publication is (or should be) subject to the U.S. Freedom of Information Act. The authors do provide a link in their paper to **R-code** that could possibly run successfully on a Linux-Cluster of computers with direct access to U.S. Government databases. My request was for a copy of their data in the form of a single **Comma**-Separated-Values "flat file" with 2,708 rows and fewer than 100 columns. Each row corresponds to data aggregated to the level of an individual U.S. County or Parish. Each column corresponds to a response or predictor variable analyzed in their paper (PM_{2.5}, SOA, Cardiorespiratory Death Rate, etc., etc.) This is **not** a "large" data-file by current-day standards. Apparently, only about 248 of the 2,708 given SOA values were actually "measured"; thus, about 2,460 SOA values were "imputed" via a model using other variables. My initial data request to the lead author was dated Jan. 19, 2022, and I have not yet received any response to my request from anyone at the EPA.