JERRETT ET AL. CRITICISM

Criticism of "Spatiotemporal Analysis of Air Pollution and Mortality in California Based on the American Cancer Society Cohort: Final Report (as revised)" by Michael Jerrett, Richard T. Burnett, Arden Pope III, Daniel Krewski, George Thurston, George Christakos, Edward Hughes, Zev Ross, Yuanli Shi, Michael Thun, et al.

William M. Briggs
Statistician
300 E. 71st A3R, New York, NY 10021
email: matt@wmbriggs.com

October 24, 2011

1. Overall

Author Michael Jerrett and nine co-investigators and four student or post-doctoral investigators prepared the report "Spatiotemporal Analysis of Air Pollution and Mortality in California Based on the American Cancer Society Cohort" prepared under Contract # 06-332 of State of California Air Resources Board Research Division.

The purpose of Jerrett was to investigate the relationship between particulate air pollution, stated as $PM_{2.5}$, and mortality in the State of California.

On p. 7 it is stated, "All-cause mortality is significantly associated with $PM_{2.5}$ exposure, but the results are sensitive to statistical model specification and to the exposure model used to generate the estimates" They derive an estimate of 1.08 hazard ratio, with a classical confidence interval between 1.00 and 1.15. They also claim that the risk associated with death due to cardiovascular disease (CVD) and $PM_{2.5}$ is significant. The risk of $PM_{2.5}$ with other causes of death they claim are insignificant.

There are three main criticisms that cast grave doubt about the conclusions of Jerret. I find further that the summary in the abstract—and therefore the only part of the report liable to be read by most—to be the result of either poor work or deliberate bias toward a predefined conclusion.

(1) The authors prepared, intensely investigated, and justified the use of a series of complex statistical models. There were nine models in total, each having particular strengths and weaknesses. Each had several subjective

"knobs" and dials to twist. Only *one* model of the nine (LUR IND+Met; Fig. 22, p. 108) showed a "statistically significant" relationship between mortality and $PM_{2.5}$, and that only barely; and in that model, only one sub-model showed "significance." The other eight models showed no relationship. Some models even hinted that $PM_{2.5}$ *reduced* the probability of early mortality. With such a large number of tests and "tweaks", the authors were practically guaranteed to find at least one "significant" result, even in the absence of any effect. Nowhere did the authors control for the multiplicity of testing, even though such controls are routine in statistical analyses of these sort.

Death by Cardiovascular disease was also said to be significant. This result is particularly suspect because of the enormous differences in death rates betwen rural and urban populations. The authors did model this distinction, but in a manner that was too simple to be conclusive.

(2) The authors only chose to report, in the Abstract (p. 8), on the one model that was "significant", ignoring all others. They also departed from the main text and inflated the size of the hazard estimate: on pp. 87-88 the estimates in Table 31 are based on a change in the interquartile range of $PM_{2.5}$, but in the summary this is inflated to present a larger effect, presumably for emphasis. This behavior makes no sense statistically and is either sloppy writing or the result of purposeful choosing a result because of bias.

(3) The models were a mixture of Bayesian and frequentist methods, but incomplete mixtures. Substantial uncertainties remain in the model constructions such that the results are too certain, i.e. the confidence and credible intervals are too narrow. It is likely that were these uncertainties properly handled, even the one model which did show "significance" would not retain that significance (see below).

(4) Even assuming the models are trouble free, and the model that indicated significance was the only model worth showing, we have to consider that the authors claimed to have shown a relationship between $PM_{2.5}$ and inhalation. Yet the authors never, not even in one case, measured the $PM_{2.5}$ inhalation of any person. How, then, could the authors claim that $PM_{2.5}$ inhalation is associated with early mortality? They cannot. At best, they can claim *(part time) residence* is associated with early mortality.

    Instead of $PM_{2.5}$ inhalation, the authors instead measured (with unaccounted for error; see Section 2) the residence of a sample of Californians. Residence was taken as a perfect, error-free, and unique proxy of $PM_{2.5}$ inhalation. This is absurd, even on the authors own reasoning. About this, more in Section 2.

At the least, these criticisms call for additional study before any decisions are made regarding $PM_{2.5}$ inhalation and mortality.

2. Detailed Criticisms

2.1. **Urban versus rural population.** Wide variances of mortality occur between urban and rural areas in California. Further, habits of life differ widely between the two. The authors write on p. 42:

> Specifically across the United States, in the 1980s there were on average 6.2 excess deaths per 100,000 in non-metropolitan areas compared to metropolitan areas, and this number increased to 71.7 excess deaths for the period 2000-2004.

This enormous and growing difference has profound consequences for any wide-region model of all-cause death. The authors' answer was to include indicators (which would change the intercept of the model only) for whether a person lived in any of fives Metropolitan areas (pp. 42-43). Even given the discrepancies in raw mortality statistics, the indicator for Los Angeles was not even significant (p. 87, Table 31). It is not clear where there others indicators (for the other metropolitan areas) went; they are not reported.

On p. 73 some of their estimates "became insignificantly elevated or were of borderline significance when the Los Angeles indicator and interaction terms were included." Table 27 later lists this as insignificant for many causes. This is odd and should be explained.

At the least, these indicators should have been included (at least for research) as a multiplier (or interaction) for the other variables in the models besides just $PM_{2.5}$. This would have changed the size of the effect of these variables inside and outside

of metropolitan areas. Better still, a hierarchical model using urban/non-urban residences would have gone a long way to quantifing these differences.

Another difficulty is the rapid change in the differences of urban-rural death rates through time. No attempt was made by the authors to incorporate this in the models. This lack of control could certainly be in favor of "significance" of $PM_{2.5}$ and all-cause mortality in the land use model.

Higher CVD deaths, incidentally, are found in *rural* populations (where ambulances and hospitals are more distant). This is most important. Since it was CVD deaths that were found significant by the authors, and since CVD made up a large proportion of overall deaths, it is likelier still that misspecification of urban versus rural populations contributed to the bare significance of one of the authors' models.

In short, what we might be seeing in these models is nothing more than a location effect unrelated to $PM_{2.5}$.

2.2. **Per-person $PM_{2.5}$ exposure.** It must be clearly understood that no person's $PM_{2.5}$ exposure was ever measured. The statements that $PM_{2.5}$ was associated with all-cause death is therefore a misnomer.

Instead of actually measuring $PM_{2.5}$, the authors created a guess of exposure based on where each person in the database (at one time) lived (see the next section). The assumption is that merely living in an area is an *error-free* proxy for actual $PM_{2.5}$ exposure. This, of course, is false.

And because it is false, it is true that the results from each model are too certain. At the least, the confidence intervals limits are too narrow. Since this is so, and

since only one model barely reached classical statistical significance, it is more than likely that *actual* $PM_{2.5}$ exposure is not significantly related to all-cause death.

Now, in creating their guess, the authors could have, but did not, create a per-person estimate of $PM_{2.5}$ exposure. They instead averaged exposure data across months or event years ("constructing 12-month moving averages from January 1988 to December 2000" p. 43). Why "moving averages"? Why not use just the numbers themselves as estimates of $PM_{2.5}$ exposure? No convincing justification is given. Excessive smoothing, caused by moving averages, tend to inflate correlations (and to improperly narrow confidence bounds).

The authors could have, but did not, create simple plots of all-cause death by exposure level, just as a sanity check. It is strange that these are missing given the plethora of other graphics.

2.3. **Uncertainty of $PM_{2.5}$ exposure.** This is a key criticism. Given that they could not directly measure $PM_{2.5}$, the authors had to make a guess. The guess was input as *certain* and true into the models. That it, the authors did not take into account the uncertainty of the exposure.

The authors used Bayesian exposure models, but only picked the means, medians, or modes of the posterior distribution of $PM_{2.5}$—and we are never sure which of these point estimates was finally used; there is more than a hint of data snooping here.

What they should have done is to pick a level of exposure implied by the posterior of $PM_{2.5}$ and then computed the rest of the model and set that result aside. They

should have then picked another level implied by the posterior, repeated the model fit and saved, etc. Then they could have weighted all these results together (the weights determined by the posteriors) and this weight would be the final answer.

No matter what, this answer derived from this proper analysis *will be less certain* than what they have shown. It is therefore extremely unlikely that any of the models would have showed statistical significance.

Curiously, the authors point out that their kriging estimates of $PM_{2.5}$ look smooth and conclude that actual values of $PM_{2.5}$ *are* smooth. But kriging, by design, produces smooth estimates. Statements like these cause concern that the authors do not fully understand the tools they are using.

2.4. **Uncertainty of land use model.** The exact same criticism can be made for the land use model. Only point estimates were used, and no account of the uncertainty of land use was made. Once again, and taking into account the previous over-certainties, it is even more likely that none of the models would have showed statistical significance.

2.5. **Uncertainty of where a person lived.** They did not control adequately for where a person lived. This is crucial because it is solely from where a person lived that the authors guessed at $PM_{2.5}$ exposure. It appears the authors used the last address only: on p. 43 they say, "We assumed that each subject resided at their home address in 1982 throughout the follow-up period to December 2000."

This will be true for some, but surely not all, persons in the database. Therefore, there must be large errors in estimating where a person lived. And that means large

errors in $PM_{2.5}$ exposure estimates, and therefore even larger errors in actual $PM_{2.5}$ exposures.

Of course, and once more, this translates into model statements that are too certain.

2.6. **Uncertainty in dietary and demographic variables.** The authors used diet and "beer, wine, and alcohol" self-report variables in their models. They also used Census-derived variables such as percent white residents (in a geographic area). All these variables are notoriously poor. These variables also changed over the period in question, but these changes were not incorporated into the models.

Using these variables as certain in the model, as before, creates over-confidence.

2.7. **Uncertainty in model diagnostics.** Fig. 5 (p. 47) is supposed to be a check on model goodness (for just one model). Why so few points in this plot? Surely the authors have many more observations of $PM_{2.5}$ than are indicated.

Further, the model does more poorly the larger $PM_{2.5}$ is. Figs. 14 (p. 58) and 15 (p. 60) are other model checks. These too indicate very poor performance at higher values of $PM_{2.5}$.

Since it is the authors' conclusion that *increasing* $PM_{2.5}$ is associated with premature death, poorer model performance at increased levels of $PM_{2.5}$ calls that conclusion seriously into question.

2.8. **Other pollutants.** The NO_x, PM_{10}, etc. models are presented as additional evidence, but they are not. These pollutants are highly correlated to $PM_{2.5}$, and each is estimated in the same way, so reporting on them in the fashion the authors

chose is essentially repeating the same information twice in the guise of independence.