UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

SUBJECT: MODELING THE PERCENT OF CHILDREN WITH HIGH BLOOD LEAD LEVELS

FROM: Joel Schwartz
Office of Policy Analysis

AUG 21 1985

TO: Vic Hasselblad
Allan Marcus
Jeff Cohen

The distribution of blood lead levels in U.S. children is a broad one, because there is a distribution of exposures those children face, a distribution of behavior patterns that affect the uptake of that exposure, and a distribution of biological absorbtion and excretion rates. Observational data bases, such as NHANES II, reflect all of those contributions to the variance of the distribution. I believe that the issue we have been grapling with on how to determine what air lead standard will protect a given percent of children from exceeding some cutoff blood lead level is the issue of how to deal with those variance components, and how they might change.

I believe that Vic and Allan's point is that the variance of the distribution is partially due to the variance in air lead levels; therefore the percent of children at high blood lead levels if all areas were at a given ambient standard should be estimated in a way that removes that variance. I believe that that point is a fair one. My point has been that the technique chosen also removes some of the variance in non-air exposure, thereby underestimating the variance at uniform air lead, and that the assumption that the tail of the distribution is well modeled by fitting a lognormal distribution to the overall distribution is less robust than some other techniques for estimating the size of the tail. I will discuss these further in the hopes that we can reach an agreement on the best way to do the modeling.

My first point is that much of the variance associated with some of the demographic variables reflects variation in baseline exposure, behavior patterns or biological factors, and not air lead, and therefore adjusting the distribution for them as a technique for removing the variance due to variation in air lead levels is excessively conservative. For example, consider race. Blacks have higher blood lead levels than whites, however the NHANES data indicates that blacks have higher blood lead levels than whites even after controlling for degree of urbanization, income, residence inside the central city, parents educational level, and any other variable that appears to test whether that difference is soley due to differential exposure to high urban air lead. Indeed, some of those variables also test whether the racial difference is due to differential exposure to paint lead.

After controlling for all of those factors there is still a substantial difference between the races in their blood lead levels. Indeed, when we did our gasoline lead regressions, which control for the major source of air lead exposure, as well as age sex, income , degree of urbanization, region of the country, educational level, dietary consumption, etc, blacks still had significantly higher blood lead levels. This difference reflects true differences in the other factors affecting blood lead levels, and therefore should remain as a real factor that contributes to the GSD.

Location is another factor that represents more than just difference in air lead exposure. While the estimates of mean baseline exposure in the criteria document are reasonable, and the best data we have, clearly food lead levels, dust lead, water lead, and other factors vary from location to location. One important problem with the modeling in the staff paper is that it ignores the distribution of city mean baseline levels. Since the percent of children at high blood lead levels is nonlinear in the mean baseline assumed, adding a distribution in mean blood lead levels about the mean baseline quoted in the CD and staff paper will in fact change the estimate of the number of children at high levels. The probability of exceeding a given blood lead level is clearly the product of the probability of the location having a mean u, given the grand mean, times the probability of having a high blood lead, given u and the gsd within a site. Since the product of log normal distributions is lognormal, this is equivalent to spreading the gsd by adding the variance of the site means to the variance of the within site population. The current modeling ignores this by assuming that the baseline in every location is the same.

If I recall what Vic told me correctly, the gsd's quoted in the CD removed the variation due to site effects; whereas they should only remove the site variations due to air lead variations, and leave the site variations due to variations in the baseline exposure.

The NHANES II data can be used to quantify this effect. Assuming that degree of urbanization is a proxy for the distribution of air lead levels (and I argued before that some of the variance due to degree of urbanization is due to other factors), I regressed the log of blood lead levels on the eight level degree of urbanization variable in the NHANES II data . I then added the mean back to the residuals and took the mean log blood lead level, by site. This is the mean after adjusting all the sites to the same population. Table 1 shows the mean, standard deviation, and "geometric standard deviation", which is exp of the stadard deviation in log space. We can see that there is an expected variation in mean blood lead levels for cities of the same size. Adding the variance of this distribution to that of a distribution with a gsd of 1.39 would give a distribution with a gsd of 1.462, which is shown in the last column.

The next line carries the analysis further, by first

controlling for degree of urbanization, income, and race, before taking the site means. Again there is a noticable residual variance in the means of locations with the same population, racial mix and income levels. Adding that variance to that of a population with a gsd of 1.39 would give a new gsd of 1.458.

The above calculations demonstrate that there is a variation in mean blood lead levels among cities, and that correcting for demographic proxies for air lead exposure leaves a residual variation in site means due to variations in non-air exposure. This is an approximation to the distribution of mean blood lead levels across sites with the same air lead exposure. Estimating a childs likelihood of having an elevated blood lead level, or the percent of children above a certain blood lead level at a given air lead level cannot ignore this variation in background level.

TABLE 1

| CASE | LOG MEAN | LOG SD | GSD | TOTAL GSD |
|---|---|---|---|---|
| Site Means, Controlling For Degree of Urbanization | 2.665 | 0.189 | 1.208 | 1.462 |
| Site Means, Controlling for Urbanization, Race and Income | 2.666 | 0.183 | 1.200 | 1.458 |

There are likely to be differences in baseline exposure that are a function of degree of urbanization, since higher population density will increase the soil contamination from other sources than air (e.g. old batteries, industrial facilities etc.). Similarly some of the racial and income differences in blood lead levels are certainly not due to airlead exposure. On the other hand, these proxies may not completely control for air variations. It would clearly be better to remove the airlead variance directly.

To do this I repeated the gasoline lead regressions, and used them to remove the variance due to gasoline lead exposure, which is about 85% of air lead exposure. By ajusting every individuals blood lead level to what it would be at zero gasoline lead exposure, we arrive at a good estimate of the distribution of blood lead levels due to baseline exposure variations, biological differences, etc. To account for the remaining 15% of air lead exposure, the resulting gsd can be reduced by 15% of the change in the gsd in going from the unadjusted distribution to the one expected at zero gasoline lead. Using the gsd from this analysis incorporates differences due to variations in baseline exposure among locations, races, etc, which must be incorporated to correctly predict the effect of of additional exposure to air lead on the percent of children that will have high blood lead levels. To maintain the distribution as log normal, the gasoline regression was done using a dependent variable of log blood lead.

The regression is shown in the appendix. Then each individuals log blood lead was reduced by the gaslead regression coefficient times their gasoline lead exposure, and a weighted univariate was performed on the resulting distribution. The results are shown on the first line of table 2. The estimated gsd is 1.428. Moreover, the principal reason that this gsd is higher than those in the CD is the natural variation in the baseline exposure among sites with similar demographic characteristics, not the variation in demographic characteristics between sites. To illustrate this, the second line in table 2 shows the same results, after adjusting by regression for the effect of degree of urbanization. Even if all children lived in towns with the same degree of urbanization, they would have a gsd of 1.414. To illustrate this further, the next two lines show the mean and gsd of location means at zero gasoline lead with and without adjusting for degree of urbanization. (The mean reflects zero gaslead, but not zero stationary source lead, the gsd reflects our correction for both factors.)

TABLE 2

| CASE | LOG MEAN | LOG SD | GSD |
|---|---|---|---|
| Controlling for Airlead (zero airlead) | 1.99 | 0.360 | 1.428 |
| Controlling for Airlead and Degree of Urbanization | 1.989 | 0.346 | 1.414 |
| Location Mean Controlling for Airlead | 1.935 | 0.187 | 1.46 * |
| Location Mean Controlling for Airlead and Urbanization | 1.978 | 0.161 | 1.44* |

Note: site means differ from overall means because the sites had different populations, which is controlled for in the overall mean.

* assuming a GSD of 1.39 after controlling for sites, and adding the site variance due to non air sources back in.

In summary, it appears that the best estimate for the GSD for children six and under, if the air lead was zero, is 1.428. To calculate the percent of children with blood lead levels above the lowest effect level, we should use this GSD and look at the effect of increasing everyones airlead exposure to 0.50, etc. This will give the percent of children expected to exceed that level if every location is at the ambient standard.

I still like the idea of doing logistic regressions to see how

the percent of children above a given level changes. Regression is a more robust technique than fitting distributions, and it can also be corrected to eliminate the variance due to airlead exposure. I will try to do that soon and send you the results.

REGRESSION RESULTS
LOG BLOOD LEAD, CHILDREN SIX AND YOUNGER

| VARIABLE | COEFFICIENT | t-STATISTIC | P-VALUE |
|---|---|---|---|
| INTERCEPT | 2.256 | 36.01 | 0.0001 |
| GASLEAD | 0.1624 | 17.67 | 0.0001 |
| URBANIZATION | -0.0265 | -10.44 | 0.0001 |
| RACE | 0.2129 | 11.24 | 0.0001 |
| SEX | -0.0307 | 2.35 | 0.0187 |
| INCOME | -0.1219 | -12.69 | 0.0001 |
| AGE | -0.0185 | -4.974 | 0.0001 |

NOTE: Regression was run in WLS, not SURREGR, since this gives the identical coefficients, and we already know the significant variables.