Carcinogenesis vol.34 no.7 pp.1689–1691, 2013
doi:10.1093/carcin/bgt100
Advance Access publication March 25, 2013

LETTER TO THE EDITOR

Ignoring and adding errors do not improve the science

Stephen M.Rappaport*, Brent A.Johnson[1], Frederic Y.Bois[2,3], Lawrence L.Kupper[4], Sungkyoon Kim[5] and Reuben Thomas

Superfund Research Program and Center for Exposure Biology, School of Public Health, University of California, Berkeley, CA 94720, USA, [1]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA, [2]Department of Toxicology, Université de Technologie de Compiègne, 60200 Compiègne, France, [3]Institut National de l'Environnement Industriel et des Risques, DRC/VIVA/METO Unit, 60550 Verneuil en Halatte, France, [4]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599-7420, USA and [5]Department of Environmental Health, School of Public Health, Seoul National University, Seoul 151–742, Republic of Korea

*To whom correspondence should be addressed. Tel: +510-642-4355; Fax: +510-642-5815;
Email: srappaport@berkeley.edu

Dear sir,

We respond to the recent 'Letter-to-the-Editor' from Price *et al.* (1) regarding our paper published in *Carcinogenesis* (2). Let us briefly review the issues. Kim *et al.* (3,4) measured benzene and its metabolites in air and urine samples from 389 Chinese workers in factories where benzene was and was not used. Natural-spline modeling showed an estimated 9-fold reduction in the dose-specific metabolism (DSM) of benzene (micromolar metabolites per p.p.m. benzene) at air concentrations between 0.03 and 90 p.p.m. with most of the reduction occurring <1 p.p.m (4). Uncertainty analyses indicated that the trend toward reduced benzene metabolism with increasing air concentrations was unlikely to be the result of chance (3,4). Upon reanalysis of the Chinese data with the same spline models, Price *et al.* (1) (Approach A) also found a 9-fold reduction in DSM. However, they discounted this result largely because Kim *et al.* had assigned the 139 subjects from the factory that did not use benzene into different groups depending on the concentrations of benzene in their urine. That is, Kim *et al.* assigned 60 subjects with the lowest benzene concentrations to a sample for estimating background metabolite levels (background sample), whereas the next 79 subjects were included in the 'modeled sample' to represent low benzene exposures from smoking, petroleum products, engine exhausts and so on.

Price *et al.* repeated the spline and uncertainty analyses using different background and modeled samples (summarized in Figure 2 of (2)); that is, Approach B assigned all workers from the factory without benzene to the background sample and Approach C assigned all subjects with benzene exposures below 0.03 p.p.m. to the background sample. These alternative approaches increased background metabolite levels and also reduced modeled-sample sizes by about 25%. Based upon analyses under Approaches B and C, Price *et al.* concluded that the data '… appear to be too uncertain to support any conclusions of a change in the efficiency of benzene metabolism with variations in exposure' (1). In our rebuttal, we showed that Price *et al.*'s reassignment of subjects with demonstrable benzene exposure to background samples—with concomitant reduction of modeled samples—obscured the ability to discern low-dose metabolic effects (2). We also reported several errors that rendered Price *et al.*'s results unreliable, namely, a mathematical error in Equation (7) (2) that was used to adjust for bias in spline models, unreported and apparently incorrect selection of knots for spline models under Approaches B and C and results from bootstrap distributions for uncertainty analyses under Approaches A–C that did not agree with the corresponding data distributions (Figure 4 in (2)). Although these serious errors compromised their major conclusion regarding the DSM of benzene (noted above), Price *et al.* do not address them in their letter. Rather, Price *et al.* revisited much of their earlier discourse in light of our paper. We will address their points in the order of importance.

Background and modeled samples

As noted above, Price *et al*. insist that all 139 workers from the factory without benzene should have been included in the background sample instead of the 60 subjects with the lowest benzene exposures. They support this contention with Figure 5 in their letter, based upon calculations employing Equation (3) in their supporting materials (which is itself derived from Equation (S4) of our paper) (2) as follows:

> The reason for this is that the selection of the 60 individuals with the lowest urinary benzene levels as the basis for estimating the population background levels of each metabolite in the workers' urine … is a plausible but arbitrary decision. As demonstrated in supplemental materials to this letter, the maximum contribution of the metabolites that occur from the air exposures for the control workers can be determined using Kim's estimates of air levels (based on the workers' urinary benzene levels and the Kim *et al*. (2). calibration model) and an assumption of a metabolic fraction of 1. When this was done the maximum contribution from air exposure to the 139 control workers' total benzene metabolites averaged only 0.28% of the observed levels of benzene metabolites (99.72% was due to background sources). Thus there is no objective reason for not using the mean of all 139 control workers as an estimate the background levels of the workers in factories that used benzene.

Well, we repeated the calculations and found that the 'maximum contribution from air exposures to the 139 control workers' total benzene metabolites' did not average 0.28% as stated by Price *et al.*, but rather averaged 6.8% (i.e. 6.7 µM from air/98.4 µM from background sources)—a 24-fold higher contribution! When we divided these 139 workers into two groups based upon subjects' benzene exposures, the corresponding 'maximum contribution from air exposures' was 0.93% (i.e. 0.80 µM from air/86.3 µM from background sources) for the 60 lowest exposed subjects and 10.5% (i.e. 11.2 µM from air/107 µM from background sources) for the remaining 79 subjects, that is, more than a 10-fold difference! Thus, calculations motivated by Price *et al.*'s letter ironically reinforce Kim *et al.*'s background adjustment *via* the 60 lowest exposed subjects, where benzene exposure contributed <1% of measured metabolite levels. As to why benzene exposures varied so much across the control workers, a major factor appears to be cigarette smoking because 11 smokers were included in the lowest 60 subjects compared with 28 smokers in the remaining 79 subjects.

Another point of contention is whether to take means or medians of benzene metabolite levels for background adjustment. It seems silly for Price *et al.* to continue to argue that Kim *et al.*'s use of median values for background adjustment was an 'error', whereas their use of mean values was correct. If an argument or proof was to be made that use of mean values is necessary for background adjustment, then it should be based on statistical principles like robustness (sensitivity to the presence of outliers), maximum likelihood (given an assumed probability distribution), loss functions (minimizing error) and so on. The theoretical proof provided in Price *et al.*'s Appendix C (1) to justify use of mean-background adjustment is nothing more than taking sums on the right- and left-hand sides of a linear equation. In fact, model estimates of DSM derived by Price *et al.* using mean-background levels from the 60 lowest exposed subjects (Approach A, DSM reduction = 9.4-fold) were quite similar to those reported by Kim *et al.*, who used median-background levels (DSM reduction = 9.2-fold) (4). This suggests that selection of mean or median metabolite levels for background adjustment has a relatively small effect on estimation of DSM, provided that a comparable measure of central tendency is used for the air concentrations.

The calibration model

Price *et al.* criticize Kim *et al.*'s use of a calibration model to estimate benzene air concentrations among low-exposed subjects (n = 22 in factories with benzene and n = 139 in factories without benzene) (3). As calibration was based on air and urinary benzene measurements in 228 workers from the benzene-using factory only, Price *et al.* contend that it was inappropriate for Kim *et al.* to predict exposures in subjects from the factory that did not use benzene. To support this contention, they summarized log-scale linear models of air and urinary benzene measurements showing somewhat different relationships across seven studies (Figure 4 in their letter). Such interstudy variability arises from methodological differences leading to losses of benzene during collection, processing and analysis of urine, and from differential effects of benzene exposures in smokers relative to the dynamic range of exposures. Because the Chinese study scrupulously applied the same collection, storage and analytical protocols for all subjects, there were no methodological differences that would have contributed to differential loss of urinary benzene between the two factories. And although underestimation of benzene exposures in smokers is unavoidable with lapel-mounted personal air monitors, Kim *et al.*'s calibration model included only subjects with air concentrations at or above 0.2 p.p.m., where any such bias would have been small. Indeed, 'true' benzene exposures predicted by Kim *et al.* for low-exposed smokers were more accurate than they would have been had they been estimated from lapel-mounted monitors, which greatly underestimate the contribution of mainstream smoke.

Price *et al.* contend that Kim *et al.*'s calibration model introduced errors into estimates of air concentrations. To clarify this issue, we will review the statistical models that were used. Kim *et al.* based their calibration of low-exposed workers on the following regression model:

$$\log(\bar{U}) = \beta_0 + \beta_1 \log(\bar{A}) + \varepsilon, \qquad (1)$$

where $\bar{U}$ and $\bar{A}$ are the geometric means of replicate urinary and air benzene measurements, respectively, and the error ε is a mean zero, normal random variable with unknown variance. The regression Model (1) led to the calibration model:

$$\bar{A} = \exp[\log(\bar{U}) - b_0]/b_1, \qquad (2)$$

where b_0 and b_1 are the estimated intercept and slope, respectively. Note that Model (1) could also be written as a model of the mean $E[\log(\bar{U})] = \beta_0 + \beta_1 \log(\bar{A})$ or conditional mean $E[\log(\bar{U}) \mid \bar{A}] = \beta_0 + \beta_1 \log(\bar{A})$ if $\bar{A}$ is regarded as fixed or random, respectively. In this case, calibration *via* Model (2) follows directly from simple algebra and is semi-parametric in the sense that it models the first moment only and does not refer explicitly to an error distribution.

If the error in Model (1) is normally distributed, then $\bar{U}$ is lognormal and one can derive a maximum-likelihood estimator based on the normal distribution. This is essentially Price *et al.*'s 'bias adjustment factor' (Equation (6) in (1)) that was applied using mean rather than geometric mean values for $\bar{U}$ and $\bar{A}$. But if the distribution of errors in Model (1) is not normal, then use of such an adjustment can lead to other biases of unpredictable magnitude and direction (for a general reference on the maximum likelihood versus moment estimators in nonlinear regression, see ref. 5). Given the skewness of the Chinese data, we were reluctant to assume a lognormal distribution and relied instead on the simple moment estimators for calibration under Model (2). We were bolstered in this approach by predicted air concentrations from the calibration model that were consistent with independent reports of benzene exposure in urban populations and smokers (3). Nonetheless, if the error distribution in Model (1) was truly normal, then the air benzene concentrations of the low-exposed subjects predicted from the calibration model by Kim *et al.* (3). would have been overestimated. One implication of overestimating air exposures from the calibration model would have been to 'underestimate' DSM at low air concentrations, suggesting that the estimated reduction of benzene metabolism would have been 'greater' than 9-fold between 0.03 and 90 p.p.m.

Price *et al.* also indicate that our updated uncertainty analyses (2) did not include 'the unexplained variance in the predictions of the calibration model.' We have no idea what they mean by this. In our uncertainty analyses, we sampled ($\bar{A}$, $\bar{U}$) pairs with replacement until the requisite 228 such pairs were obtained. Then, we fit Model (1) to the data and used the estimated parameters *via* Model (2) to do the calibration for subjects missing $\bar{A}$ values. Finally, the full set of 386 ($\bar{A}$, $\bar{U}$) pairs was used to construct a natural-spline model. This complete analysis (from calibration through spline-model fitting) was repeated several thousand times (bootstrapping) and confidence intervals were obtained from the resulting distributions of estimators. Thus, our uncertainty analyses appear to account for all sources of variation in our methodology—including variation in prediction from the calibration model—but not for possible bias (see above). Apparently, Price *et al.* regard 'unexplained variance in the predictions of the calibration model' in a different way but they offer no references for justification, despite our earlier criticism (2).

Direct versus indirect measurement of internal dose

In supplemental materials to our paper (2), we showed that the steady-state ratio of benzene concentrations in exhaled and inhaled air, $C_{\text{exh}}/C_{\text{inh}}$, can be related to the corresponding metabolized fraction of inhaled benzene $Q_{\text{met}}/Q_{\text{inh}}$ (Q signifying quantities of benzene), and—with assumptions about rates of alveolar ventilation (V_{alv}) and urinary excretion—to DSM (Equation (S4)) (2). Then, we used values of $C_{\text{exh}}/C_{\text{inh}}$ from four human studies to estimate $0.4 \leq Q_{\text{met}}/Q_{\text{inh}} \leq 0.7$ over the benzene exposure range 0.02 p.p.m. $\leq C_{\text{inh}} \leq$ 57 p.p.m. Interestingly, the corresponding range of 506 μM/p.p.m. ≥ DSM ≥ 86 μM/p.p.m. showed that DSM decreased about 6-fold between 0.02 and 57 p.p.m. (Figure 6) (2). Thus, even though $C_{\text{exh}}/C_{\text{inh}}$ is an imprecise predictor of $Q_{\text{met}}/Q_{\text{inh}}$ (2), these data offer independent evidence that DSM is enhanced at benzene concentrations <1 p.p.m.

In their comments and Figures 1–3 of their letter, Price *et al.* attempt to turn this result on its head by suggesting, on the one hand, that $Q_{\text{met}}/Q_{\text{inh}}$ is a more appropriate measure of benzene metabolism than DSM, and, on the other hand, that metabolic data derived from estimates of $C_{\text{exh}}/C_{\text{inh}}$ are comparable to those from direct measurements of benzene metabolites. Because DSM reflects the quantity of benzene metabolites produced per unit time and the magnitude of V_{alv}, both of which can vary greatly across subjects, DSM is a more objective measure of internal dose (and risk) than $Q_{\text{met}}/Q_{\text{inh}}$. This was illustrated in Table S5 and Figure 6 of our paper (2), where a 3-fold increase in estimated DSM was observed at a given $Q_{\text{met}}/Q_{\text{inh}}$ due to the higher breathing rate in automobile mechanics (Egeghy *et al.* (6) study, V_{alv} = 17 l/min) compared with sedentary volunteer subjects (other three studies, V_{alv} = 6 l/min). Regarding the quality of data used to estimate DSM or $Q_{\text{met}}/Q_{\text{inh}}$, there can be little doubt that 'direct' measurement of benzene metabolites, as performed by Kim *et al.*, provides vastly more accurate and precise metabolic information. Indeed, 'indirect' estimation of human metabolism *via* $C_{\text{exh}}/C_{\text{inh}}$ is very imprecise because a range of $C_{\text{exh}}/C_{\text{inh}}$ is compatible with a given value of $Q_{\text{met}}/Q_{\text{inh}}$, and *vice versa* (2). Price *et al.* are essentially arguing that we should turn the clock back 30 years and abandon direct measurements of internal dose (*via* urinary benzene metabolites) in favor of indirect and imprecise dose surrogates such as $C_{\text{exh}}/C_{\text{inh}}$, coupled with population-averaged kinetic parameters.

In closing, we reiterate that Price *et al.* have not clarified their use of spline and uncertainty models so as to counter obvious errors and permit independent confirmation of their results (2). Here, we have identified additional errors in Price *et al.*'s Letter-to-the-Editor that further diminish the validity of their arguments. As for the tone of our discourse, it is fair to say that Price *et al.* questioned the integrity of our workmanship and we responded in turn. Perhaps, the most important message from this exchange is that investigators should carefully examine competing results derived from complex data under different analytical approaches and assumptions. We will rely upon the discerning reader to judge the strengths and weaknesses of arguments presented in these letters and the publications that preceded them.

Funding

National Institute of Environmental Health Sciences grant (P42ES04705 to S.M.R.).

References

1. Price,P.S. *et al.* (2012) A reanalysis of the evidence for increased efficiency in benzene metabolism at airborne exposure levels below 3 p.p.m. *Carcinogenesis*, **33**, 2094–2099.
2. Rappaport,S.M. *et al.* (2013) Low-dose metabolism of benzene in humans: science and obfuscation. *Carcinogenesis*, **34**, 2–9.
3. Kim,S. *et al.* (2006) Using urinary biomarkers to elucidate dose-related patterns of human benzene metabolism. *Carcinogenesis*, **27**, 772–881.
4. Kim,S. *et al.* (2006) Modeling human metabolism of benzene following occupational and environmental exposures. *Cancer Epidemiol. Biomarkers Prev.*, **15**, 2246–2252.
5. Carroll,R.J. *et al.* (1988) *Transformation and Weighting in Regression*. Chapman and Hall, New York, NY.
6. Egeghy,P.P. *et al.* (2002) Self-collected breath sampling for monitoring low-level benzene exposures among automobile mechanics. *Ann. Occup. Hyg.*, **46**, 489–500.

Received March 6, 2013; accepted March 16, 2013