Bayesian Analysis (2007) **2**, Number 1, pp. 37–42

Comment on Article by Dominici et al.

David Ruppert[*] and Raymond J. Carroll[†]

We thank the authors for a very thought-provoking paper. They address a problem of great importance and have proposed an interesting and ingenious solution. It is quite challenging to develop a model that is suitable for the complex data structure presented by this problem, and the authors should be congratulated for their success.

We believe in the philosophy of using whatever works best, and for complex problems such as this one, a fully Bayesian analysis seems much more satisfactory than anything else we know of. As the authors show, a Bayesian analysis allows one to multiply impute missing data, the true values of mismeasured covariates, and counterfactuals and to adjust inference for uncertainty in the imputations. Figure 4 shows the importance of the latter. As we will discuss shortly, there are a few places where the authors could have improved their analysis by taking further advantage of Bayesian techniques. For example, Bayesian modeling would allow for a more data-driven choice of the spline model. We also believe that their measurement error model (9) could be improved by a fully Bayesian analysis rather than using a regression that is fit outside the Gibbs sampler.

The authors use counterfactuals to relate treatment effects on birth weight to treatment effects on survival. This is a very interesting technique and relatively new to us. However, as discussed at the end of these comments, we wonder to what extent the results, which are based on counterfactuals, really prove a causal relationship between these treatment effects.

The paper starts with a simple analysis based upon model (1) for $W_{it_i}^{obs}$ that is conditional upon covariates including the outcome Y_i^{obs}. Because models such as (1) do not distinguish between cause and effect, e.g., do not say whether $W_{it_i}^{obs}$ influences Y_i^{obs} or vice versa, we find them less satisfactory than hierarchical Bayesian models such as equation (4). The authors appear to be in agreement with us here.

A common approach taken in the measurement error literature is to develop a hierarchical model with three basic components. Using, for the moment, the notation of Carroll et al. (2006), the three parts of the model are

- an "exposure model" for the true values, $\boldsymbol{X}$, of the error-prone covariates conditional upon correctly measured covariates, $\boldsymbol{Z}$,
- a "measurement error model" for the surrogates, $\boldsymbol{W}$, given $(\boldsymbol{Z}, \boldsymbol{X})$, and
- an "outcome model" for the responses, $\boldsymbol{Y}$, given $(\boldsymbol{Z}, \boldsymbol{X})$, which is assumed to be

[*] School of Operations Research & Industrial Engineering,Cornell University, Ithaca, NY, mailto:dr24@cornell.edu

[†] Department of Statistics, Texas A&M University, College Station TX, mailto:carroll@stat.tamu.edu

the same as the distribution of $\boldsymbol{Y}$ given $(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{W})$—this assumption is called nondifferential measurement error.

Given these three components and a prior, it is straightforward to implement an MCMC to sample from the posterior—see Richardson (1998), Gustafson (2004), or Carroll et al. (2006). The vector $\boldsymbol{X}$ is an unknown "latent variable" or "missing data" and treated in the same way as the parameters, that is, one samples from its full conditional given the observed data and the parameters. Other missing data can be treated in this way as well.

Since the authors' data have a more complex structure than a typical measurement error model, the hierarchical model must be somewhat more complicated. Returning to the authors' notation, the approach when applied to the authors' study has the following three components:

a) a model for $\{W(0), W(1)\}$ conditional on $\boldsymbol{x}_i$, where $\boldsymbol{x}_i$ is a vector of covariates, e.g., the authors' model (7),

b) a model for $\{W_{it_i}(0), W_{it_i}(1)\}$ conditional upon $\{W(0), W(1)\}$ and t_i, and

c) a model for $\{Y_i(0), Y_i(1)\}$ given $\boldsymbol{x}_i$ and $\{W(0), W(1)\}$, e.g., the author's model (5) and (6).

Equation (9) is an attempt at supplying part b). For the reader's convenience we repeat that equation here:

$$W_{it_i}(z)|W_i(z), t_i \sim N(\gamma_{0i} + \gamma_1 t_i, \tau^2).$$

As the authors explain, the parameter γ_{0i} on the right hand side is $\gamma_0 + \delta_i$ where δ_i is "known and equal to $W_{it_i}(z) - \widehat{W}_{it_i}(z)$. This seems to make the conditional distribution $W_{it_i}(z)|W_i(z), t_i$ dependent on $W_{it_i}(z)$, which of course cannot really be true, so we asked the authors for some clarification. In email correspondence, Professor Dominici provided us with further details about how the birth weights at time 0 are predicted. The authors' methodology includes estimation of certain parameters "… outside the Gibbs sampling by fitting a regression on $(W_{it}(z_i), t_i)$ … ." It our belief that a fully Bayesian approach here would both be conceptually simpler and just as effective, if not more so. We still do not understand whether model (9) as implemented by the authors produces a distribution that is dependent only upon $(W_i(z), t_i)$. Also, it may be that this non-Bayesian component does not allow for all uncertainties to be reflected in the posterior, as a fully Bayesian analysis would allow. Figure 3 shows the potential danger of not using a fully Bayesian methodology. At the Case Studies Workshop, Professor Zeger mentioned that the authors fell back on (9) because of a convergence problem when the fitting was done inside the Gibbs sampler. We wonder whether this problem might be due to non-identifiability; see below.

We suggest replacing equation (9) by

$$W_{it_i}(z)|W_i(z), t_i \sim \text{Normal}\{W_i(z) + \gamma_1 t_i, \tau^2 t_i^{2\alpha}\}, \tag{1}$$

$\alpha > 0$, which allows for heteroscedastic measurement error, reflecting the belief that there should be more uncertainty about $W_i(0)$ the longer the time interval between birth and the time an infant is weighed. One advantage of (1) is that there is no need to fit a regression model outside the Gibbs sampler. Instead, the parameters in (1) and the unobserved birth weights at time 0 are sampled by the MCMC using information from all observed data, observed weights, covariates, and Y^{obs}. Any uncertainty due to prediction from this model is incorporated in the posterior. Another advantage of (1), one due to the heteroscedasticity, is that the conditional standard deviation τt_i^{α} is zero when $t_i = 0$, so our model satisfies the obvious "boundary condition" that $W_{it_i}(z)|_{t_i=0} = W_i(z)$. The obvious linear trends in Figure 2 suggest that γ_1 is well-determined from the data. It is unclear whether α can be accurately estimated without additional data or an informative prior. We would be surprised if there were no data in Nepal or a similar third world country where birth weights are measured longitudinally so that one could get informative priors for τ and α. Even data from a developed country might provide some useful information. However, locating and obtaining such data may be difficult, especially from a developing country, and in its absence we suggest a somewhat informative prior such as $\alpha \sim \text{Uniform}(.5, 1)$ or perhaps a beta distribution supported on [.5, 1] with more probability near 1. The values $\alpha = .5$ and $\alpha = 1$ seem to us to be extreme cases of the likely values of α. The former is implied by the model where $W_{it}(z)$ is a Brownian motion in t with drift that is constant across individuals. The latter is implied by model where each baby's weight grows linearly in t with a rate that varies among individuals and is normally distributed. Compared to the Brownian motion model, this model seems more realistic to us, which suggests that the prior put more probability near 1. Both models imply the linear mean in (1) as well as the normal distribution. The combination of Brownian motion and individual-specific drifts might be reasonably approximated by an α between .5 and 1. Of course, if we really believed that a random-drift Brownian motion model held, we could use this to obtain a variance function model.

The heteroscedasticity in model (1) has another advantage besides realism—it makes the parameter τ identified. To appreciate this, note that by equation (7) of the paper and (1),

$$\text{var}\{W_{it_i}(z)|\boldsymbol{x}_i, t_i\} = \tau^2 t_i^{2\alpha} + \sigma_z^2.$$

In the homoscedastic case where $\alpha = 0$, only the sum of $\tau^2 + \sigma_z^2$ is identified, not the individual components. Allowing $\alpha > 0$ is an example of model expansion inducing identifiability (Gustafson (2005)). As Gustafson shows, model expansion can be ill-advised if the parameters are near the region of non-identifiability. Thus, we would be concerned if we α was close to 0. However, as just explained, α should be close to 1 and certainly at least 0.5. Non-identifiability is a common problem when there is measurement error and no validation data or replications, but fortunately it can be avoided here by using a suitable model. We wonder whether the non-identifiability when $\alpha = 0$ could be the cause of convergence problems with the Gibbs sampler that the authors experienced and was mentioned above. Despite their identifiability here, τ and α may be poorly determined by the data, so we reiterate that longitudinal data on the growth of infants could be useful.

It might be mentioned in passing that measurement error model (1) is somewhat unusual. First, the measurement error seems to be in an outcome, not a covariate. Often response measurement errors can be ignored and incorporated into equation error. However, in this case, the measurement error has a bias depending on t_i, so correction for this bias is needed. Moreover, although weight is an outcome, it is also used as a covariate in the model for survival, so in that sense there is covariate measurement error.

Choosing a likely value of ρ does not seem easy, and will depend upon whether it is viewed as a conditional correlation, as in (7), or an unconditional correlation, as in (10)—we will assume the former. One possible way to view ρ is that it measures the heterogeneity in the response to the treatment. We gained some insight by considering the model

$$W_i(z) = \beta_0 + z^* \beta_z + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_x + z^* \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_{zx} + \omega_{i,1} + z^* \omega_{i,2}, \tag{2}$$

where $z^* = 1$ or -1 for treatment and control, respectively, $z_i^* \beta_z$ is the treatment main effect, $\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_x$ is the main effect of the covariates, $z^* \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_{zx}$ is the interaction between the covariates and treatment, $\omega_{i,1}$ is a random effect due to the infant, and $z^* \omega_{i,2}$ is an infant-treatment interaction. Assume that $\omega_{i,1}$ is Normal$(0, \sigma^2_{\omega,1})$, $\omega_{i,2}$ is Normal$(0, \sigma^2_{\omega,2})$, and $\omega_{i,1}$ and $\omega_{i,2}$ are independent. Then (2) is equivalent to model (8) of the paper with $\sigma_0^2 = \sigma_1^2 = \sigma^2_{\omega,1} + \sigma^2_{\omega,2}$ and

$$\rho = \frac{\sigma^2_{\omega,1} - \sigma^2_{\omega,2}}{\sigma^2_{\omega,1} + \sigma^2_{\omega,2}}. \tag{3}$$

Notice that if there is no interaction between infants and treatment, meaning that and $\sigma^2_{\omega,2} = 0$, then $\rho = 1$. For this reason, ρ may be even higher than the correlation between successive children with the same mother or even identical twins. However, ρ could be small if some children respond much more to the treatment than others. Depending upon the variance, $\sigma^2_{\omega,1}$, of the infant effects and the variance, $\sigma^2_{\omega,2}$, of the infant-treatment interactions, ρ might be less than 0.5 and perhaps even negative. Consideration of model (2) might help practitioners elicit a prior for ρ. Model (2) implies that $\sigma_0^2 = \sigma_1^2$ but it could be easily generalized to remove this constraint, e.g., to

$$W_i(z) = \beta_0 + z^* \beta_z + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_1 + z^* \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_2 + \exp(\theta z^*) \left\{ \omega_{i,1} + z^* \omega_{i,2} \right\}. \tag{4}$$

Equation (3) continues to hold under model (4), though $\omega_{i,1}$ and $\omega_{i,2}$ lose their simple interpretations, since if $\theta \neq 0$ then $W_i(1) + W_i(0)$ depends on $\omega_{i,2}$ as well as $\omega_{i,1}$ and $W_i(1) - W_i(0)$ depends on $\omega_{i,1}$ as well as $\omega_{i,2}$.

An interesting paper by Gustafson (2005) argues in favor of using an informative prior on non-identified nuisance parameters such as (ρ, ψ) rather that using a sensitivity analysis with several fixed values of the nuisance parameters. It would be interesting to see how this strategy would work on the authors' case study.

The authors use splines in several places. These splines could be natural cubic splines with three knots, as the authors have used in their work, but such splines have only three parameters and might not be flexible enough to fit the data. They could also be too flexible and overfit the data. It is difficult to know a priori whether more or less degrees of freedom are needed. An alternative is to use more knots, e.g., 5 to 8, and a penalty which allows the "effective" number of parameters to be determined by the data. The penalty can be imposed by an appropriate prior of the coefficients of spline basis functions. Then the smoothing parameter is a ratio of variances which are sampled during the MCMC. This not only allows for data-driven smoothing parameter, but also adjusts the inference for uncertainty about the smoothing parameter. Moreover, simultaneous confidence bands are easy within this framework, and simultaneous bands on derivatives allow inference about where the function is increasing or decreasing (Ruppert et al. (2003)). We have had much success with Bayesian penalized splines for nonparametric regression, both with and without measurement error (Ruppert et al. (2003); Berry et al. (2002)).

Although there is a large and increasing causal inference literature, many statisticians are not very familiar with the area. It would be helpful if the authors could more fully separate inferences that depend on counterfactuals from those that do not. Marginal treatment effects such as those illustrated in Figure 5 should be independent of the counterfactuals and therefore independent of assumptions about ρ and ψ. Figure 6 shows the average of treatment effects for subpopulations defined by both $\{W(1), W(0)\}$. These effects are shown under four pairs of assumed values of (ρ, ψ), and it appears that the assumed value of (ρ, ψ) has little effect on the stratified treatment effect. A first we thought this was a remarkable robustness property, but now this is not so clear to us. Perhaps the authors can comment.

The point we do not fully understand can be explained by considering the second from the left set of boxplots in Figure 6(a). These show that for the stratum with low birth weight, i.e., $W(0) < 2500$, and a large imputed treatment effect, i.e., $\{W(1) - W(0)\} > 50$, there is large beneficial imputed treatment effect. This is a wonderful result, to the extent that it is true. Remember that the treatment effects on both birth weight and survival are imputed. The imputed values should be consistent with the *marginal effects*, and, in fact, might be largely determined by the marginal effects. It seems to us that given the *marginal* treatment effects, those low birth weight babies with a large *imputed* effect on birth weight *must* have a large *imputed* effect on survival, regardless of (ρ, ψ). This is a issue where further discussion by the authors of the meaning of these results and perhaps cautions about their use would be welcome. Model (2) might be a useful aid to thinking about this issue. Model (2) decomposes the covariance between $W_i(0)$ and $W_i(1)$ into a component $\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}_x + z^*\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}_{zx}$ due to the covariate effects and a subject-specific component $\omega_{i,1} + z^*\omega_{i,2}$. The variance of $W_i(0)$ and $W_i(1)$ of course have the same types of decomposition. The components due to the covariates are identified, and, if they dominate, then conclusions will be robust to misspecification of (ρ, ψ); perhaps this is what is happening here.

In any application of statistics, the biggest mistake is to ask the wrong questions. Fortunately, this is a mistake that the authors have been careful to avoid. The key questions are,

1. does the treatment shift the lower end of the birth weight distribution upwards without shifting the upper end?

2. does the survival probability increase with increasing birth weight at the lower end of the birth weight distribution?

The authors have asked both questions and found positive answers. The answers depend upon the distribution of (birth weight, survival) marginally within the the treatment and control groups, and these distributions can be estimated because the assignment to control or treatment has been randomized. The joint distribution of (birth weight, survival) under both treatment and control within an individual is not needed to address these questions. Because of this, the conclusions of the paper in regards to questions 1. and 2. should be robust to misspecification about the counterfactuals.

References

Berry, S. A., Carroll, R. J., and Ruppert, D. (2002). "Bayesian Smoothing and Regression Splines for Measurement Error Problems." *Journal of the American Statistical Association*, 97: 160–169. 41

Carroll, R. J., Crainiceanu, C. M., Ruppert, D., and Stefanski, L. A. (2006). *Measurement Error in Nonlinear Models*. Boca Raton: Chapman & Hall/CRC Press, second edition. 37, 38

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology*. Boca Raton: Chapman & Hall/CRC Press. 38

— (2005). "On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables (with discussion)." *Statistical Science*, 20: 111–140. 39, 40

Richardson, S. (1998). "Measurement Error." In *Markov Chain Monte Carlo in Practice*, 401–418. London and New York: Chapman & Hall. 38

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press. 41

Acknowledgments

David Ruppert acknowledges support by NSF Grant DMS 04-538 and NIH Grant CA57030. Raymond Carroll was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106)