An approximate method for population toxicokinetic analysis with aggregated data

Weihsueh A. Chiu, Frédéric Y. Bois

▶ **To cite this version:**

Weihsueh A. Chiu, Frédéric Y. Bois. An approximate method for population toxicokinetic analysis with aggregated data. Journal of Agricultural Biological and Environmental Statistics, 2007, 12 (3), pp.346-363. 10.1198/108571107X229340 . ineris-00961914

HAL Id: ineris-00961914

https://ineris.hal.science/ineris-00961914

Submitted on 20 Mar 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An Approximate Method for Population Toxicokinetic Analysis With Aggregated Data

Weihsueh A. Chiu *
U.S. Environmental Protection Agency
National Center for Environmental Assessment
Washington, DC 20460, USA
and
Frédéric Y. Bois
Institut National de l'Environnement Industriel et des Risques
Unité de Toxicologie Expérimentale Parc Alata, BP2, 60550, Verneuil-En-Halatte, France

February 26, 2007

Abstract

Standard statistical models for analyzing inter-individual variability in clinical pharmacokinetics (non-linear mixed effects; hierarchical Bayesian) require individual data. However, for environmental or occupational toxicants only aggregated data are usually available, so toxicokinetic analyses typically ignore population variability. We propose a hierarchical Bayesian approach to estimate inter-individual variability from the observed mean and variance at each time point, using a bivariate normal (or lognormal) approximation to their joint likelihood. Through analysis of both simulated data and real toxicokinetic data from 1,3-butadiene exposures, we conclude that given information on the form of the individual-level model, useful information on inter-individual variability may be obtainable from aggregated data, but that additional sensitivity and

*Weihsueh A. Chiu is Environmental Health Scientist, U.S. Environmental Protection Agency (EPA), Washington, DC 20460, USA (email: chiu.weihsueh@epa.gov); Frédéric Y. Bois is Scientific Officer, Institut National de l'Environnement Industriel et des Risques Unité de Toxicologie Expérimentale Parc Alata, BP2, 60550, Verneuil-En-Halatte, France (email: Frederic.Bois@ineris.fr). Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA.

identifiability checks are recommended.

Keywords: 1-3 butadiene; Bayesian; inter-individual variability; Markov chain Monte Carlo simulation; population pharmacokinetics

1 INTRODUCTION

Population analyses of toxicokinetic data are designed to characterize inter-individual variability in the parameters or predictions of models describing the absorption, distribution, metabolism, and excretion of toxicants in the body. While such analyses are standard in pharmaceutical research (Sheiner, Rosenberg, and Melmon 1972; Sheiner 1984; Radine-Poon 1985; Yuh et al. 1994; Wakefield, Smith, Racine-Poon, and Gelfand 1994; Wakefield 1996; FDA 1999), it has only been since the work of Bois et al. (1996a, b) and Gelman, Bois, and Jiang (1996) that similar methods have been applied in environmental health sciences. However, there is only a handful of published population toxicokinetic analyses (recent examples include Hack, Chiu, Zhao, and Clewell 2006; Marino et al. 2006; Yokley et al. 2006; and Covington et al. 2007). Population analysis methods such as non-linear mixed effects or hierarchical Bayesian modeling require individual-level data. However, individual data from toxicokinetic studies (commonly performed in laboratory animals, but also sometimes in humans) are often unavailable due to the common practice of summarizing data in the form of average and SD (which are sufficient for some classical analyses).

Toxicokinetic analyses of older, often under-exploited, datasets therefore typically ignore population variability because models are fit using reported means and standard errors from multiple individuals, with those parameters interpreted as representing an "average individual." The increasing use of physiologically-based toxicokinetic (PBTK) models, which have many more parameters than classical pharmacokinetic models, has led to the additional practice of fixing "known" physiological variables and estimating the remaining chemical-specific parameters either from *in vitro* measurements or by "fitting" to aggregated data. There are several problems with such practices. First, except for the simplest models, the concept of

"average" parameters is difficult to interpret because of the non-linear relationships between model parameteres and predictions. Moreover, the general presumption that population variability can be ignored in laboratory animal experiments because they are performed on groups of in-bread, genetically similar strains is questionable. Such data still may show a fairly wide range of kinetic response, especially evident in experiments that report single data points for individual animals (e.g., Prout, Provan, and Green 1985). Furthermore, there is typically a large amount of both uncertainty and variability in parameters in PBTK models that are treated as "known," and the selection of which parameters to fix and which to fit is difficult, especially when combining information from various sources and multiple datasets. Therefore, the practice of fitting to aggregate data in this setting can at the very least underestimate overall uncertainty and may lead to inaccurate and biased estimates (Racine-Poon and Smith 1990; Sheiner 1984; Sheiner and Ludden 1992; Woodruff and Bois 1993).

We demonstrate here that population inferences may still be made from aggregated data if both the observed mean and variance at each time point are available and there is *a priori* information about the model and variance structures. In particular, we propose that the usual hierarchical Bayesian approach, briefly reviewed in Section 2, be extended by treating individual data as missing (or latent) and marginalizing over it. As described in Section 3, we approximate the corresponding (generally intractable) joint likelihood of the observed means and variances with a multivariate normal (or lognormal) distribution having the correct first and second-order moments. The form of the likelihood depends in part on the underlying measurement model for the individual data and the assumed variance structure. We derive approximations based on normal, proportional, and lognormal errors and a single level of inter-individual variability, which are commonly used in toxicokinetic models. In Section 4, we present comparisons of analyses based on individual data using the full population approach to analyses based on aggregated data using our proposed approximations for three simulated data sets and one published human toxicokinetic dataset of controlled exposures to 1,3-butadiene. We find in our examples that the aggregated analyses provide posterior

predictions quite similar to individual analyses, albeit with greater posterior uncertainty, as should be expected. In Section 5, we discuss our conclusion that substantial information on inter-individual variability may remain in the aggregated data, and that such information can be recovered through appropriate analyses. Given that some information is still lost in data aggregation, we suggest that posterior analyses, including checking of model fit, sensitivity, and parameter identifiability, are of great importance to increasing confidence in conclusions drawn from analyses of aggregated data.

2 POPULATION MODELING OF TOXICOKINETICS

As described in Gelman et al. (1996), population modeling of toxicokinetics involves setting up a model in several stages. A nonlinear toxicokinetic model, with predictions denoted f, describes the absorption, distribution, metabolism, and excretion of a compound and its metabolites in the body. This model depends on several, usually known, parameters such as measurement times t, exposure E, and measured covariates ϕ. Each subject i in a population has a set of unknown parameters θ_i. A population model describes their distribution in the population, and incorporates existing scientific knowledge about them through prior distributions on the population mean μ and variance Σ^2. Finally, a "measurement error" model describes deviations ϵ (with variance σ^2) between the data y and model predictions f. This level of the hierarchical model typically also encompasses intraindividual variability as well as model misspecification, but for notational convenience we refer to it here as "measurement error." All these components are illustrated graphically in the left part of Figure 1.

***Figure 1 about here.

The posterior distribution for the unknown parameters is obtained in the usual manner by multiplying (A) the prior distribution for the population mean and variance and the "measurement" error $P(\mu, \Sigma^2|I)P(\sigma^2|I)$, (B) the population distribution for the indi-

vidual parameters $P(\theta|\mu,\Sigma^2,I)$, and (C) the likelihood $P(y|\theta,\sigma^2,I)$, where for notational convenience, we collapse the knowledge of f, ϕ, E, t, and n into prior information I:

$$P(\theta,\mu,\Sigma^2,\sigma^2|y,I) \propto P(\mu,\Sigma^2|I)P(\sigma^2|I)P(\theta|\mu,\Sigma^2,I)P(y|\theta,\sigma^2,I) \tag{1}$$

Here, each individual's parameters θ_i have the same sampling distribution (i.e., they are *iid*), so their joint prior distribution is

$$P(\theta|\mu,\Sigma^2,I) = \prod_{i=1}^{n} P(\theta_i|\mu,\Sigma^2,I). \tag{2}$$

We consider three different measurement models for the likelihood function, normal errors (Model I), proportional errors (Model II), and lognormal errors (Model III), as shown in Table 1. Note that Models II and III are heteroscedastic, a common concern for toxicokinetic data. Different types of measurements $j = 1 \ldots m$ may have different errors, but errors are otherwise assumed to be *iid*. Since the individuals are treated as independent given $\theta_{1\ldots n}$, the total likelihood function is simply

$$P(y|\theta,\sigma^2,I) = \prod_{i=1}^{n}\prod_{j=1}^{m}\prod_{k=1}^{N_j} P(y_{ijk}|\theta_i,\sigma_j^2,t_{jk}) \tag{3}$$

where n is the number of individuals and m is the number of different types of measurements, N_j is the number of measurements of type j, and t_{jk} are the times at which measurements of type j were made. Note we have assumed that the individuals each have the same experimental design, as would be expected if data were to be aggregated.

3 ANALYSIS OF AGGREGATED DATA

If individual data have been aggregated, and one only has the number of individuals n and the sample mean m_{jk} and variance s_{jk}^2 of individual measurements at time-point k, then one must modify the statistical model used. The individual data y_{ijk} are considered missing or latent, and therefore treated as parameters rather than data in a Bayesian context. Thus, the standard data model becomes part of the population model, and a new data model for m and s^2 is needed. One therefore has a posterior distribution given by

$$P(\theta,\sigma^2,\mu,\Sigma^2,y|m,s^2,I) \propto P(\theta|\mu,\Sigma^2,I)P(\mu,\Sigma^2|I)P(\sigma^2|I)P(y|\theta,\sigma^2,I)P(m,s^2|y,I). \tag{4}$$

The additional term $P(m, s^2|y, I)$ is formally a δ-function (or 0-1 indicator) specifying the arithmetic relationship between the observed values y_{ijk} and their mean m_{jk} and variance s^2_{jk}. This full statistical model is illustrated in the middle of Figure 1. A possible treatment of the problem would be to consider the y_{ijk} as latent variables and sample (i.e., impute) them as the other estimands, through Monte Carlo techniques, for example. The logical relationship between y_{ijk} and (m_{jk}, s^2_{jk}) would complicate such a treatment for continuous measures (see Marjoram, Militor, Plagnol, and Taveré 2003 for discussion of such methods). The approach we take is to marginalize over the individual measured values y_{ijk}, which here may be considered nuisance parameters, *prior* to sampling via Monte Carlo. The posterior distribution we are aiming for, then, has the form

$$P(\theta, \sigma^2, \mu, \Sigma^2|m, s^2, I) \propto P(\theta|\mu, \Sigma^2, I)P(\mu, \Sigma^2|I)P(\sigma^2|I)P(m, s^2|\theta, \sigma^2, I), \tag{5}$$

with

$$\begin{aligned} P(m, s^2|\theta, \sigma^2, I) &= \prod_{j,k} P(m_{jk}, s^2_{jk}|\theta, \sigma^2_j, I) &(6)\\ &= \prod_{j,k} \int P(m_{jk}, s^2_j|y_{1jk} \ldots y_{njk}) \prod_i P(y_{ijk}|\theta_i, \sigma^2_j, I) dy_{ijk} &(7) \end{aligned}$$

This marginalization is illustrated graphically on the right part of Figure 1. Because the likelihood function $P(m_{jk}, s^2_j|\theta, \sigma^2_j, I)$ is conditional on θ, it is independent of the population model for inter-individual variability and only depends on the "measurement" model. Below, we present approximations to $P(m, s^2|\theta, \sigma^2, I)$ for the different measurement models considered above. Note that from here on we generally suppress the indices j, k for clarity, and concentrate on approximations for the integral in equation (7).

3.1 Likelihood Functions for Various Measurement Models

Our general approach is to approximate (for fixed observation type j and time-point t_k) the joint distribution of m and s^2 (m and $m_2 \equiv (m^2 + s^2)$ for Model III), conditional on $\{\theta_i\}$ $(i = 1 \ldots n)$, with a bivariate normal (lognormal for Model III) distribution by matching the first and second order moments. In particular, we require that the distributions match

in terms of $\mathrm{E}[m]$, $\mathrm{E}[m^2] - \mathrm{E}[m]^2$, $\mathrm{E}[s^2]$, $\mathrm{E}[s^4] - \mathrm{E}[s^2]^2$, and $\mathrm{E}[ms^2] - \mathrm{E}[m]\mathrm{E}[s^2]$, where E is the expectation conditional on the values of $\{\theta_i\}$. The results of these calculations are summarized in Table 1.

Table 1 about here.

The moment matching derivations for Models I and II are straight-forward, though tedious (and available as supplementary material online). Note that in both cases, the marginal distributions for the observed mean m are exactly normal, as they are weighted sums of normal deviates. The covariance between m and s^2 is zero for Model I, and non-zero for Model II.

The derivation for lognormal errors, Model III, merits additional discussion. In this case, the mean, given by $m = (1/n)\sum f_i \exp(\epsilon_i)$, has no simple closed form solution for its distribution (see Barakat 1976 and Leipnik 1991 for series approaches to calculating the characteristic function). Several approaches may be taken to approximate it. For large n and/or small σ^2, the central limit theorem can be invoked to approximate the distribution for m by a normal distribution. For many applications, however, n is quite small (~ 4), and Barakat (1976) shows that the coefficient of skewness of the sum distribution decays only as $\sim n^{-1/2}$. Moreover, m and $\sqrt{s^2}$ are often of the same order, so it may be important to incorporate the fact that m must be positive. For these reasons, in many telecommunications and engineering applications, the sum of lognormal deviates is commonly approximated by a lognormal distribution by matching moments (e.g., Fenton 1960). Simulations appear to indicate that this approximation is useful for values up to $\sigma^2 \sim 1$ (Fenton 1960; Schwartz and Yeh 1982), and we adopt this approximation here.

So as to enable use of the same approach, we take as our second "observed" value not the variance s^2 but the second moment $m_2 \equiv (1/n)\sum f_i^2 \exp(2\epsilon_i) = s^2 + m^2$, as it is also a sum of lognormal deviates. Thus, we also approximate its distribution by a lognormal distribution by matching moments, noting that it is the same form as m with $\sigma^2 \to 4\sigma^2$ and $f_i \to f_i^2$. From the work cited above, this means that the lognormal approximation for m_2 is only useful for $4\sigma^2 \leq 1$. Therefore, in general the approximate likelihood for (m, m_2) is

only accurate if $\sigma \leq 0.5$.

The final step, then, is to define the correlation coefficient r of the bivariate lognormal distribution. We use the same approach as for Models I and II, matching the product moment $\mathrm{E}[mm_2]$ (equivalent to matching the central moment because we have already matched $\mathrm{E}[m]$ and $\mathrm{E}[m_2]$). We have investigated this approximation in a limited number of simulations for values of $n \leq 10$ and values of $\sigma \leq 0.5$, and have found it to be acceptable as long as the f_i has a relatively small coefficient of variation. If f_i are too widely dispersed (e.g., $\mathrm{VAR}[\ln f] > 1$), then the above formula can sometimes give a value of $r > 1$. In this case, however, we have found that using the above formula, but taking $\sigma \to 0$ (i.e., the limit in which measurement error is negligible relative to inter-individual variability), so that $r \to \mathrm{E}[f^3]/\sqrt{\mathrm{E}[f^2]\mathrm{E}[f^4]}$, gives reasonable results. We take the minimum value of r from the above two formulae as the one we use in our distribution for m and m_2, as shown in Table 1. We find that this approach sometimes slightly underestimates r by up to a few percent, but ensures that $r \leq 1$. This constraint on r should be checked in posterior simulations.

3.2 Additional Issues for Consideration

3.2.1 Sufficiency and Identifiability

For most non-linear models, there will undoubtedly be some loss of information in data aggregation, so there may be concerns about whether there is still sufficient information to ensure parameter identifiability in the toxicokinetic and population models. However, even with individual data, a toxicokinetic model may not be fully identifiable (e.g., the estimation of decay times of a mixture of exponentials is ill-conditioned, Acton 1970), so this problem is not unique to aggregated data. At the very least, some *a priori* information on the system being analyzed should exist to motivate the formulation of the model and, preferably, informative prior distributions for its parameters. For instance, in their analysis of tetrachloroethylene toxicokinetics, Gelman et al. (1996) remarked that both the physiological model and prior distributions were necessary to ensure identifiability. At best, data from other experiments

where data were not aggregated could be used to establish the appropriate model form. In toxicokinetics, if one's goal is to characterize population variability, then one presumably already has sufficient information to justify the structure of the individual-level model. So we focus in particular on parameter identifiability rather than model discrimination.

For non-linear models, these issues are necessarily case-specific, and it may not be possible to know whether parameter identifiability is a problem prior to performing an analysis. However, posterior and sensitivity analyses can be done to check that the population parameters are identifiable given the data. The first evidence for non-sufficiency would be if the priors and posterior distributions are identical. If the posteriors are narrower, then the data add some information. A scalar measure τ of the "overlap" between marginal prior and posterior distributions was proposed by Garrett and Zeger (2000) for latent class models, and is adapted here for more general use. In particular, for a parameter θ, data y, and prior information I, prior distribution $P(\theta|I)$, and posterior distribution $P(\theta|y, I)$, the "overlap diagnostic" is defined as

$$\tau = \int \min \{P(\theta|I),\ P(\theta|y, I)\}\, d\theta. \tag{8}$$

Values of τ near unity (i.e., 100% overlap) indicate weak identifiability from the data, whereas lower values (Garrett and Zeger 2000 proposed 0.35 as a heuristic threshold) indicated strong identifiability from the data.

Additionally, it can be determined whether parameters are *uniquely* identifiable. For example, in the case of a model $y_i(t) = (\alpha_i + \beta_i)t + \epsilon_{it}$, the parameters α_i and β_i (and their corresponding population mean and variance) are not uniquely identifiable without informative priors, whether one has individual data or aggregated data. If one does not have informative priors, then this can be detected *a posteriori* by examining sensitivity of the results to different diffuse priors, and more directly through examining correlations in the posterior samples. While visual inspection of the posterior correlation matrix can be useful in simple cases such as this, in more realistic applications in which more than two parameters may be involved simultaneously, principal component analysis (PCA) should be a useful tool.

In particular, performing PCA on the posterior samples, and then comparing the posterior distributions of the principal components with *their* priors (i.e., applying the same centering, scaling, and rotation to samples from their joint *prior* distribution) should identify any non-diagonal components which are not identifiable. This should also be applicable to cases such as $y_i(t) = \theta_i + \epsilon_{it}$, where variance contributions from different levels of the hierarchical model are indistinguishable after data aggregation.

3.2.2 Choice of Measurement Model

In standard population analyses, the choice of measurement model is influenced only by the underlying hypothesis for how the individual data deviate from model predictions. In our case, how the data are aggregated also plays a role. For most toxicokinetic data, a lognormal measurement model has been traditionally assumed for individual data points. One could thus either use Model I by transforming $y \rightarrow \ln y$, or use Model III. *A priori,* Model I would be preferable because it is simpler. However, one could only use this model if the aggregation was done on the transformed data — e.g., the reported values are *geometric* mean and standard deviation. Typically, it is the *arithmetic* mean and standard deviation that are reported, so that model III would be the most appropriate. Unfortunately, model III is also the least robust of the models because of the requirement for $\sigma^2 \leq 0.5$ and the possible lack of stability of the derived correlation coefficient. Model II then is a more robust alternative approximation that still allows for errors to be proportional to the measured value. On a practical matter, data are probably most useful in a regime where proportional normal errors (Model II) and lognormal errors (Model III) are not easily distinguishable. Moreover, the measurement model also encompasses model misspecification, and a value of $\sigma > 0.5$ (i.e., a $> 50\%$ error) may be reason to rethink the model.

3.2.3 Model Checking and Model Choice

As with standard population analyses, model checking is important here. In addition to the identifiability checks described above, the most basic check is whether the model is consistent with the data. In the standard population analysis, a typical method to perform this check is to compare simulated (replicated) data $y_{i,\text{rep}}$ with the observed data $y_{i,\text{obs}}$ (Gelman, Carlin, Stern, and Rubin 2003, chap. 6). Analyses of aggregated data are no different except that the quantities being compared are $(m, s^2)_{\text{rep}}$ and $(m, s^2)_{\text{obs}}$, as in the examples we give below.

One important limitation of aggregated analyses, however, is that the structural assumptions in the population model $P(\theta|\mu\Sigma^2)$ and the measurement model $P(y|\theta, \sigma^2)$, which are not easy to check with individual data, are even more difficult to validate with only aggregated data. Typical techniques for assessing these assumptions include both posterior simulations as well as sensitivity analyses. For aggregated analyses, posterior simulation is of somewhat limited usefulness when checking structural assumptions because of the additional layer of latency. Sensitivity analyses take on greater importance. We illustrate this below by considering all three measurement models in our analyses.

On the related issue of model choice, there are a large number of statistical methods for model selection in both a frequentist (e.g., log-likelihood ratio test, Akaike information criterion) and Bayesian context (e.g., Bayes factors, Bayesian information criterion). These methods may be less reliable with only aggregated data. Fortunately, the formulation of toxicokinetic models is motivated primarily by *a priori* biological and chemical information rather than statistical measures.

In checking and choosing models, it is also important to make a distinction between the "statistical" and "practical" significance of model errors. That is, because models such as those used in toxicokinetics can never be thought of as strictly "true," statistical lack of fit may or may not have an impact on the substantive inferences for which the model is used. Thus, the use of the model should be kept in mind when assessing either the consistency between model predictions and data or the impact of different model assumptions.

4 APPLICATION TO DATA ON 1,3-BUTADIENE

We first performed a number of simulations with a simple model and simulated data, as summarized in Table 2. These simulations covered all three measurement models. We generated simulated data and compared full population analysis of the individual data with analyses of the same data aggregated. Computations were performed using WinBUGS version 1.4 and MCSim (Version 5.0.0, Bois, Maszle, Revzan, Tillier, and Yuan 2005). In each case, the results of the analysis of aggregated data were quite consistent with the results of the full population analysis as well as with the underlying "true" values from which the data were generated. This offered support for the accuracy of the approach developed here.

***Table 2 about here.

4.1 Butadiene data and model

We then applied our aggregation model to actual toxicokinetic data. The data and model are described in detail in Bois, Smith, Gelman, Chang, and Smith (1999), and are summarized briefly here. Eight human volunteers were recruited and tested at National Cheng Kung University in Taiwan. The tests were conducted, under informed consent, with an Institutional Review Board-approved human subjects protocol. They were exposed to an ambient concentration $C_{in}(t)$ of five ppm of 1,3-butadiene for two hours and then zero thereafter. This exposure was the minimum that could be precisely measured, and was well below Taiwan's allowable occupational exposure of 10 ppm per eight hour work day for a working lifetime. For each individual, measurements of body weight (BDW), minute-ventilation rate K_{in}, and blood-air partition coefficients P_{ba} were made along with exhaled breath measurements $C_{ex}(t_k)$ at a series of times t_k from the beginning of exposure to about one hour post-exposure. The full dataset for $C_{ex}(t_k)$ is displayed in all three panels of Figure 2.

***Figure 2 about here.

The toxicokinetic model was a standard two-compartment model, with a central (volume V_c) and peripheral compartment (volume V_p) governed by the following differential equations

for the amount of 1,3-butadiene in each respective compartment $Q_c(t)$ and $Q_p(t)$:

$$\dot{Q}_c(t) = K_{in}C_{in}(t) + \frac{K_{cp}Q_p(t)}{P_{pc}V_p} - \left(K_{cp} + \frac{K_{in}}{P_{ca}}\right)\frac{Q_c(t)}{V_c} - K_{met}Q_c(t) \quad (9)$$

$$\dot{Q}_p(t) = K_{cp}\frac{Q_c(t)}{V_c} - \frac{K_{cp}Q_p(t)}{P_{pc}V_p}, \quad (10)$$

where, in addition to the parameters defined above, P_{ca} is the central-to-air partition coefficient (assumed to be the measured blood-air partition coefficient P_{ba}); K_{cp} is the rate constant for distribution from central to peripheral compartments; P_{pc} is the peripheral to central partition coefficient; and K_{met} is the metabolic rate constant. The measurement of covariates BDW, K_{in}, and P_{ba} greatly improves parameter identifiability in this model.

The exhaled concentration at observed time t_k is given by

$$C_{ex}(t_k) = \frac{0.7\,Q_c(t_k)}{P_{ca}V_c} + 0.3\,C_{in}(t_k), \quad (11)$$

where a physiological dead space of 30% is assumed. Because the product $P_{pc}V_p$ is not separable, it is treated as a single parameter to ensure identifiability. In addition, the central volume and the minute-volume are assumed to follow the scaling relations

$$V_c = sc_V_c\,BDW \quad (12)$$

$$K_{in} = sc_K_{in}\,BDW^{0.7}, \quad (13)$$

so the sc_V_c and sc_K_{in} are the actual parameters in the model. Finally, one of the important uses of the model is the prediction of the amount metabolized AMET, which is given by $\text{AMET} = \int K_{met}Q_c(t)\,dt$. The characterization of the uncertainty and variability in dose metrics such as this is important for making inferences about the population distribution of risks from exposure.

In the population model, all the individual-level parameters are assumed to be lognormally distributed with geometric mean $\exp(\mu)$ and log-variance parameter Σ^2. Two modifications from the Bois et al. (1999) full population analysis were made to allow comparison between analyses of individual and aggregated data. First, population body weight parameters were estimated as part of the model. The population model was lognormal, but the

likelihood function was assumed to be normal with a measurement standard deviation of 0.25 kg. Second, only time points for which measurements were available in all 8 individuals were included. Although it is possible to include the missing data points in the model, it adds a level of complexity that obscures the point of this comparison, which is to compare full population analysis with an analysis using only aggregated data.

To summarize, each individual i has seven parameters $\theta_i = (sc_V_c,\ K_{cp},\ P_{pc}V_p,\ K_{met},\ P_{ca},\ sc_K_{in},\ BDW)_i$ and four types of observations $y_i = (P_{ba,obs},\ K_{in,obs},\ C_{ex,obs}(t_k),\ BDW_{obs})_i$. All errors were assumed lognormal except for BDW, for which errors were assumed to be normal:

$$P_{ba,obs} = P_{ca}\exp(\epsilon_{P_{ca}}) \tag{14}$$

$$K_{in,obs} = K_{in}\exp(\epsilon_{K_{in}}) \tag{15}$$

$$C_{ex,obs}(t_k) = C_{ex}(t_k)\exp(\epsilon_{C_{ex}}) \tag{16}$$

$$BDW_{obs} = BDW + \epsilon_{BDW} \tag{17}$$

where in the middle two cases, the model predictions are given by equations (13) and (11), respectively, and $\epsilon_{P_{ca}} \sim N(0, \log 1.17)$, $\epsilon_{K_{in}} \sim N(0, \log 1.02)$, $\epsilon_{C_{ex}} \sim N(0, \log GSD_{ex})$, and $\epsilon_{BDW} \sim N(0, 0.25\text{ kg})$ (the first two were based on replicate samples).

Data aggregation was performed using log-transformed measurements (thus using Error Model I) as well as un-transformed measurements (thus using Error Models II and III), the last of which are also shown in the left panel of Figure 2. Simulations for all cases were performed using the MCSim software. To ensure consistency, the original MCSim code used in Bois et al. (1999) was obtained and modified only where needed.

Prior distributions were the same as in Bois et al. (1999), except for the body weight parameters, which are new. The prior mean body weight was assigned a uniform distribution with min and max equal to the min and max measured body weight. The prior variance was assigned the inverse Gamma distribution with shape parameter of unity and scale parameter corresponding to a 20% coefficient of variation. All priors are listed in Table 3. As was done in Bois et al. (1999), 50,000 iterations performed for two independent chains in each

case. The first 10,000 iterations were discarded, and only every 10 iterations were stored for analysis. Convergence was monitored through the method of Gelman and Rubin (1992), and potential scale reduction factors were ≤ 1.03 in the case of individual data, and ≤ 1.01 in the aggregated data cases.

4.2 Comparisons between full population and aggregated analyses

***Figure 3 and Table 3 about here.

Table 3 and Figure 3 summarize the statistical results for both full population and aggregated analyses. There is substantial agreement between the different analyses, as shown visually in Figure 3. Posterior estimates of all population parameters substantially overlap for all parameters except for the "measurement" error, which is signficantly larger in the aggregated analyses (discussed below). While difficult to see in the Figure due to the logarithmic scale, estimates of the population mean parameters (μ), summarized in Table 3, are not greatly affected by the use of aggregated data, both in terms location and scale. In most cases, the population variances (expressed as geometric standard deviations $\exp \Sigma$ in Table 3) are only slightly affected as well, with 95% confidence intervals substantially overlapping among the four analyses. The uncertainties in the population variance parameters $\exp \Sigma$, however, are consistently greater in the aggregated analyses. Because the inverse-Gamma priors on the population variances Σ^2 have shape parameter of unity, they have infinite dispersion. Thus, greater posterior uncertainty generally leads to higher central estimates and upper confidence limits. This is reflected in the results, particularly for parameters sc_V_c and sc_K_{in}. It is clear that in these cases, some information was lost in the aggregation process. The parameter for "measurement" error GSD_{ex} was slightly greater in the aggregated analyses. Bois et al. (1999) reported that the analytical errors were estimated to be about 7%, so this increased "measurement" error reflects additional model error (e.g., due to the approximations necessary to derive the likelihood functions) and/or intraindividual variability. Table 3 also shows the posterior uncertainty in the population variability in the

dose metric AMET for the 8 individuals. The posterior predictions were very similar among the full population and aggregated analyses, although as with the population parameters, the uncertainty from the aggregated analyses is slightly greater. Overall, the results of the population analysis based on aggregated data, using the statistical models developed above, are consistent (albeit with greater overall uncertainty) with those based on individual data.

4.3 Parameter identifiability

Checks for non-identifiability were conducted on the full population analysis as well as each aggregated analysis, concentrating on the 15 population parameters (mean and variance for each of the 7 model parameters, plus residual "measurement" variance). Given the narrowing of all of these distributions from prior to posterior, one would not expect any parameters to be completely unidentifiable. Indeed, for population means, values for the overlap diagnostic τ were ≤ 0.2 except for the metabolic rate constant K_{met} ($\tau \sim 0.5$) and the mean body weight ($\tau \sim 0.6$). Moreover, the difference between τ in individual and aggregated analyses was ≤ 0.05, indicating little information loss for estimated population means. Population variances were not as well identified, with $\tau \geq 0.4$. In addition, the differences between individual and aggregated analyses were greater. The largest changes in overlap were in the estimated variances of V_c ($\tau = 0.55$ for the individual analysis and $\tau = 0.83 - 0.90$ for the aggregated analyses) and P_{pc} ($\tau = 0.46$ for the individual analysis and $\tau = 0.61 - 0.64$ for the aggregated analyses). For the other variance parameters, the overlap diagnostic τ changed by ~ 0.1.

Further checks examined whether some parameter combinations are only weakly constrained by the data. As a visual check, the correlation matrix was calculated with mean parameters log-transformed (this is more "natural" since the population model and most of the likelihoods are lognormal) and variability parameters transformed, if necessary, to variances (e.g., $GSD \rightarrow (\ln GSD)^2$). All non-diagonal correlation coefficients for the population parameters were in the interval (-0.4,0.6); of the $15 \times 14/2 = 105$ unique non-diagonal

elements, only 8 were outside the interval [-0.2,0.2]. Thus, there were some correlations, but none were extremely strong.

PCA was performed on the posteriors with the R statistical package, with centering to zero mean and scaling to unit variance (i.e., equivalent to determining the eigenvectors of the correlation matrix). Each analysis (full population and the three aggregated) had a slightly different rotation matrix, as should be expected for a non-linear model and the approximate likelihood functions used. Priors for each set of principal components were then generated by applying the same transformations to random samples from the joint prior distribution. All prior principal component [2.5%,97.5%] confidence intervals included zero (which, by definition, is the posterior mean of the principal components), so there is no conflict between priors and posteriors even after transformation. Moreover, the posterior confidence intervals were *wholly* contained within the prior confidence intervals, with the prior intervals substantially wider (i.e., by at least 5.9-fold). Thus, we conclude that there are no substantial parameter identifiability problems.

4.4 Checking Model Fit

Inspection of Figure 2 (middle and right panel) of the scatter in the data points relative to the scatter of the predictions suggests that full population and aggregated analyses give similar inter-individual variance in their predictions. Additional checks using the full posterior distribution of predictions for the measured mean and variance for the exhaled air concentration C_{ex} showed good qualitative consistency between the model predictions and the underlying (aggregated) data (not shown). However, the scatter during the period of exposure ($t = 0 - 120$ minutes) appears underestimated, perhaps due to intraindividual variability, which here was not modeled separately but lumped with "measurement" error. This feature also appears in the full population analysis, as can be seen in the middle panel of Figure 2. We also note that, for Model III, the posterior range of the "measurement" error variance $\sigma < 0.5$, as required by our approximation for the likelihood function. In addition,

for Model III, we checked the approximation for the correlation parameter r, and found that in no case did the moment-matching formula give $r > 1$, so the $\sigma \to 0$ approximation was never used.

Sensitivity analysis is illustrated through comparison of the results from different error models. Summary statistics are quite similar, with the exception of the variance of sc_V_c, which had greater uncertainty in Model II. In particular, the posterior estimates for the dose metric AMET were remarkably consistent. Furthermore, graphs of the posterior distributions as well as the comparisons between data and model predictions showed little sensitivity to different error models (not shown).

Finally, we note that checks for parameter identifiability (§4.3), model fit and sensitivity analyses are not unique to the analysis of aggregated data. They could (and should, in our opinion) be more widely applied to standard population analyses as well. For instance, parameter identifiability is usually not checked in any formal manner. Model fit is typically checked only through scatter plots of data and a single posterior prediction (e.g., using a "random sample" or using population mean parameters), as is shown in Figure 2. As reported above, we performed additional checks using the full posterior distribution. Checking of the assumed "measurement" error model, as was done here as part of the sensitivity analyses, is rarely done.

5 DISCUSSION

Through the use of some conceptually simple approximations, we have developed likelihood functions for the observed sample mean and variance of individual measurements, given a hierarchical population model. Our results illustrate that individual data, while *preferable,* are not necessarily *essential* to analyze population variability in a hierarchical Bayesian framework. In the cases we analyzed, the resulting posterior distributions between analysis of individual and aggregated data are very similar. This may not be too surprising with very simple normal models. Yet, we have found in our examples that the mean and

variance are nearly sufficient even with a nonlinear model. However, it is important to check model fit, sensitivity to the likelihood approximation used, and parameter identifiability. For this last check, we propose principal component analysis as a broadly applicable method. We have not yet applied our approach to more complex (e.g., PBTK) models, but results from the butadiene example are encouraging.

There are several important limitations to our approach, some of which are fundamental and some more practical. First of all, in the examples we have examined, the assumption has been made (by design for the simulated data and checked in the case of the butadiene analysis) that other sources of variability (intraindividual variability, measurement error, model uncertainty, for example) are small compared to inter-individual variability. If this assumption were not true, the analysis may not be able to disentangle inter- and intra-individual variability. However, this can be checked *a posteriori* with the identifiability checks described above. In addition, structural assumptions, which are not easy to test even with individual data, are even more difficult to check with aggregated data, so use of aggregated data for statistical model discrimination is not recommended. Fortunately, for toxicokinetic models there is usually *a priori* information on model formulation. Finally, one should always consider evaluations of the performance of aggregated analyses in the context of what the data are to be used for, as a mild loss of information or parameter non-identifiability may not necessarily have a significant impact on posterior inferences of interest. In our example with 1,3-butadiene, the predicted mean and variance for the dose metric AMET differed very little among the different analysis, even though some information on population variability was lost.

On a practical level, the analysis of aggregated data is more computationally burdensome than if individual data were available. This is due to the more complicated likelihoods, particularly for Models II and III where there is covariance, and possibly slower convergence. For instance, on an Intel Pentium 4 2.8 GHz processor running Windows XP with 512 MB of RAM, two chains of length 50,000 run in MCSim took about 1 hour to complete for the individual data for butadiene, and about 4.5 hours for the aggregated data using Model

III. While these times are not long, more complicated models and data could substantially increase the computation time and the chain length necessary for convergence. These practical limitations could be alleviated by either more efficient MCMC algorithms or faster computing power than we have used here.

Ideally, of course, individual data would always be available for analysis, and it is sound practice to use all such data. It should not be inferred from our analysis that summary data should be used when individual data are available, as data aggregation always entails a degree of information loss. However, especially in a toxicological/environmental health setting where data are often gathered from multiple, usually historical, sources, the original data may be unavailable. The effort, then, should be to maximize the use of the available information, particular in the case of human data where unnecessary exposure to toxicants should be minimized. Typically, no attempt is made to extract population variability information from aggregated data, perhaps because it is presumed to be unimportant or to have been lost in the aggregation process. Our analysis shows this information is not necessarily completely lost — that when both the mean and variance are reported, significant information on population variability may remain. We have presented here an example of application to the analysis of toxicokinetic data, but aggregation of data for publication has been pervasive in biology, and occasions to test the approach we propose should be plentiful. For example, in cancer bioassays, groups of animals are exposed to predetermined doses of a carcinogen and the onset of tumors is observed. Each animal is expected to react differently to the exposure. Such inter-individual variability can be studied if time-to-event reporting of tumors is available. However, most of the time (e.g. the carcinogenic potency database of Lois Gold at Berkeley, Gold, Manley, Slone, and Rohrbach 1999) only the number of animals bearing tumors at the end of the experiment is kept and analyzed with binomial or Poisson regressions. A better understanding of the variabilities and uncertainties involved in such experiments would probably benefit society as a whole.

ACKNOWLEDGEMENTS

This work is partially supported by the Ministry of the Environment and Sustainable Development (BCRD-AP2004-DRC05). The authors would like to thank B. Amzal, C. Chen, T. Choi, C. Diack, J. Fox, H. Kahn, E. Parent, W. Setzer, R. Subramaniam, and P. White for helpful discussions and comments, as well as the editors and referee whose suggestions improved this paper.

References

[1] Acton, F.S. (1970), *Numerical Methods That Work*, New York: Harper and Row.

[2] Barakat, R. (1976), Sums of independent lognormally distributed random variables. *Journal of the Optical Society of America*, **66,** 211-216.

[3] Bois, F.Y., Gelman, A., Jiang J., Maszle, D., Zeise, L., and Alexeef, G. (1996), Population Toxicokinetics of tetrachloroethylene. *Archives of Toxicology*, **70,** 347-355.

[4] Bois, F.Y., Maszle, D.R., Revzan, K., Tillier, S., Yuan Z. (2005), MCSim: a Monte Carlo Simulation Program, version 5.0.0. available at `http://fredomatic.free.fr/page_mcsim.html` (visited 1/22/07).

[5] Bois, F.Y., Smith, T.J., Gelman, A., Chang, H.Y., and Smith, A.E. (1999), Optimal design for a study of butadiene toxicokinetics in humans. *Toxicological Sciences*, **49,** 213-224.

[6] Covington TR, Robinan Gentry P, Van Landingham CB, Andersen ME, Kester JE, Clewell HJ. (2007), The use of Markov chain Monte Carlo uncertainty analysis to support a Public Health Goal for perchloroethylene. *Regulatory Toxicology and Pharmacology*, **47,** 1-18.

[7] Fenton, L.F. (1960), The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions of Communications Systems*, **8,** 57-67.

[8] Garrett, E.S., Zeger, S.L. (2000), Latent class model diagnosis. *Biometrics*, **56,** 1055-1067.

[9] Gelman, A., Bois, F., Jiang, J. (1996), Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions. *Journal of the American Statistical Association*, **91,** 1400-1412.

[10] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004), *Bayesian Data Analysis (2nd ed.)*, Boca Raton, FL: Chapman & Hall/CRC.

[11] Gelman, A., and Rubin, D. (1992), Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, **7,** 457-511.

[12] Gold, L. S., Manley, N. B., Slone, T. H., and Rohrbach, L. (1999), Supplement to the Carcinogenic Potency Database (CPDB): Results of Animal Bioassays Published in the General Literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996. *Environmental Health Perspectives*, **107,** (suppl. 4), 527-600.

[13] Hack CE, Chiu WA, Jay Zhao Q, Clewell HJ. (2006), Bayesian population analysis of a harmonized physiologically based pharmacokinetic model of trichloroethylene and its metabolites. *Regulatory Toxicology and Pharmacology*, **46,** 63-83.

[14] Leipnik, R.B. (1991), On Lognormal Random Variables: I-The Characteristic Function. *Journal of the Australian Mathematical Society B*, **32,** 327-347.

[15] Lin Y.S., Smith T.J., Wang P.Y., (2002), An automated exposure system for human inhalation study. *Archives of Environmental Health*, **57,** 215-223.

[16] Marino DJ, Clewell HJ, Gentry PR, Covington TR, Hack CE, David RM, Morgott DA. (2006), Revised assessment of cancer risk to dichloromethane: part I Bayesian

PBPK and dose-response modeling in mice. *Regulatory Toxicology and Pharmacology*, **45,** 44-54.

[17] Marjoram, P., Militor, J., Plagnol, V., and Tavaré, S. (2003), Markov Chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, **100,** 15324-15328.

[18] Prout. M.S., Provan, W.M., and Green, T. (1985), Species Differences in response to Trichloroethylene. *Toxicology and Applied Pharmacology*, **79,** 389-400.

[19] Racine-Poon, A. (1985), A Bayesian approach to nonlinear random effects models. *Biometrics*, **41,** 1015-1023.

[20] Racine-Poon, A. and Smith, A.F.M. (1990), Population Models. In *Statistical Methodology in the Pharmaceutical Sciences*, D.A. Barry (ed), 139-162. New York: Decker.

[21] Schwartz, S., and Yeh, Y.S. (1982), On the distribution function and moments of power sums with log-normal components. *Bell System Technical Journal*, **61,** 1441-1462.

[22] Sheiner, L.B. (1984), The population approach to pharmacokinetic data analysis: rationale and standard data analysis methods. *Drug Metabolism Reviews*, **15,** 153-171.

[23] Sheiner, L.B., and Ludden T.M., (1992), Population pharmacokinetics/dynamics. *Annual Review of Pharmacology and Toxicology*, **32,** 185-209.

[24] Sheiner, L.B., Rosenberg, B., Melmon, K.I. (1972), Modeling of Individual Pharmacokinetics for Computer-Aided Drug Dosage. *Computers and Biomedical Research*, **5,** 441-459.

[25] U.S. Food and Drug Administration (FDA). (1999), Guidance for Industry: Population Pharmacokinetics.

[26] Wakefield, J.C. (1996), The Bayesian Analysis of Population Pharmacokinetic Models. *Journal of the American Statistical Association*, **91,** 62-75.

[27] Wakefield, J.C., Smith, A.F.M., Racine-Poon, A., and Gelfand, A.E. (1994), Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Journal of the Royal Statistical Society C*, **43,** 201-221.

[28] Woodruff, T.J., and Bois, F.Y. (1993), Optimization issues in physiological toxicokinetic modeling: a case study with benzene. *Toxicology Letters*, **69,** 181-196.

[29] Yokley K, Tran HT, Pekari K, Rappaport S, Riihimaki V, Rothman N, Waidyanatha S, Schlosser PM. (2006), Physiologically-based pharmacokinetic modeling of benzene in humans: a Bayesian approach. *Risk Analysis*, **26,** 925-943.

[30] Yuh, L., Beal, S., Davidian, M., Harrison, F., Hester, A., Kowalski, K., Vonesh, E., Wolfinger, R. (1994), Population pharmacokinetic/pharmacodynamic methodology and applications: a bibliography. *Biometrics*, **50,** 566-575.

Table 1: Approximate Likelihood Functions for Aggregated Data

Error Model	Approximate Likelihood Function
I: $y_i \sim N(f_i, \sigma)$	$(m, s^2) \sim N(\mu_m, \sigma_m, \mu_{s^2}, \sigma_{s^2}, r)$
	$\mu_m = \mathrm{E}[f]$
	$\sigma_m^2 = \sigma^2/n$
	$\mu_{s^2} = \mathrm{E}[f^2] - \mathrm{E}[f]^2 + (n-1)\sigma_m^2$
	$\sigma_{s^2}^2 = 4\sigma_m^2 \left\{\mu_{s^2} - (n-1)\sigma_m^2/2\right\}$
	$r = 0$
II: $y_i \sim N(f_i, f_i\sigma)$	$(m, s^2) \sim N(\mu_m, \sigma_m, \mu_{s^2}, \sigma_{s^2}, r)$
	$\mu_m = \mathrm{E}[f]$
	$\sigma_m^2 = \mathrm{E}[f^2]\sigma^2/n$
	$\mu_{s^2} = \mathrm{E}[f^2] - \mathrm{E}[f]^2 + (n-1)\sigma_m^2$
	$\sigma_{s^2}^2 = 4\sigma^2 \left(\mathrm{E}[f^4] - 2\mathrm{E}[f^3]\mathrm{E}[f] + \mathrm{E}[f^2]\mathrm{E}[f]^2\right)/n$
	$\quad +2\sigma^4 \left\{(n-2)\mathrm{E}[f^4] + \mathrm{E}[f^2]^2\right\}/n^2$
	$r = 2\sigma^2 \left(\mathrm{E}[f^3] - \mathrm{E}[f^2]\mathrm{E}[f]\right) / \left(n\sqrt{\sigma_m^2\sigma_{s^2}^2}\right)$
III: $\ln y_i \sim N(\ln f_i, \sigma)$	$(\ln m, \ln m_2) \sim N(\mu_m, \sigma_m, \mu_{m_2}, \sigma_{m_2}, r)$
	$\mu_m = \ln \mathrm{E}[f] + (\sigma^2 - \sigma_m^2)/2$
	$\sigma_m^2 = \ln\{1 + (\exp\sigma^2 - 1)\mathrm{E}[f^2]/(\mathrm{E}[f]^2 n)\}$
	$\mu_{m_2} = \ln \mathrm{E}[f^2] + (4\sigma^2 - \sigma_{m_2}^2)/2$
	$\sigma_{m_2}^2 = \ln\{1 + (\exp 4\sigma^2 - 1)\mathrm{E}[f^4]/(\mathrm{E}[f^2]^2 n)\}$
	$r = \min\left[\mathrm{E}[f^3]/\sqrt{\mathrm{E}[f^2]\mathrm{E}[f^4]}\ ,\right.$
	$\quad \left.\ln\{1 + (\exp 2\sigma^2 - 1)\mathrm{E}[f^3]/(\mathrm{E}[f^2]\mathrm{E}[f]n)\}/\sqrt{\sigma_m^2\sigma_{m_2}^2}\right]$

Note: $f_i \equiv f(\theta_i)$ and $\mathrm{E}[f^q] \equiv \sum_{i=1}^n f_i^q/n$

Table 2: Comparison of Individual and Aggregated Analyses of Simulated Data. The model is $f_1 = \exp(-\kappa_1 t)$ and $f_2 = (1 - f_1)\exp(-\kappa_2 t)$, with the population distributions of $\kappa_{1,2}$ given by $\ln \kappa_{1,2} \sim N(\mu_{1,2}, \Sigma_{1,2})$, and measurement error models $y_{1,2} \sim N(f_{1,2}, \sigma_{1,2})$, $y_{1,2} \sim N(f_{1,2}, f_{1,2}\sigma_{1,2})$, $\ln y_{1,2} \sim N(\ln f_{1,2}, \sigma_{1,2})$, for error models I, II, and III, respectively. Prior distributions are $\mu_{1,2} \sim N(0, 2)$, $\Sigma^2_{1,2} \sim\sim \text{Inv}\Gamma(1, 0.01)$, and $\sigma^2_{1,2} \sim\sim \text{Inv}\Gamma(1, 0.01)$. Posterior values are $\text{median}^{97.5\%}_{2.5\%}$, based on 5 chains, each of length 11,000, with the last 7000 for inference. Simulated data is cross-sectional, with 8 individuals at each of $N = 6$ time points. Similar results (not shown) are obtained from longitudinal data ($N = 6$, $n = 8$).

Error		Parameter					
Model	Analysis	μ_1	$\exp \Sigma_1$	$\exp \sigma_1$	μ_2	$\exp \Sigma_2$	$\exp \sigma_2$
All	True Value:	0.50	1.28	1.064	-0.50	1.13	1.064
All	Sample Value:	0.49	1.26	1.068	-0.52	1.13	1.064
I	Individual:	$0.49^{0.57}_{0.43}$	$1.25^{1.32}_{1.2}$	$1.072^{1.12}_{1.047}$	$-0.53^{-0.49}_{-0.57}$	$1.1^{1.14}_{1.07}$	$1.074^{1.118}_{1.047}$
	Aggregate:	$0.49^{0.56}_{0.42}$	$1.26^{1.33}_{1.21}$	$1.068^{1.121}_{1.042}$	$-0.53^{-0.49}_{-0.56}$	$1.11^{1.15}_{1.08}$	$1.071^{1.127}_{1.044}$
II	Individual:	$0.49^{0.53}_{0.43}$	$1.25^{1.29}_{1.2}$	$1.071^{1.115}_{1.046}$	$-0.53^{-0.5}_{-0.57}$	$1.11^{1.13}_{1.08}$	$1.071^{1.035}_{1.044}$
	Aggregate:	$0.48^{0.54}_{0.42}$	$1.23^{1.3}_{1.18}$	$1.07^{1.127}_{1.043}$	$-0.53^{-0.49}_{-0.57}$	$1.11^{1.16}_{1.08}$	$1.069^{1.121}_{1.043}$
III:	Individual:	$0.49^{0.57}_{0.43}$	$1.25^{1.32}_{1.2}$	$1.072^{1.12}_{1.047}$	$-0.53^{-0.49}_{-0.57}$	$1.1^{1.14}_{1.07}$	$1.074^{1.118}_{1.047}$
	Aggregate:	$0.48^{0.54}_{0.42}$	$1.23^{1.3}_{1.18}$	$1.07^{1.129}_{1.043}$	$-0.53^{-0.48}_{-0.57}$	$1.11^{1.16}_{1.08}$	$1.068^{1.117}_{1.043}$

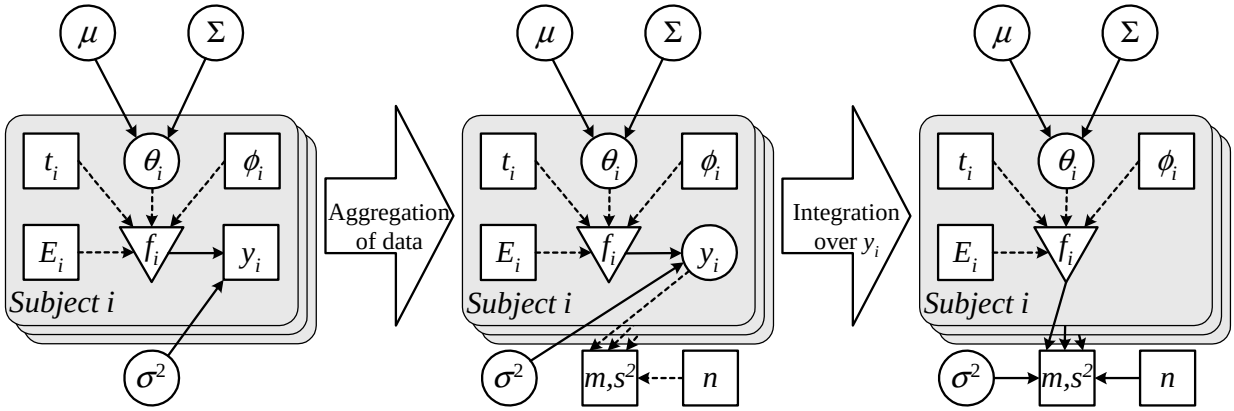

Figure 1: Graphical representation of population statistical model describing dependence relationships between variables. Square nodes are known or measured quantities, circle nodes are unknown or unobserved, solid arrows indicate a stochastic dependence, and dashed arrows indicate a logical (functional) dependence. The inverted triangle f represents the nonlinear pharmacokinetic model prediction. Individuals are labeled by the index i. It should be noted that all of the nodes may have additional dimensions in addition to i (e.g., multiple time points, tissues). The left model represents the situation where individual data y_i are available. Aggregation of data leads to the middle model, where only m, s^2, and n are available (note y_i is still part of the model, but has changed from a square node, denoting a measured quantity, to a circle node, indicating an unobserved quantity). Marginalization over y_i leads to the model on the right, which is used in our analyses. Approximations for the results of this marginalization for various measurement models are described in the text.

Table 3: Comparison of Individual and Aggregated Analyses of 1,3-Butadiene Data. Each pharmacokinetic model parameter has an associated population mean μ and variance Σ^2. Individual analyses used error model III (Bois et al. 1999). Aggregated analyses for error model I used log-transformed measurements; those for error models II and III used untransformed measurements. Prior distributions on μ and Σ^2 are specified below, as are the posterior median$^{97.5\%}_{2.5\%}$ on μ and $\exp\Sigma$ based on two independent chains of 50,000 iterations, thinned by 10, with the first 10,000 iterations discarded. For the prediction AMET, GM_8 and GSD_8 are the geometric mean and geometric standard deviation of 8 individuals, respectively.

Parameter or Prediction	Priors (μ, Σ^2)	Posteriors (μ, $\exp\Sigma$) from Analyses:			
		Individual	Model I	Model II	Model III
sc_V_c: μ	Unif$(0.01, 0.5)$	$0.0637^{0.0772}_{0.0533}$	$0.0654^{0.084}_{0.0517}$	$0.0682^{0.0928}_{0.0517}$	$0.0665^{0.086}_{0.0526}$
sc_V_c: Σ^2	Inv$\Gamma(1, 0.039)$	$1.16^{1.33}_{1.09}$	$1.21^{1.62}_{1.1}$	$1.28^{1.97}_{1.11}$	$1.22^{1.65}_{1.1}$
K_{cp}: μ	Unif$(0.5, 5)$	$1.09^{1.32}_{0.919}$	$1.12^{1.39}_{0.923}$	$1.13^{1.39}_{0.934}$	$1.13^{1.4}_{0.925}$
K_{cp}: Σ^2	Inv$\Gamma(1, 0.039)$	$1.16^{1.34}_{1.09}$	$1.18^{1.4}_{1.1}$	$1.19^{1.41}_{1.1}$	$1.18^{1.4}_{1.1}$
$P_{pc}V_p$: μ	Unif$(10, 100)$	$27.9^{33.2}_{23.9}$	$28.8^{35.3}_{23.8}$	$29.2^{35.1}_{24.2}$	$29.2^{35.7}_{24.1}$
$P_{pc}V_p$: Σ^2	Inv$\Gamma(1, 0.039)$	$1.14^{1.28}_{1.08}$	$1.17^{1.38}_{1.09}$	$1.18^{1.39}_{1.1}$	$1.18^{1.38}_{1.1}$
K_{met}: μ	Unif$(0.05, 0.5)$	$0.224^{0.346}_{0.144}$	$0.23^{0.37}_{0.136}$	$0.228^{0.37}_{0.137}$	$0.23^{0.368}_{0.136}$
K_{met}: Σ^2	Inv$\Gamma(1, 0.693)$	$1.66^{2.52}_{1.38}$	$1.74^{2.93}_{1.4}$	$1.75^{2.89}_{1.4}$	$1.72^{2.88}_{1.4}$
P_{ca}: μ	Unif$(0.1, 5)$	$1.3^{1.57}_{1.09}$	$1.29^{1.52}_{1.09}$	$1.3^{1.53}_{1.09}$	$1.28^{1.53}_{1.08}$
P_{ca}: Σ^2	Inv$\Gamma(1, 0.039)$	$1.21^{1.44}_{1.12}$	$1.18^{1.38}_{1.1}$	$1.18^{1.39}_{1.1}$	$1.18^{1.39}_{1.1}$
sc_K_{in}: μ	Unif$(0.1, 1)$	$0.372^{0.422}_{0.329}$	$0.372^{0.444}_{0.315}$	$0.373^{0.444}_{0.319}$	$0.373^{0.443}_{0.319}$
sc_K_{in}: Σ^2	Inv$\Gamma(1, 0.039)$	$1.17^{1.33}_{1.11}$	$1.24^{1.5}_{1.14}$	$1.23^{1.49}_{1.13}$	$1.23^{1.48}_{1.13}$
BDW: μ	Unif$(48, 71)$	$61.7^{68.6}_{54.3}$	$61.7^{68.5}_{54.8}$	$61.7^{68.5}_{54.7}$	$61.6^{68.5}_{54.5}$
BDW: Σ^2	Inv$\Gamma(1, 0.039)$	$1.19^{1.36}_{1.12}$	$1.18^{1.34}_{1.12}$	$1.18^{1.34}_{1.12}$	$1.18^{1.34}_{1.11}$
GSD_{ex}	LogUnif$(1.01, 1.30)$	$1.08^{1.09}_{1.07}$	$1.11^{1.13}_{1.09}$	$1.11^{1.13}_{1.09}$	$1.11^{1.13}_{1.09}$
AMET: GM_8	—	$0.023\ ^{0.028}_{0.017}$	$0.023\ ^{0.028}_{0.016}$	$0.024\ ^{0.031}_{0.017}$	$0.024\ ^{0.031}_{0.017}$
AMET: GSD_8	—	$1.37\ ^{1.76}_{1.17}$	$1.40\ ^{1.92}_{1.18}$	$1.41\ ^{2.00}_{1.18}$	$1.39\ ^{1.91}_{1.17}$

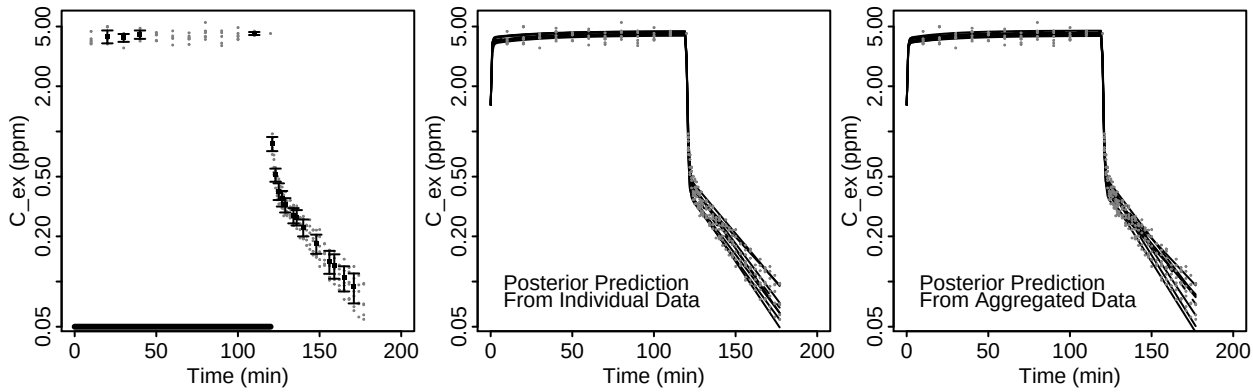

Figure 2: Exhaled breath concentration data used in 1,3-butadiene analysis. Individual data are shown in the points. Aggregated data are shown in the left panel (square, with one-standard deviation error bars), but only at data points where all 8 subjects had measurements. The solid bar indicates the time during which exposure (at 5 ppm) occurred. The middle and right panels show comparison of the original individual data with posterior simulations (without "measurement" error) for the samples with the highest posterior likelihood (solid lines) for analysis of individual data (center panel) and aggregated data using Model III (right panel).

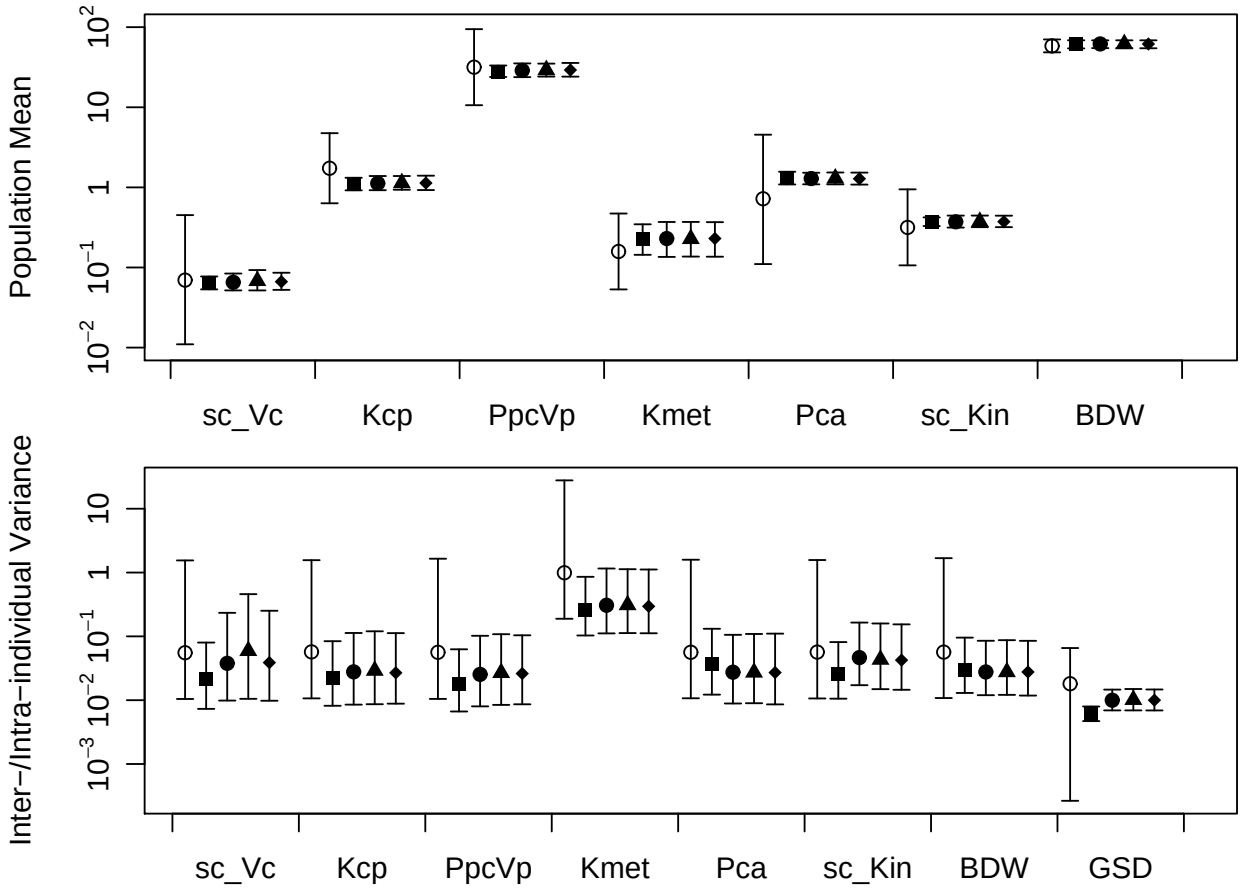

Figure 3: Median (symbols) and [2.5%,97.5%] confidence interval (error bars) for prior (open circle) and posterior (solid symbols) distributions for the population means (upper panel) and inter-/intra-individual variances (lower panel). Solid symbols represent full population analysis of individual data (square), and aggregated analyses for Models I-III (solid circle, triangle, and diamond, respectively).