How can statistical theory help with statistical practice? Example of a Bayesian analysis in toxicokinetics

Andrew Gelman and Frédéric Y. Bois
Columbia University and Lawrence Berkeley National Laboratory

Abstract: Statistical theory affects statistical practice in two ways: by suggesting methods for us to try, and by giving us confidence in some of our conclusions. We discuss both of these issues in the context of a Bayesian analysis in toxicokinetics. Key areas of statistical theory that have been helpful to us so far are Bayesian inference, Markov chain simulation, and hierarchical modeling. Other areas, for which more theory is needed, are parameter transformations for model simplification, modeling of correlations, and posterior predictive model checking and sensitivity analysis.

Keywords: Bayesian data analysis, model checking, posterior predictive checks, sensitivity analysis, toxicokinetics, transformations

1 Introduction

The theme of this volume is "good statistical practice in scientific data analysis." Of particular interest to statisticians is the relation of statistical theory to good statistical practice. We explore this issue in the context of a particular problem in toxicokinetics (the study of the flow and metabolism of toxins in the body) studied by a toxicologist (Bois) with the help of a statistician (Gelman). We begin in Section 1 with some background (see BOIS ET AL., 1996, and GELMAN, BOIS and JIANG, 1996, for details of the problem and our analysis) and a discussion of why we felt the need to use elaborate statistical methods. In Section 2, we discuss the key places where statistical theory was crucial in allowing us to follow the principles of good statistical practice. In Section 3, we discuss areas of "good statistical practice" (in this example) where we feel that there is room for more statistical theory to be developed. We conclude in Section 4 with a short discussion.

Very briefly, we identify statistical theory with the following topics: probability and distribution theory; algorithms for statistical computation; formal procedures for inference, testing, and prediction; and evaluation of the probability distributions of statistical procedures. As "good statistical practice," we include: data collection; inclusion of all relevant information in the statistical analysis; relating statistical models to scientific quantities of interest; recognizing variability in populations; recognizing uncertainty in inference, testing, and prediction; checking the adequacy of statistical procedures; and using graphical displays to understand data and inferences.

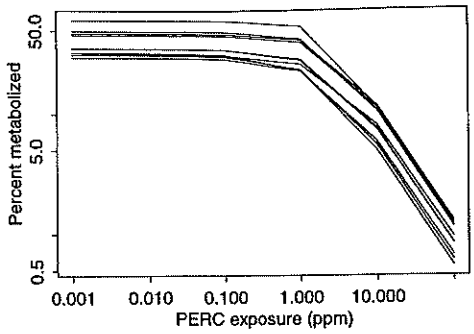

Figure 1: Estimated fraction of PERC metabolized, as a function of steady-state concentration in inhaled air, for 10 hypothetical individuals randomly selected from the estimated population of young adult white males.

1.1 Background

Tetrachloroethylene (also called perchloroethylene and abbreviated PERC) is one of many industrial products that cause cancer in animals and is believed to do so in humans as well. PERC is breathed in, and the general understanding is that it is metabolized in the liver and that its metabolites are carcinogenic. Thus, a relevant "dose" to study when calibrating the effects of PERC is that metabolized in the liver. Not all the PERC that a person breathes will be metabolized and, because of saturation mechanisms, the fraction metabolized depends on the rate of intake. We were focusing on estimating the fraction of PERC metabolized as a function of its concentration in the breathed air, and how this function varied across the population. This particular work was funded by health and environmental regulatory agencies that were interested in the exposure to PERC of the general population. To give an idea of what we are talking about, we skip ahead to show some output from our analysis. Figure 1 displays the estimated fraction metabolized as a function of concentration in air, for 10 randomly selected draws from the estimated population of young adult white males (the group on which we had data).

Figure 1 has two key features: extrapolation to low exposures, and population variability. It was not possible to estimate either of these with reasonable confidence using simple procedures such as direct measurement of metabolites (difficult even at high doses and not feasible at low doses) or extrapolation from animal results. So it was decided to fit a *toxicokinetic* model; that is, a mathematical model of the flow of the toxin through the bloodstream and body organs, and of its metabolism in the liver. "Figure 1" could then be estimated in two steps, which we in fact followed:

- Estimate the parameters of the toxicokinetic model from indirect measurements (concentrations of PERC in the blood, inhaled air, and exhaled air, over time) in experimental data. Use data on several individuals to estimate

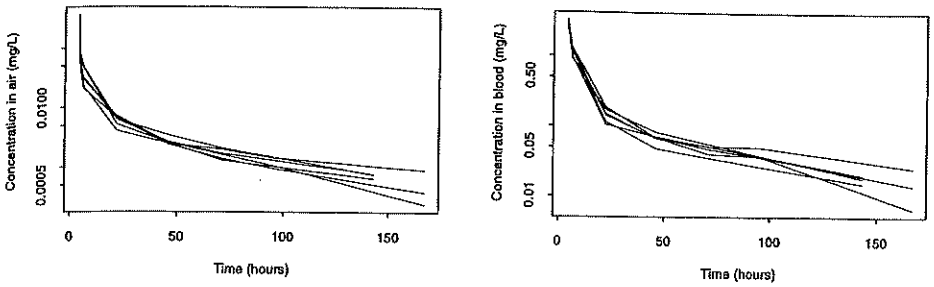

Figure 2: Concentration of PERC in exhaled air and in blood, over time, for one of two replications in each of six experimental subjects.

the population variability of the parameters.

- Compute the curves in Figure 1 given the toxicokinetic model and the estimated parameters.

In the standard model, which we used, the toxin enters and leaves through the breath, is distributed by blood flow to four "compartments"—well-perfused tissues, poorly-perfused tissues, fat, and the liver—and is metabolized in the liver. About 15 toxicokinetic parameters govern the model behavior: the equilibrium concentration ratios between air and blood, or blood and bodily tissues; the volumes of the compartments; the blood flows between compartments, and the rate and capacity of the metabolism in the liver. The model specifies, through differential equations, the time evolution of the concentration in the blood, lungs, and the four compartments and the quantity of metabolites formed, as a function of the input which is the concentration in the inhaled air. We can use the model's predictions about concentration in exhaled air and blood to estimate its parameters.

A sample of the experimental data we used is shown in Figure 2. Each of several volunteers was exposed to PERC at a high level for four hours (believed long enough for the PERC concentrations in most of their bodily organs to come to equilibrium) and then measured over a period of a week. In addition, the data on each subject were replicated, and measurements such as body weight and fat content were recorded.

A key issue in the modeling is: how many compartments should be included? The more compartments, the more realistic the model but the harder it is to fit. At one extreme is the one or two-compartment model which, given data from several subjects, can be estimated reasonably precisely without the need for prior information (see WAKEFIELD, 1996). However, these simpler models provide a poor fit to our data and, more importantly, do not have the complexity to accurately fit varying exposure conditions. (For example, if the concentration of PERC is high in

your workplace and low at home, then your fatty tissues can store excess PERC at the workplace, then release it into your bloodstream at home.) For this problem, four compartments seemed necessary: the liver (where metabolism occurs), fatty tissues (where the toxin can be stored), and other tissues (divided into well and poorly-perfused to model immediate and longer-term reactions to changes in input conditions). This model has a long history in solvent toxicology modeling and has been showed to reproduce most features of such data[1].

1.2 Why complicated statistical methods were needed

The preceding section described why an elaborate model was used. Before going on, let us discuss briefly why fancy statistics (rather than just something simple like least squares) were needed to estimate it. First, multicompartment models are hard to estimate from indirect data. That is, given a set of measurements such as shown in Figure 2, there are many combinations of parameter values, some of which make no scientific sense, that will fit the data about equally well. In numerical analysis terms, this is an "ill-posed problem"—actually, close to the problem of estimating the parameters in a mixture of declining exponential distributions. So we need to find parameter values that fit both the data and our prior understanding. Second, we are interested in estimating population variability, which means, in particular, distinguishing this from estimation uncertainty. At this point, formal Bayesian inference seemed to be the easiest way to attack the problem.

Following GELMAN ET AL. (1995), we divide the statistical analysis into three steps: modeling, inference, and model checking. Throughout, there were three key issues we had to deal with, which roughly corresponded to the three steps of analysis:

1. A physiological pharmacokinetic model, which has many parameters and involve differential equations, is inherently complicated and much harder to understand than statistical standbys such as regressions and mixture models.

2. As noted above, the model's parameters are hard to estimate accurately from indirect data.

3. Finally, the model is inherently oversimplified, most notably in that it assumes that each compartment is in internal equilibrium at all times, and that it assumes the parameters are constant over time.

2 Help from statistical theory so far

Without modern statistical theory, we would not have been able to have come close to the amount of scientific progress that we achieved in this problem. Our most important theoretical tools were Bayesian inference, Markov chain Monte Carlo

[1]However, for other compounds (e.g., butadiene), a two-compartment model can fit well.

sampling, and hierarchical modeling. To see why, let us describe the state of the analysis before the statistical theory had arrived.

2.1 Toxicokinetics without statistical theory

What is done today by toxicologists who do not team up with statisticians? The toxicokinetic models make scientific sense, but they are often unnecessarily complicated. To cope with their complexity, highly uncertain parameters are assumed known, so that parameters can be estimated. With no good plan for estimation, the designs for data collection are not focused on the outcomes of interest. Model parameterization can be based on visual fits to data, and model checking reduces to statements like "the fit looks nice" when no formal fitting has been done. No measures of uncertainty are given for parameters or predictions except purely from independent prior distributions on parameters.

To be clear, we should distance this from pharmacology, where the drugs more often have simple kinetics (for reasons linked to their molecular structure), the drugs usually act directly, and measuring or predicting the drug level in blood or plasma does a good job at getting at an effective dose. In addition, pharmacologists have a strong motivation to accurately measure drug effects, and relatively clear evidence: they give high levels of compounds to already sick people, and these chemicals are pretty active. Money and lives are directly at stake, and the researchers must present clear proofs of safety and efficacy, and statisticians have been vetting these analyses for quite a while. So, in pharmacology they can use (often) simple models[2] and they have been trained to deal with them rigorously.

In contrast, toxicologists often have study complicated pathways, have never thought about issues of variability (rarely having to deal with real people in bad shape and able to sue them, usually using healthy animals which look all the same, so the toxicologists forgot about inter-individual variability). Usually the effort goes to trying to prove that the chemical is ineffective (that is the industry's interest, and it has a big bearing on experimental design) or understand qualitatively how it works, rarely thinking in quantitative terms. This is changing, but slowly. Impetus for change comes in part from the interest in risk assessment for the range of individuals in the population.

2.2 Areas of theory that have directly improved the practice of data analysis in toxicokinetics

So where did statistical theory help in our analysis? First, as described in Section 1.2, it was possible to estimate the four-compartment model only with the use of

[2]More complicated models do have their place in pharmacokinetics, however. For example, LUDDEN, GILLESPIE and BACHMAN (1995) discuss the possibility of physiologically-based pharmacokinetic models to help with complex drug problems (some drugs do act via several metabolites, have complicated kinetics, and so forth).

prior information (from the biomedical literature) on the physiological parameters. Bayesian inference provided a good framework for incorporating this prior information, when augmented with some special techniques of our own (for example, bounding the prior distributions of parameters at ± 3 standard deviations, and then checking that the posterior distributions were not concentrated at the boundary of parameter space—that is, that the model and data fit the prior distributions). Other, less theoretically-developed ways of including the information—for example, setting some parameters to fixed values and estimating the others—did not allow the model to fit the data accurately.

Second, it was only possible to find parameter values that fit both the data and the prior knowledge (in Bayesian terms, to explore the posterior distribution) by using Markov chain Monte Carlo simulation, an area of active theoretical research in statistics (see GILKS, RICHARDSON and SPIEGELHALTER, 1996). The previously-tried strategy of sampling from the prior distribution simply did not discover good fits to the data. In addition, new research on efficient simulation algorithms (GELMAN, ROBERTS and GILKS, 1996), motivated by this project, was helpful in reducing the computation time required for this and subsequent projects.

Third, hierarchical modeling was crucial in distinguishing population variability from parameter uncertainty, as discussed in Section 1.2. Statistical theory for hierarchical models, developed only in the past few decades, allows us to simultaneously estimate individual parameters along with their population distribution.

Fourth, many individual technical developments smoothed our path and allowed us to fit a model that corresponded as closely as possible to our scientific understanding. These included transformations of the parameter space to allow a natural modeling for constrained parameters (which required theory in the form of a new computational method; see GELMAN, 1995), and simulation-based posterior predictive model checking (discussed more in Section 3.2).

3 Help from future statistical theory?

In this section we discuss the areas in our data analysis where we believe there is the potential for new statistical theory to be helpful.

3.1 Constructing the model

To be able to construct informative prior distributions, it was necessary to parameterize the problem in a meaningful way so that, for example, the four compartments referred to recognizable bodily organs (this would not be reasonable in a one or two-compartment model). As mentioned at the end of Section 2.2, the parameters were transformed so that their values would be more similar across the population. For example, instead of the volumes each of the four components, we worked with their volumes relative to total body volume. In general, the practical advice is always to

transform data and parameters so that they are more understandable and easier to model accurately. Existing statistical theory, which focuses on transformations to additivity and normality, has not yet caught up with good practice. We believe that more theoretical work needs to be done on the transformations to parameterizations where a given set of prior knowledge will be maximally informative.

In addition to reducing the population variance (thus allowing more information to be shared among individuals in the Bayesian analysis), the transformations made it more plausible for us to assume independent prior distributions on our 15 individual-level parameters. This relieved a potential burden of modeling and computing with a big covariance matrix, but left us awaiting more statistical theory on the effects of ignoring parameter correlations in a multivariate hierarchical model. Our understanding goes something like this: suppose parameters α_j and β_j are positively correlated across the population of individuals j, but we ignore this, assuming a zero correlation in our model. Then we will certainly be mistaken in our predictions about additional individuals in the population—but for the individuals on which we have data, we are merely reducing the efficiency of inference (compared to the analysis that accounts for the true correlation) by not allowing information about α_j to be useful in estimating β_j and vice-versa. But we are not aware of any theoretical results that directly address this question. One problem is that the notion of "bias" is not easily translatable to hierarchical models, where classical "unbiased" estimates are generally undesirable even if possible (see, e.g., pp. 108–109 of GELMAN ET AL., 1995).

Now here's an issue we've swept under the rug: our model, as described so far, is deterministic! Given the parameters and input conditions, it predicts the outputs exactly. This of course will not fit any real data. We then simply assumed the errors were independent and normally distributed (on the log scale). The relevant practical advice is to put "error terms" at all levels of your model, to reflect that it won't fit reality and so that it will not overfit your data. For our example, the problem was certainly model misfit and not measurement error. Much statistical theory has been developed on the topic of model errors, from simple (distributions of error terms) to complex (stochastic differential equations that put the error inside of the model)—although in this example, we only felt the need to do the simplest thing. If short-term kinetics were of interest, it might have been necessary to model the errors more carefully. For our purposes, we were concerned not with the distribution of the errors but with their magnitude. Figure 3 shows a scatterplot of the relative prediction errors of all our observed data (that is, observed data divided by their predictions from the model) along with their estimated values from the deterministic part of the model. (Since the analysis was Bayesian, we have many simulation draws of the parameter vector, each of which yields slightly different predicted data. Figure 3, for simplicity, shows the predictions from just one of these simulation draws, selected at random.) The magnitude of the errors (most no larger than a factor of 1.5) was reasonably low compared to other fits of this kind of data, so we were satisfied.

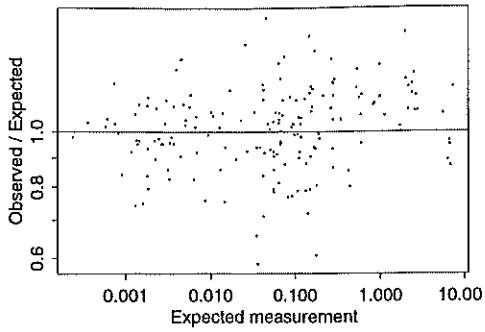

Figure 3: Observed PERC concentrations (for all individuals in the study) divided by expected concentrations, plotted vs. expected concentrations. Note the different scales on the (logarithmic) x and y-axes: observations vary by a factor of 10000, but the relative errors are mostly between 0.8 and 1.25. Because the expected concentrations are computed based on a random draw of the parameters from their posterior distribution, the figure shows the actual misfit estimated by the model, without the need to adjust for fitting.

3.2 Checking and understanding the model

A key step of data analysis, historically often ignored in the theory of Bayesian statistics, is to check the fit of the model to data and other substantive knowledge. In the toxicokinetics example, this began with the realization that fitting the model freely to data, with no constraints on parameters, led to nonsense such as an 8-kilogram liver, because the model was not well constrained by the indirect data used to fit it. The first step in checking the Bayesian model was thus to check that the posterior inferences were reasonable—that they agreed with the prior distribution. They did so. Although the Bayesian approach is intended to compromise between the data and the prior distribution, it can happen that the model is such a poor fit that the result close to *neither* the prior distribution nor the data. Thus, we also checked that the model predictions were close to the data. The fit of the data to their predictions, from a single random draw from the posterior distribution, as shown in Figure 3, is acceptable.

The theory that has been developed for this "good statistical practice" is based on the idea of hypothesis testing: comparing observed data to what would be expected under the prior (BOX, 1980) or posterior (RUBIN, 1984) distributions. One theoretical issue we still do not understand is when is it appropriate to compare to prior and when to posterior distributions. In many cases, a posterior check seems more appropriate (see GELMAN, MENG and STERN, 1996), but in our example it was important to check the fit to the prior distribution as well.

In addition, we checked the sensitivity of our inferences to the prior distributions using scatterplots showing the relation between the quantity of interest—the percent

of PERC metabolized at low and high doses—and various key parameters in the model. Each dot in such a plot represents a different draw from the posterior distribution. The plots show how the posterior distribution for the quantity of interest would be affected by changes in the marginal posterior distributions of the various parameters. This can be considered a "static" sensitivity analysis, in contrast to the usually-recommended "dynamic" version that requires the model to be re-fit with different prior distributions. The advantage of the static analysis is that it does not require re-fitting the model. Theoretical work needs to be done to understand the effectiveness of this approach.

Finally, we checked the model assumption of zero population correlations among the 15 model parameters (recall from Section 3.1). This was done in a fairly elaborate way, making use of the replication in the data. For each of the $15 \times 14/2$ pairs of parameters and for each posterior simulation draw, we computed the correlation of the parameters across the six experimental subjects. Then, for each pair of parameters, we examined the distribution of the simulated correlations to see if they were substantially and statistically significantly far from zero. We in fact found some high correlations—three parameters with correlations of about 0.8—which we were able to remove by a further, relatively minor adjustment to the model. This check is good practice, we believe, but the relevant theory is still somewhat elusive. It seems to be related to predictive checking (since, under the model, the correlations would be expected to be zero, on average).

4 Discussion

In general, statistical theory can and should lead to good statistical practice. In fact, much of our analysis—setting up a hierarchical structure and assigning them prior distributions, performing Bayesian inference, and some of the model checking—was performed following statistical theories. Other crucial steps in the analysis—using transformations to allow the prior distribution to be more informative, understanding the consequences of assuming zero correlations, and some aspects of model checking—clearly must have some theoretical justification, but more work needs to be done. In addition, the applied work itself motivated new theoretical ideas in Markov chain simulation, sensitivity analysis, parameterization of constrained prior distribution. In applied work, statistical theory is often needed to transform principles of "good data analysis" into practically-useful methods.

Acknowledgments

We thank the U.S. National Science Foundation for grant DMS-9404305 and Young Investigator Award DMS-9457824, and the Research Council of Katholieke Universiteit Leuven for fellowship F/96/9.

References

Bois, F.Y., Gelman, A., Jiang, J., Maszle, D., and Alexeef, G. (1996). Population toxicokinetics of tetrachloroethylene. *Archives of Toxicology* **70**, 347–355.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A* **143**, 383–430.

Cox, D.R., and Snell, E.J. (1981). *Applied Statistics: Principles and Examples.* London: Chapman and Hall.

Gelman, A. (1995). Method of moments using Monte Carlo simulation. *Journal of Computational and Graphical Statistics* **3**, 36–54.

Gelman, A., Bois, F.Y., and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association* **91**, 1400–1412.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Gelman, A., Meng, X.L., and Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6**, 733–807.

Gelman, A., Roberts, G.O., and Gilks, W.R. (1996). Efficient Metropolis jumping rules. In: *Bayesian Statistics 5*, ed. J. Bernardo et al., 599–607. Oxford University Press.

Gilks, W.R., Richardson, S., and Spiegelhalter, D., eds. (1996). *Practical Markov Chain Monte Carlo.* London: Chapman and Hall.

Ludden, T.M., Gillespie, W.R., and Bachman, W.J. (1995). Physiologically based pharmacokinetic modeling as a tool for drug development—commentary. *Journal of Pharmacokinetics and Biopharmaceutics* **23**, 231–235.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**, 1151–1172.

Wakefield, J.C. (1996). The Bayesian analysis of population pharmacokinetic models. *Journal of the American Statistical Association* **91**, 62–75.

Wakefield, J., and Bennett, J. (1996). The Bayesian modeling of covariates for population pharmacokinetic models. *Journal of the American Statistical Association* **91**, 917–927.