A Primer On Hierarchical Models

Dalene K. Stangl

Hierarchical or multi-level models are useful in many applications. Meta-analysis, nested data structures, longitudinal data, heterogeneous populations and model selection are the primary areas of application. This paper provides a conceptual overview of Bayesian hierarchical models and discusses making inferences from a Bayesian perspective. References to papers that make comparisons between Bayesian and other inferential methods will be provided, as will references to software for implementing these methods. While the utility of hierarchical models is recognized in many fields including agriculture, biology, education, social and natural sciences, and policy, the examples presented here are mostly from health-related research.

Why Use Hierarchical Models?

You use batteries for your CD player, walkman, calculator and other small appliances. Suppose you want to compare brands of batteries to determine which last longest. How would you design an experiment to help answer this question? One possibility is to enlist the help of 15 friends. You give 15 of your friends identical CD players and money to buy 10 Everready and 10 Duracell batteries. Each friend records the play-time of their batteries. How do you analyze the data to determine which brand lasts longer? From a statistical modeling perspective, this raises interesting methodological questions.

What is the best way to make inferences from the 300 observed batteries, and how do you use these inferences to make purchasing decisions for your next batteries?

Although your friends have 'identical' CD players, there may be unobserved differences in the CD players and in the ways your friends use their players (e.g., volume) and record their play-times that may affect battery life. How do we take this into account? Is it possible that conditions of use are so different between friends that Everready works better for one friend and Duracell works better for another? If your friends buy their batteries in packages of 2 or use their batteries in pairs, should these pairs be treated as independent observations?

This type of analysis problem calls for the use of hierarchical or multi-level models. They are called hierarchical models because there is a sequence of nested probability models. For pedagogical purposes, assume we are interested in only one of the two brands of batteries and that friends use the batteries solo. At stage I, there is a model for battery life within each of your 15 friends. At stage II, there is a model that takes into account the variability in average battery life across your 15 friends. At stage III we incorporate any information we might have (from sources other than the data at hand) about the average battery life and the variability we expect to see between friends. One possibility would be to model the log of battery lifetimes for friend i as independent normally distributed data points, t_{ij}, with mean μ_i and known error variance σ^2. For the convenience of conjugacy, assume that the average battery lifetime for each friend, μ_i, is drawn from a normal distribution with mean μ_0, and variance τ. This model says that the μ_i are conditionally independent given (μ_0, τ). Finally we assign a prior distribution to μ_0 and τ. In summary:

$$\begin{array}{lll} \text{Stage I} : & t_{ij} \sim N(\mu_i,\sigma^2) & i=1,K\ ,15;\ j=1,K\ ,10 \\ \text{Stage II} : & \mu_i \sim N(\mu_0,\tau) & i=1,K\ ,15 \\ \text{Stage III}: & p(\mu_0,\tau) & \end{array}$$

While this is a hypothetical example, the world is full of examples that have similar structure. In medicine we carry out multi-center clinical trials in which 'identical' treatment protocols are imple-

Dalene Stangl is Associate Director of the Institute of Statistics and Decision Sciences and Professor of the Practice of Statistics at Duke University. She has taught statistics at Duke University since 1992. Her professional interests are Bayesian methods and statistics education reform. She will serve as Executive Editor of Chance *magazine during 2002–2004. (www.stat.duke.edu/~dalene).*

mented at multiple sites. In education we examine teaching practices by observing students clustered within classrooms. In policy areas we examine state/federal government interventions by observing citizens clustered within counties/states. Next we will examine a health-related example that shares this same structure. It is the data from a meta-analysis assessing the effect of an antidepressant drug called S-adenosylomethionene.

A Meta-Analysis Example

In many research areas, including but not limited to medicine, education, and policy, clustered data arises in meta-analysis. The goal of meta-analysis is to combine information from a number of studies examining the same phenomena to make inferences and predictions. In such an analysis the data is clustered within studies. Here we will look at a meta-analysis of nine clinical trials.

An analysis was conducted to examine the effects of the antidepressant drug S-adenosylomethionene (SAMe), (DuMouchel 1989; Berry and Stangl, 1996). Nine study sites participated in the trial. Each site had characteristics setting it apart from the other sites that affected the distribution of outcomes at that site. The outcome of interest was the rate of successes observed with SAMe. The data are presented in the table below.

Table 1. Data from 9 sites in the study of the antidepressant drug S-Adenosylmethionine.

Site	s_i	n_i	s_i/n_i
1	20	20	1.00
2	4	10	.40
3	11	16	.69
4	10	19	.53
5	5	14	.36
6	36	46	.78
7	9	10	.90
8	7	9	.78
9	4	6	.67
Totals	106	150	.71

An analysis that simply pooled the data across the 9 sites would give a maximum likelihood estimate of 0.71, and more than 95% of the area under the likelihood is between 0.6 and 0.8. Compare this to the maximum likelihood estimates for each site given in the fourth column of Table 1. Five of the nine sites had success proportions outside the interval (0.6, 0.8). While sampling variability accounts for some differences, the variability seen here is greater than what would be expected from sampling alone. This suggests that the success probability of SAMe is not equal across the 9 sites. Should we simply present 9 estimates or 9 confidence intervals? This assumes nothing is learned about the effect of the drug at a particular site from the observations at the other 8 sites. So, if naïve pooling is not satisfactory, and separate estimates are not satisfactory, how should we model the effects across the 9 sites to account for the extra variability and come up with answers that can help us make predictions about future observation at the nine observed sites and at sites not included in the study? One possibility is a Bayesian hierarchical model. Before demonstrating such a model, let's review the basics of a Bayesian model.

Bayesian Models: Let's suppose that the data in Table 1 came from a single study with 150 patients. Our goal is to estimate the success rate, p, of the treatment. From a Bayesian perspective p is a random variable with a probability distribution. The distribution we assign to the success rate before seeing the data is called a prior distribution. Because the success rate must fall between 0 and 1, one possibility for the prior distribution would be a beta. The beta distribution has the following form:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}.$$

This distribution has support on the interval [0,1], and can take a variety of shapes. The mean of the beta distribution is

$$\frac{\alpha}{\alpha+\beta},$$

And the variance is

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Figure 1 shows three possibilities for the choices of α and β. We see that for Beta(α = 2, β = 2), there is visible probability mass on the interval (.03,.97), for Beta(α = 5, β = 5) on (0.1, .9), and for Beta(α = 10, β = 10) on (.2,.8). When $\alpha + \beta$ is large, the variance of the beta distribution will be small, so the distribution will be highly concentrated signifying high certainty about the location of p before we observe the current data. However, when $\alpha + \beta$ is small the uncertainty in p is considerable. At the extreme, α = 1 and β = 1, we have the uniform distribution. Apriori, we do not favor any value of p in the interval [0,1] over any other.

The densities in Figure 1 show examples where $\alpha = \beta$, so the beta distribution is symmetric around 0.5. Figure 2 shows a few alternative beta distributions (mean = 0.25) to demonstrate the diversity of belief that a beta prior can represent, and hence its flexibility as a prior distribution.

Let's suppose that we have chosen values of α and β for the beta prior distribution. The data are

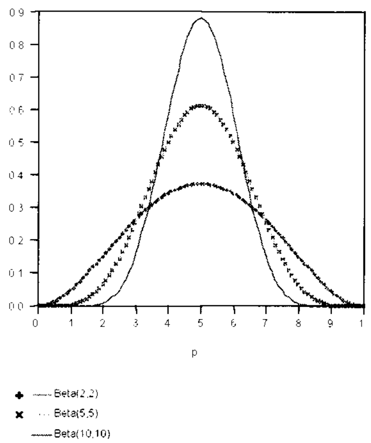

Figure 1. Beta distributions symmetric around $p = .5$.

generated from a binomial distribution with sample size n and success probability p.

$$s \sim \text{Binomial}(n,p).$$

How does one arrive at the posterior distribution of p and the predictive distribution for future observations? The answer is Bayes theorem. Bayes theorem relates the posterior and prior distributions for p through the following formula:

$$f(p \mid data) = \frac{f(data \mid p) f(p \mid \alpha,\beta)}{f(data)}.$$

The denominator of Bayes theorem is the marginal distribution of the data. It is often referred to as the normalizing constant. Factoring out this constant, the posterior distribution, $f(p|data)$, is proportional to the likelihood, $f(data|p)$, times the prior distribution, $f(p|\alpha,\beta)$.

Given binomial data and a beta prior distribution, we have the following posterior distribution for p:

$$f(p \mid data) \propto p^s (1-p)^{n-s} p^{\alpha-1} (1-p)^{\beta-1}$$
$$= p^{s+\alpha-1} (1-p)^{n-s+\beta-1}.$$

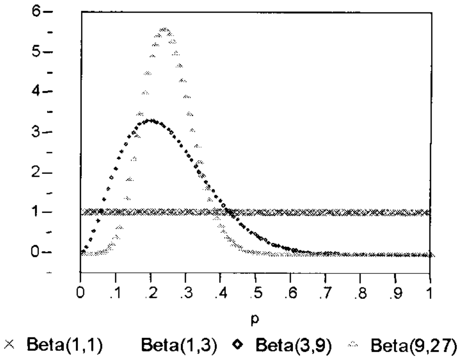

Figure 2. Alternative Beta distributions.

This is the kernel of a Beta distribution with updated parameters $s + \alpha$ and $n - s + \beta$. So, the posterior distribution of p is a Beta($s + \alpha$ and $n - s + \beta$). From this equation we can easily see how conjugacy allows us to interpret $\alpha + \beta$ as the prior sample size.

$$\frac{\alpha+s}{\alpha+\beta+n} = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right) + \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{s}{n}\right)$$

posterior mean = prior weight x prior mean
+ sample weight x sample mean

The posterior mean is a weighted average of the sample mean, s/n, and the prior mean $\alpha/(\alpha + \beta)$. The weights are the ratios of prior and observed sample sizes to the total. From this representation it is easy to see that if $\alpha + \beta$ is small relative to n then the prior will have little impact on the posterior mean.

In the SAMe example, suppose that the prior distribution for p was a Beta(3,2). Then the posterior distribution would be Beta(109, 46). The prior and posterior distributions are shown in Figure 3. Using this model, the posterior mean for p is about 0.70, and the posterior standard deviation is about 0.04. A 95% highest posterior density (hpd) interval is approximately (0.62, 0.78). Given the relatively diffuse prior distribution this interval is about the same as a 95% confidence interval.

This is an unsatisfactory result, because 5 of the 9 success probabilities fall outside this interval. This suggests that an important variance component is left out of the model, the between study variability. It suggests the need for a hierarchical model to incorporate the variability across studies.

Before adding this additional level of variability, it should also be noted that we do not want to fix values of α and β, because we do not know with certainty the value of these parameters. Instead we want to assign a distribution that coheres with our beliefs about plausible values of these parameters.

Hierarchical Model: A multi-level, hierarchical or random-effects formulation avoids the homogeneity assumption by modeling a random effect, p_i for study i. Each p_i is assumed to be selected from a distribution of study effects. Here we will use a Beta(α,β) distribution for the study effects. Response at study i is

$$s_i \sim \text{Binomial}(n_i, p_i),$$

and the individual-study effects are exchangeable. Conditional on α and β, the p_i are independent draws from a beta distribution.

$$p_i \sim \text{Beta}(\alpha,\beta).$$

The likelihood function of the p_i is:

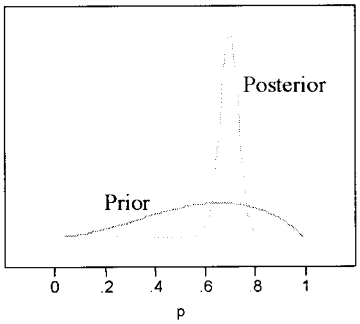

Figure 3. Prior and posterior distributions for *p* assuming patients are exchangeable across all 9 sites.

$$\prod_{i=1}^{I} p_i^{s_i}(1-p_i)^{n_i-s_i}$$

At the last level of the hierarchy, a prior distribution is placed on α and β. We will leave it unspecified for now, and simply denote it by

$$\pi(\alpha,\beta).$$

The joint posterior distribution of all the parameters is:

$$\begin{aligned} f(p,\alpha,\beta \mid s) &\propto f(s \mid p,\alpha,\beta) f(p \mid \alpha,\beta)\pi(\alpha,\beta) \\ &\propto \prod_{i=1}^{I} p_i^{s_i}(1-p_i)^{n_i-s_i} \\ &\times \prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1}(1-p_i)^{\beta-1} \\ &\times \pi(\alpha,\beta). \end{aligned}$$

Given α and β, each of the p_i has an independent posterior beta distribution. Their joint density is

$$f(p \mid \alpha,\beta,s) \propto \prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta+n_i)}{\Gamma(\alpha+s_i)\Gamma(\beta+n_i-s_i)} p_i^{\alpha+s_i-1}(1-p_i)^{\beta+n_i-s_i-1}.$$

The marginal posterior of (α,β) is

$$f(\alpha,\beta \mid s) \propto \pi(\alpha,\beta)\prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+s_i)\Gamma(\beta+n_i-s_i)\Gamma(\alpha+\beta+n_i)}{\Gamma(\alpha+\beta+n_i)}.$$

The posterior distribution of the study effects is the average Beta density with respect to the posterior distribution of (α, β). This is also the predictive distribution for an unobserved study site. Before calculating this distribution, we now must choose a prior distribution for α and β. Berry and Stangl (1996) present two possibilities. They consider independent geometric distributions for α and β, as well as independent uniform distributions on the integers between 1 and 10, inclusive. For a similar model, Gelman et al (1995) demonstrate choosing a parameterization and setting up a diffuse prior distribution continuous in α and β. For simplicity, results will be shown using the independent uniform distributions on the integers between 1 and 10, inclusive.

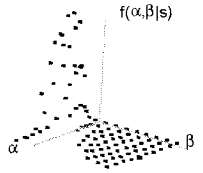

Figure 4. Posterior distribution for α and β assuming uniform prior on integer grid 1,...,10. Posterior mode is α = 7, β = 3.

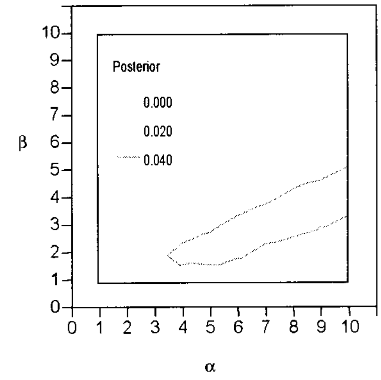

Figure 5. Posterior contours for α and β assuming uniform prior on integer grid 1,...,10.

Figures 4 & 5 show the joint posterior distribution of (α,β), using the uniform prior distribution. This prior puts probability 0.01 on each (α,β) pair. The figure shows the discrete joint posterior evaluated at each point on the grid. The posterior mode of the joint distribution is α = 7, β = 3, and this point has probability .063. Figure 5 shows the contours of the joint posterior. The figure shows the contours as continuous, but they are really only defined on the grid of integers between 1 and 10. We can see that each of the pairs (α,β) in the set {(4,2), (5,2), (6,3), (7,3), (8,3), (8,4), (9,3), (9,4), (10,4)} have posterior probabilities greater than or equal to .04.

Figure 6 shows the posterior distribution of the study effects, or the predictive distribution for an unobserved study effect. This distribution is a mixture of 100 beta distributions averaged across the joint posterior distribution of α and β shown in Figure 4. Compare this to posterior of Figure 3 in which the model assumed all 150 patients were exchangeable. Contrary to the posterior in Figure 3, here a 95% interval covers all but one of observed success probabilities of the 9 sites.

In this particular example, this posterior distribution has about the same variability as an empirical-Bayes estimate. Empirical-Bayes proce-

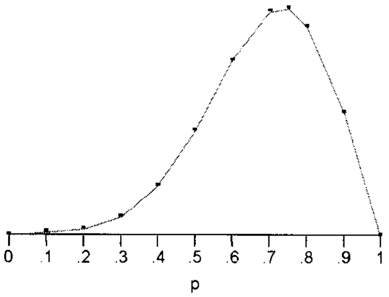

Figure 6. Posterior distribution of *p*.

dures use point estimates of α and β based on the current data to estimate this distribution. For example, if we used the maximum likelihood estimates of α and β (assuming the discrete parameter space), the distribution of study effects would be Beta(7,3). This distribution has a mean of 0.7 and a standard deviation of 0.138. The distribution presented in Figure 6 incorporates the additional uncertainty from α and β, but it is quite close to a Beta(7,3).

Table 2 provides the posterior means for study specific p_i or the equivalently predictive probability of success for the next patient in each of the 9 sites and for a patient at an unobserved site (column total).The table shows that the study-specific p_i are shrunk from their observed rates toward the overall mean. The observed probabilities for sites 2, 4, 5, and 9 are pulled upward toward 0.68, while the observed probabilities for sites 1, 6, 7, 8 are pulled down toward 0.68. Comparing the shrinkage between site 6 and 8, demonstrates that shrinkage increases as sample size decreases. Both studies had observed success probabilities of .78. However, the sample size of site 6 was 46, while the sample size of site 8 was only 9. To approximate, we can use the mode of the joint posterior of (α,β) which was (7,3). At this mode, $\alpha + \beta = 10$. Hence the mean of site 6 will be shrunk toward the overall mean (0.68) with weight 10/(10 + 46) = 0.18, while the mean of site 8

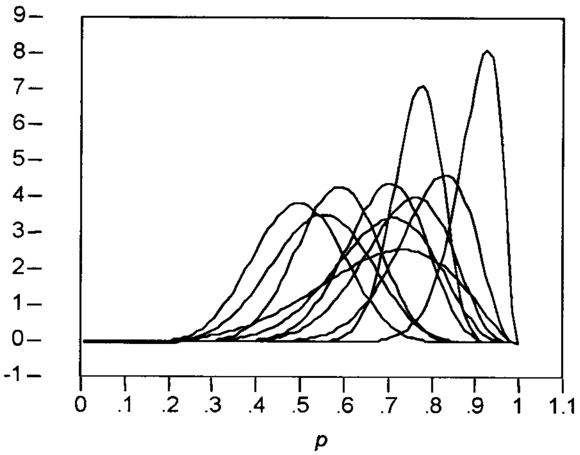

Figure 7. Posterior distributions of study specific effects for sites 1–9 and an unobserved center.

will be shrunk toward the overall mean with weight 10/(10 + 9) = 0.53. This is only an approximation, because we have used the posterior mode of (α,β) rather than average across the posterior distribution of (α,β). However, from the values in the table, it can be seen that it is quite close.

This example demonstrates a hierarchical model for the meta-analysis of a single treatment with a dichotomous outcome. Modeling is more complicated when two treatments are considered and the relative treatment effect is of interest. Then one must decide whether shrinkage should occur within each treatment, or within the relative treatment effect. Smith et al. (1996), Berry (2000), and Brophy and Joseph (2000) demonstrate the use of Bayesian hierarchical models in the context of meta-analyses comparing two treatments with dichotomous outcomes. All three use very different models. The paper by Smith is especially interesting because it compares the conclusions based on Bayesian hierarchical models to those from a Mantel-Haenszel-Peto method. The paper by Berry is especially interesting

Table 2. Posterior Means for Site-Specific Success Probabilities

Site	s_i	n_i	s_i/n_i	Predictive Probability Uniform prior on α and β
1	20	20	1.00	.90
2	4	10	.40	.53
3	11	16	.69	.69
4	10	19	.53	.57
5	5	14	.36	.48
6	36	46	.78	.77
7	9	10	.90	.80
8	7	9	.78	.73
9	4	6	.67	.68
Totals	106	150	.71	.68

in that it demonstrates how hierarchical models help dispel the controversy arising from perceived conflicting evidence in mega-clinical trials and meta-analyses. Lastly, the Brophy and Joseph paper is especially interesting, because it examines the controversial GUSTO clinical trial, and demonstrates how Bayesian hierarchical models can be adapted to assess the impact of protocol biases.

Modeling is also more complicated when outcomes are not discrete and covariates are included. Examples of this appear in Stangl (1995, 1996), Stangl and Huerta (2000) and Sargent et al. (2000). Stangl uses exponential, exponential mixture, and exponential changepoint models to compare the time-to-recurrence of depression for two treatment regimens. Stangl and Huerta use log-normal models to compare length-of-hospital-stay, pre versus post, implementation of a new managed care strategy incorporating both patient and hospital covariates. Sargent et al. demonstrates the use of random effects in Cox proportional-hazards models.

While the example presented here and the references cited thus far use hierarchical models for multi-center trials and meta-analyses, hierarchical models are also well suited to repeated measures and growth curve analyses. Racine-Poon and Wakefield (1996) use hierarchical models to fit pharmacokinetic-pharmacodynamic models. Albert and Chib (1996) demonstrate the use of hierarchical models for binary repeated measures data.

While the examples presented and referenced here are all from medical contexts, hierarchical models are abundant in agriculture, education (Bryk and Raudenbush, 1992, and Spiegelhalter and Marshall, 1999), engineering, policy (Daniels and Gatsonis, 1999), biology, social sciences (Draper, 1995), and other areas as well.

Software

The example presented in this paper was simple and calculations were carried out using JMP software (SAS Institute, 1989-2000). Models that are more complicated including nonconjugate distributions, continuous parameter spaces, and covariates at various stages of the model require more sophisticated software. Currently, the most widely used is BUGS (Spiegelhalter et al., 1996). This acronym stands for Bayesian Inference Using Gibbs Sampling. The software uses Markov chain Monte Carlo methods and allows models to be specified in a manner analogous to a graphical model. Smith et al. (1996) presents a straight forward example, as does the tutorial that comes with the software. Along with the user's manual there are two volumes of examples, most of which demonstrate the use of Bayesian hierarchical models to analyze data from multi-site trials, meta-analyses, repeated measures and longitudinal datasets.

The education version of the software can be downloaded from the internet free of charge at *http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml*

The website also lists published papers that used BUGS for their analyses. These papers are full of examples of hierarchical models in diverse applications including disease mapping, prevalence and incidence estimation, genetic linkage analysis, stock assessments, meta-analysis, longitudinal data, multi-center trials, and pharmacokinetics.

Conclusions

This paper has presented an elementary introduction to Bayesian methods and Bayesian hierarchical models. More complete descriptions can be found in *Bayesian Data Analysis* by Gelman et al. (1995) and *Bayes and Empirical Bayes Methods* by Carlin and Louis (1996).

References

Albert, J. and Chib S. (1996). Bayesian modeling of binary repeated measures data with application to crossover trials. In *Bayesian Biostatistics*, (eds. D.A. Berry and D.K. Stangl), Marcel Dekker, 577–600.

Berry, D.A., and Stangl, D.K. (1996). Bayesian methods in health-related research. In *Bayesian Biostatistics*, (eds. D.A. Berry and D.K. Stangl), Marcel Dekker, 3–66.

Berry, S. (2000). Meta-analysis versus large trials: Resolving the controversy. In *Meta-Analysis in Medicine and Health Policy*. (eds. D.K. Stangl and D.A. Berry), Marcel Dekker, 65–82.

Brophy, J. and Joseph, L. (2000). A Bayesian meta-analysis of randomized mega-trials for the choice of thrombolytic agents in acute myocardial infarction. In *Meta-Analysis in Medicine and Health Policy*. (eds. D.K. Stangl and D.A. Berry), Marcel Dekker, 83–104.

Bryk A. S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* London: Sage.

Carlin, B. and Louis, T. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

Clyde, M., Mueller, P., and Parmigiani, G. (1996). Inference and design strategies for a hierarchical logistic regression model. In *Bayesian Biostatistics*, (eds. D.A. Berry and D.K. Stangl), Marcel Dekker, 297-320.

Daniels, M. and Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, 94: 29–42.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20, 115–147.

DuMouchel W. (1989). Bayesian metaanalysis. *Statistical Methodology in the Pharmaceutical Sciences*, (ed. Berry, D.A.), New York: Marcel Dekker, 509–529.

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

Racine-Poon, A. and Wakefield, J. (1996). Bayesian analysis of population pharmacokinetic and instantaneous pharmacodynamic relationships. In *Bayesian Biostatistics*, (eds. D.A. Berry and D.K. Stangl), Marcel Dekker, 355–388.

Sargent, D. Zee, B., Milan, C, Torri, V, and Francini G. (2000). In *Meta-Analysis in Medicine and Health Policy*. (eds. D.K. Stangl and D.A. Berry), Marcel Dekker, 255–276.

SAS Institute Inc. JMP version 4. (SAS Institute Inc., Cary, NC, 1989-2000).

Smith, T., Speigelhalter, D. and Parmar, M. (1996). Bayesian meta-analysis of randomized trials using graphical models and BUGS. In *Bayesian Biostatistics*, (eds. D.A. Berry and D.K. Stangl), Marcel Dekker, 411–428.

Spiegelhalter, D. and Marshall, E. (1999). Inference-robust institutional comparisons: A case study of school examination results. In *Bayesian Statistics* 6, (eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith), Oxford Press, 613–630.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). *BUGS: Bayesian inference Using Gibbs Sampling*, Version 0.5, MRC Biostatistics Unit, Cambridge, U.K.

Stangl, D. (1995). Prediction and decision-making using Bayesian hierarchical models, *Statistics in Medicine,* 14:2173–2190.

Stangl, D. and Huerta G. (2000). Assessing the impact of managed-care on the distribution of length-of-stay using Bayesian hierarchical models. *Lifetime Data Analysis*, 6: 123–139.

Stangl, D. and Berry, D.A. (2000). *Meta-Analysis in Medicine and Health Policy*. New York: Marcel Dekker.

Stangl, D. (1996). Hierarchical analysis of continuous-time survival models. In *Bayesian Biostatistics*, (Eds. D.A. Berry and D.K. Stangl), Marcel Dekker, 429–451.