Classifying individuals as physiological responders using hierarchical modeling

Richard J. Barker and Matthew R. Schofield

Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand

Submitted 12 December 2007; accepted in final form 13 May 2008

Barker RJ, Schofield MR. Classifying individuals as physiological responders using hierarchical modeling. J Appl Physiol 105: 555–560, 2008. First published May 29, 2008; doi:10.1152/japplphysiol.01317.2007.—We outline the use of hierarchical modeling for inference about the categorization of subjects into “responder” and “nonresponder” classes when the true status of the subject is latent (hidden). If uncertainty of classification is ignored during analysis, then statistical inference may be unreliable. An important advantage of hierarchical modeling is that it facilitates the correct modeling of the hidden variable in terms of predictor variables and hypothesized biological relationships. This allows researchers to formalize inference that can address questions about why some subjects respond and others do not. We illustrate our approach using a recent study of hepcidin excretion in female marathon runners (Roecker L, Meier-Buttermilch R, Brechte L, Nemeth E, Ganz T. Eur J Appl Physiol 95: 569–571, 2005).

Hepcidin is a peptide hormone associated with iron metabolism; levels are increased by inflammation, and its excess is associated with low iron levels. Roecker et al. noted increased hepcidin levels after the marathon for most, but not all, of their sample and speculated that the sample included two types of runners: “responders” and “nonresponders.” Roecker et al. carried out no formal analysis of the potential responder vs. nonresponder classification. Instead, they simply identified as responders the subjects who showed a 4- to 27-fold increase in urinary hepcidin. Any subjects under this threshold were classified as nonresponders.

Studies in which there is interest in making inference about responders and nonresponders are reasonably common. Subjects are usually classified on the basis of response variables, with statistical testing procedures often used for formal discrimination. Townsend et al. (16) considered the response of trained endurance athletes to simulated altitude-related hypoxia. They classified subjects as “strong responders,” “moderate responders,” and nonresponders according to the outcome of statistical analysis. Baylor and Hackney (1) classified their subjects as responders if their measured response was >1.5 SD from the mean and nonresponders if their response was <1.5 SD from the mean. Mahler et al. (12) examined the response to exercise training of patients with chronic obstructive pulmonary disorder. They classified their subjects as responders and nonresponders according to peak O2 uptake measurements before and after a training program.

A common feature of the above-mentioned studies is that the true state of the subject is unknown and must be inferred from formal statistical analysis or use of ad hoc classification methods in lieu of formal analysis. Inference then proceeds on the assumption that the classification is correct. There are three related problems with this approach. 1) Subjects may be misclassified. 2) Uncertainty of classification is not fully accounted for in the analysis. Ignoring sources of uncertainty leads to parameter estimates that appear to be more precise than they really are. This results in incorrect inference. 3) Incorrect or inefficient statistical procedures may be used, leading to lost learning opportunities.

It is well known in the statistical literature that misclassification of a binary response variable can lead to biased estimation (4), particularly in small samples (5). Even low misclassification rates can have major effects on inference (2, 4). One approach to dealing with misclassified data is use of robust methods of inference (5, 17).

The fundamental issue with misclassification is that the nominal classification, which is observed, is not the same as the true classification, which is latent. Instead of robust methods of inference, a more natural approach is an explicit hierarchical model, in which the problem is broken down into simple components. One component involves the distribution of the observed response conditional on the true classification, and the other component models the true classification in terms of explanatory variables of interest. This approach facilitates understanding by users and can be easily adapted to incorporate additional sources of information, for example, if classification into responder and nonresponder classes is based on repeated measures or other multiple sources of information.

Using the study of Roecker et al. (14) as an example, we illustrate how hierarchical models can be used to make inference when the classification of subjects into the categories responder and nonresponder is latent. Hierarchical models are
particularly useful for complex models that can be factored into a sequence of simple conditional models. Our approach to inference is Bayesian (7), although we discuss potential frequentist approaches to this problem.

**DATA AND MODEL**

The data used by Roecker et al. (14) are not publicly available. Instead, we transcribed data from their Fig. 1. We define $Y_i$ as the log-scale difference between the maximum hepcidin level measured in subject $i$ after the marathon and the level measured before the race and $R_i$ as an indicator for whether subject $i$ is a responder ($R_i = 1$) or a nonresponder ($R_i = 0$). Note that $R_i$ is latent.

We assume that 1) $Y_i$ has a normal $[\mathcal{N}(\mu_i,\sigma^2)]$ distribution, with mean $\mu_i = R_i \beta$ and variance $\sigma^2$ and 2) $R_i$ is a Bernoulli random variable with parameter $p$. That is, the log-scale differences in the hepcidin data are assumed to have been drawn from a mixture of normal distributions, with $p$ defining the proportion of responders. For simplicity, we have assumed a common variance for the two normal distributions in the mixture; however, this is not necessary and can be relaxed if required, but more complex models will require more data for reliable inference.

Departing slightly from the definition of Roecker et al. (14) of a responder as an athlete whose peak measured hepcidin level was at least four times the level measured before the race, we defined responders as athletes whose expected hepcidin levels were elevated after the race and nonresponders as athletes whose expected (mean) hepcidin levels were unchanged. Note that our classification refers to the expected, not the measured, hepcidin level. The parameter $\beta$ expresses the size of the difference in responders. Our interest is in predicting group membership for an individual athlete with an assessment of reliability of this allocation.

The model can be written as the product of two factors: the density of the hepcidin response conditional on the value for $R$ $[f(Y|R, \beta, \sigma^2)]$ and the distribution of the latent variable $R$ $[f(R)]$. Bayesian inference methods are well suited to fitting hierarchical models with inference about the parameters and latent variables based on the appropriate posterior distributions. Inasmuch as the exact form of the posterior distributions required for inference is difficult to compute, we draw a large sample from the posterior distribution using a computational technique called the Gibbs sampler (3). Using posterior samples obtained from the Gibbs sampler, we can sketch the posterior distribution and summarize it as accurately as we like simply by ensuring that the sample we have taken is sufficiently large. The Gibbs sampler has been implemented in the computer program WinBUGS (11). Input to the model is in the form of a model description (see Appendix), which is then compiled.

Priors for the parameters were chosen to represent vague prior knowledge subject to the constraint that $\beta$ is positive, inasmuch as a responder has elevated hepcidin levels by definition. To assess prior sensitivity, two sets of priors were chosen. Set I is a uniform distribution for each of $p$ $[U(0,1)]$, $\beta$ $[U(0,10)]$, and $\sigma$ $[U(0,100)]$. Set II is a $Beta(1/2,1/2)$ prior for $p$, a half-normal $[HN(0.01)]$ prior for $\beta$, and a $\gamma[Ga(0.001,0.001)]$ prior for $\sigma$.

**RESULTS**

For the hepcidin data, the posterior densities are shown in Fig. 1 and summary statistics are shown in Table 1 for a sample of 100,000 drawn from the joint posterior distribution. Of particular interest is $p$, for which the posterior median was 0.59 with a 95% credible interval of $(0.22,0.96)$ for prior set I and 0.61 with a 95% credible interval of $(0.16,1.00)$ for prior set II. This suggests that ~60% of athletes in the study population are responders, although this value could be as low as 16–22% or as high as 96–100%. Comparison of results for the two sets of priors indicates modest sensitivity to the prior. The wide credible intervals reflect the small sample sizes. For precise inference about $p$ or $\beta$, >14 subjects are required. The posterior probabilities of $R = 1$ for the 14 subjects (Table 2) indicate the strength of evidence that a particular subject is a responder. For example, the maximum postrace measurement of hepcidin for subject 8 was 27.3 times the prerace measurement; the probability that subject 8 is a responder is 0.96 for both sets of priors. The maximum postrace hepcidin measurement for subject 1 was only 4.4 times the prerace measurement, and the probability that subject 1 is a responder is 80% (prior set I) or 83% (prior set II). Thus there is a 20% (prior set I) or 17% (prior set II) chance that subject 1 is a nonresponder, in which case the apparent postrace elevation of hepcidin is explained by within-subject variation in measured hepcidin. The posterior probabilities for $R = 1$ appear to be relatively insensitive to the choice of prior, agreeing to within 3%.

On the basis of our analysis, the grouping of subjects by Roecker et al. (14) is good: all the subjects they classified as responders have a $\approx$80% chance of being a responder under the model, despite the relatively wide credible intervals for the parameters in the model. This good discrimination among subjects reflects the separation of the hepcidin responses identified by Roecker et al. between subjects with a $\geq$4-fold increase after the race and subjects with a $\leq$1.6-fold increase. The strength of classification is not as high for the nonresponders as for the responders: no better than 82% (prior set I) or 83% (prior set II) for the 14 subjects (Table 2) indicate the strength of evidence that a particular subject is a responder. For example, the maximum postrace measurement of hepcidin for subject 8 was 27.3 times the prerace measurement; the probability that subject 8 is a responder is 0.96 for both sets of priors. The maximum postrace hepcidin measurement for subject 1 was only 4.4 times the prerace measurement, and the probability that subject 1 is a responder is 80% (prior set I) or 83% (prior set II). Thus there is a 20% (prior set I) or 17% (prior set II) chance that subject 1 is a nonresponder, in which case the apparent postrace elevation of hepcidin is explained by within-subject variation in measured hepcidin. The posterior probabilities for $R = 1$ appear to be relatively insensitive to the choice of prior, agreeing to within 3%.

On the basis of our analysis, the grouping of subjects by Roecker et al. (14) is good: all the subjects they classified as responders have a $\approx$80% chance of being a responder under the model, despite the relatively wide credible intervals for the parameters in the model. This good discrimination among subjects reflects the separation of the hepcidin responses identified by Roecker et al. between subjects with a $\geq$4-fold increase after the race and subjects with a $\leq$1.6-fold increase. The strength of classification is not as high for the nonresponders as for the responders: no better than 82% (prior set I) or 83% (prior set II) for the 14 subjects (Table 2) indicate the strength of evidence that a particular subject is a responder. For example, the maximum postrace measurement of hepcidin for subject 8 was 27.3 times the prerace measurement; the probability that subject 8 is a responder is 0.96 for both sets of priors. The maximum postrace hepcidin measurement for subject 1 was only 4.4 times the prerace measurement, and the probability that subject 1 is a responder is 80% (prior set I) or 83% (prior set II). Thus there is a 20% (prior set I) or 17% (prior set II) chance that subject 1 is a nonresponder, in which case the apparent postrace elevation of hepcidin is explained by within-subject variation in measured hepcidin. The posterior probabilities for $R = 1$ appear to be relatively insensitive to the choice of prior, agreeing to within 3%.

On the basis of our analysis, the grouping of subjects by Roecker et al. (14) is good: all the subjects they classified as responders have a $\approx$80% chance of being a responder under the model, despite the relatively wide credible intervals for the parameters in the model. This good discrimination among subjects reflects the separation of the hepcidin responses identified by Roecker et al. between subjects with a $\geq$4-fold increase after the race and subjects with a $\leq$1.6-fold increase. The strength of classification is not as high for the nonresponders as for the responders: no better than 82% (prior set I) or 83% (prior set II) for the 14 subjects (Table 2) indicate the strength of evidence that a particular subject is a responder. For example, the maximum postrace measurement of hepcidin for subject 8 was 27.3 times the prerace measurement; the probability that subject 8 is a responder is 0.96 for both sets of priors. The maximum postrace hepcidin measurement for subject 1 was only 4.4 times the prerace measurement, and the probability that subject 1 is a responder is 80% (prior set I) or 83% (prior set II). Thus there is a 20% (prior set I) or 17% (prior set II) chance that subject 1 is a nonresponder, in which case the apparent postrace elevation of hepcidin is explained by within-subject variation in measured hepcidin. The posterior probabilities for $R = 1$ appear to be relatively insensitive to the choice of prior, agreeing to within 3%.
Table 1. Posterior density summaries for parameters in the two-part normal-mixtures model fitted to hepcidin data of Roecker et al. (14)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
<th>2.5</th>
<th>Median</th>
<th>97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior set I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\beta)</td>
<td>1.898</td>
<td>0.568</td>
<td>0.702</td>
<td>1.973</td>
<td>2.856</td>
</tr>
<tr>
<td>(P)</td>
<td>0.592</td>
<td>0.189</td>
<td>0.216</td>
<td>0.587</td>
<td>0.961</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.987</td>
<td>0.411</td>
<td>0.466</td>
<td>0.893</td>
<td>1.954</td>
</tr>
<tr>
<td>Prior set II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\beta)</td>
<td>1.994</td>
<td>1.162</td>
<td>0.749</td>
<td>2.003</td>
<td>2.917</td>
</tr>
<tr>
<td>(P)</td>
<td>0.606</td>
<td>0.211</td>
<td>0.162</td>
<td>0.596</td>
<td>0.995</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.910</td>
<td>0.387</td>
<td>0.445</td>
<td>0.800</td>
<td>1.844</td>
</tr>
</tbody>
</table>

Prior set I has uniform priors for each of \(p\;[U(0,1)], \beta; [U(0,10)],\) and \(\sigma\; [U(0,100)]. Prior set II has a beta \([Beta(1/1,1/2)]\) prior for \(p\), a half-normal \([HN(0.01)]\) prior for \(\beta\), and a gamma \([Gamma(0.001,0.001)]\) prior for \(\tau = 1/\sigma^2\).

set I) or 81% (prior set II) for subject 10. Subject 12, with a maximum postrace hepcidin measurement that was 1.6 times higher than the prerace measurement, has only a 68% (prior set I) or 81% (prior set II) chance of correct classification by Roecker et al.

Although the analysis we have described, that of predicting the status of subjects, is a step forward, it is not especially interesting. Our real interest should be in modeling responses in terms of higher-level biological processes. That is, if we knew for sure which subjects were responders and which were not, we would almost certainly be interested in modeling the true classification in terms of explanatory variables and processes. We can integrate this type of modeling into the framework we have outlined.

Roecker et al. (14) commented that the nonresponders had high baseline hepcidin values and that this was possibly related to intensity of prerace training. We can use the prerace information to model the classification. We can do this to 1) better understand the classification process biologically and 2) improve the classification by taking advantage of a known covariate. It is straightforward to modify the model to include the dependence between the classification variable \(R\) and any covariates. For example, we can use a logistic regression-type model to allow the logit of the probability that a subject is a responder to be a linear regression on prerace hepcidin levels (see APPENDIX).

For analysis, we modeled the logit of \(p_i\) as a simple linear regression on \(z_i\), the standardized natural logarithm of the prerace hepcidin level for subject \(i\)

\[
\logit(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \gamma_0 + \gamma_1z_i
\]

The covariate \(z\) has been added to the data input section and is the prerace hepcidin level (on the log scale) standardized by subtracting the mean value of 3.112 and then dividing by the standard deviation of 1.029. For priors, we used the uniform priors from the first analysis with the substitution of logistic \([logit(0,1)]\) priors on \(\gamma_0\) and \(\gamma_1\). This use of the logistic prior induces an approximately uniform prior on \(p_i\). Posterior summaries from fitting this model are provided in Tables 3 and 4. The parameter 1 indicates quite a strong effect of the prerace hepcidin level on the probability of classifying a subject as a responder. The effect of including the covariate is to increase certainty that the six subjects classified by Roecker et al. (14) as nonresponders are, in fact, nonresponders and to reduce the certainty of the responder classification for subjects 1 and 2. We can state that, given her prerace level of hepcidin, there is a posterior probability of 100 – 31 = 69% that subject 1 is a nonresponder.

We can also use this example to compare the analysis described above with a naive analysis in which logistic regression is carried out using the classifications of Roecker et al. (14) as though they were correct. We fitted the logistic regression model using the same prior on the parameters \(\gamma_0\) and \(\gamma_1\) used in the preceding analysis.

Ignoring the uncertainty in classification leads to a posterior density that is too narrow and shifted toward zero (posterior mean = –2.058 (SD 0.916); Fig. 2), indicating a tendency for the naive analysis to be biased, in this case, underestimating the effect of the covariate, and for it to underestimate the uncertainty in parameter estimates. Although there is considerable overlap in the posterior distributions, this is due to the small sample sizes.

To remove some of the sample-size effect, we repeated the comparison of the two modeling approaches by simulating data under the model using the posterior means in Table 3 as the true parameters and employing 100, instead of 14, subjects. For

Table 2. Posterior predictive probabilities that an individual is a responder for hepcidin data of Roecker et al. (14)

<table>
<thead>
<tr>
<th>Subject No.</th>
<th>Roecker Classification</th>
<th>(\Delta\ln(\text{hepcidin}))</th>
<th>Max(Post)/Pre</th>
<th>Pr(Resp)</th>
<th>(\text{Pr(Resp)})</th>
<th>Prior Set I</th>
<th>Prior Set II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Resp</td>
<td>1.5</td>
<td>4.4</td>
<td>0.80</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Resp</td>
<td>1.5</td>
<td>4.4</td>
<td>0.80</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Resp</td>
<td>2.1</td>
<td>7.9</td>
<td>0.90</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Resp</td>
<td>1.5</td>
<td>4.4</td>
<td>0.80</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Resp</td>
<td>2.8</td>
<td>16.0</td>
<td>0.94</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Resp</td>
<td>2.5</td>
<td>12.7</td>
<td>0.93</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Resp</td>
<td>1.9</td>
<td>6.6</td>
<td>0.89</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Resp</td>
<td>3.3</td>
<td>27.3</td>
<td>0.96</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Nonresp</td>
<td>0.0</td>
<td>1.0</td>
<td>0.24</td>
<td>0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Nonresp</td>
<td>–0.6</td>
<td>0.6</td>
<td>0.18</td>
<td>0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Nonresp</td>
<td>–0.3</td>
<td>0.7</td>
<td>0.21</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Nonresp</td>
<td>0.5</td>
<td>1.6</td>
<td>0.32</td>
<td>0.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Nonresp</td>
<td>0.3</td>
<td>1.3</td>
<td>0.28</td>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Nonresp</td>
<td>0.4</td>
<td>0.7</td>
<td>0.20</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pr(Resp), predictive probability that a subject is a responder; Resp, responder; Nonresp, nonresponder; \(\Delta\ln(\text{hepcidin})\), difference between the maximum value of hepcidin recorded for an individual in 3 postrace measurements and the prerace measurement. Uniform priors for sets I and II are given in Table 1 footnote.

Table 3. Posterior density summaries for parameters in the two-part normal-mixtures model fitted to hepcidin data of Roecker et al. (14), including the prerace level of hepcidin as a covariate that influences the probability that a subject is a responder

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
<th>2.5</th>
<th>Median</th>
<th>97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\gamma_0)</td>
<td>0.050</td>
<td>0.911</td>
<td>–1.822</td>
<td>0.065</td>
<td>1.802</td>
</tr>
<tr>
<td>(\gamma_1)</td>
<td>–2.820</td>
<td>1.413</td>
<td>–6.180</td>
<td>–2.613</td>
<td>–0.682</td>
</tr>
<tr>
<td>(\beta)</td>
<td>2.203</td>
<td>0.349</td>
<td>1.514</td>
<td>2.198</td>
<td>2.902</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.780</td>
<td>0.225</td>
<td>0.470</td>
<td>0.739</td>
<td>1.334</td>
</tr>
</tbody>
</table>

J Appl Physiol • VOL 105 • AUGUST 2008 • www.jap.org
the naive analysis, we used the criterion of Roecker et al. (14) to classify the subjects. The larger sample size has resulted in a greater contrast between the two posterior distributions, and the bias resulting from the naive analysis is now more marked (Fig. 3). Even with 100 subjects, the posterior distributions are quite wide, indicating that precise inference for covariate effects requires large sample sizes.

**DISCUSSION**

Hierarchical models are especially useful for latent (hidden) variable modeling, as in the hepcidin example. These models allow researchers to better understand relationships between predictor variables and the hidden variable of interest, and they also facilitate individual prediction as highlighted by the example.

![Fig. 2](image-url) Posterior density for the parameter $\gamma_1$ in the logistic regression $\logit(p) = \gamma_0 + \gamma_1 z_i$, where $z_i$ is the prerace hepcidin level, and compared with a simple logistic regression ignoring the uncertainty in classification. Responses for 100 subjects were simulated using the posterior means in Table 3 as true values of the parameters. True value for $\gamma_1$ was $-2.82$.

It is important to understand that there is a distinction between the true classification of a subject and nominal classification when classification methods that are subject to error have been used to distinguish among subjects. Statistical analysis that assumes that the data have been correctly classified can potentially lead to misleading analyses if classification is uncertain.

The analysis that we have outlined accounts for the inherent uncertainty in classification through the use of a joint model, with one component that models the hidden classification and the other that models the response conditional on the hidden classification. In effect, the analysis integrates across all possible classifications, but at the same time it allows prediction of the classification with an assessment of uncertainty. Note that this is quite different from a two-stage analysis, in which the classification is first predicted and then these predictors are used as data, as in the naive logistic regression analysis discussed above. Such an approach is almost certainly biased, as illustrated by our simulation, and does not fully account for all uncertainties, leading to overconfidence in the results of the analysis.

The approach that we have outlined makes use of what is known as the “complete data likelihood,” in which the statistical model explicitly includes the hidden variable of interest. In this case, the hidden value is the true state of the subject. The value of the complete-data-likelihood approach to modeling is that it facilitates proper modeling of variables of interest, despite the fact that they are not observed. Thus emphasis can move away from simply predicting the hidden state of the subject and move to formal modeling in terms of questions of interest. Although in the hepcidin example our model for response status was a simple logistic regression model, exploration of more complex models is relatively straightforward. It is fitting that these models allow researchers to formalize inference that can address questions about why some subjects respond and others do not.
One extension that is of interest is modeling of the $z_i$ covariates when they are measured with error, as will be common in practice. With use of the hierarchical model approach, it is simple to add a factor that models the distribution of the observed covariate $z_i$. If the true value of the covariate is $\xi_i$, and the relationship between $z_i$ and $\xi_i$ is expressed through the model $[z_i|\xi_i, \theta]$, then we simply add this component to the model along with priors for $\xi$ and $\theta$. Also, we would now model $p$ in terms of $\xi_i$ instead of $z$. For example, it might be that $z_i$ is normally distributed with mean $\xi_i$ and variance $\theta$. If there is sufficient information to allow estimation of the $\theta$ parameters, for example, if measurements of $z_i$ are replicated for each individual, then the extended model allows inference about the latent variable $R$ incorporating uncertainty about $\xi$ and $\theta$.

Once a model has been specified, the next problem is how to make statistical inference. Our approach is Bayesian. Bayesian methods are well suited to fitting hierarchical models with latent and missing variables and are flexible in allowing arbitrary choices of distributions and forms of response functions involving covariates. A necessary step in Bayesian analysis is the specification of prior uncertainty about parameters. As a general rule, prior choice will be most influential when we have few observations. In such cases, a good approach is evaluation of the sensitivity of inference to choices of priors. Although there were only 14 subjects in this study, there was little difference in results between the two sets of priors that we considered. These priors were chosen to represent vague prior knowledge about the parameters. Prior choice is a major issue in Bayesian inference (for more discussion see Refs. 7 and 9).

There are frequentist alternatives to Bayesian inference for latent variable models based on structural equation models and extensions such as the generalized linear latent variable model (8). The generalized linear latent variable model extends generalized linear models to allow modeling of latent responses, and models can be fitted by maximum likelihood. Each of the examples we have considered can be expressed as a generalized linear latent variable model; however, the Bayesian approach allows extension of these to nonlinear cases and can easily include modeling of uncertainty associated with covariate measurement.

Frequentist software, such as Mplus (13), is also readily available. Methods such as maximum likelihood are based on the observed data likelihood, which is found by integrating out latent components. This can be done using the expectation-maximization algorithm (6). In simple cases, including the two models discussed in RESULTS, this can be done explicitly, and maximum likelihood inference about parameters is straightforward. Given parameter estimates, inference about latent variables, such as $R$, can then be computed as functions of these parameter estimates.

A disadvantage of maximum likelihood is that, except in a few cases, inference is asymptotic and only valid if the actual sample size is large. Estimators are also assumed to follow a normal distribution, an assumption that is also only valid in large samples. A result is that estimates of parameters and predictions for latent variables may be unreliable in studies based on small samples, as is common in practice. Another disadvantage of maximum likelihood inference is that the likelihood function may have multiple peaks, especially in hierarchical models. The presence of multiple peaks and posterior densities that are far from normally distributed are features that can, with care,1 be readily identified using Bayesian inference methods. For example, the posterior densities in Fig. 1 appear far from normal. Bayesian inference thus has the advantage of not requiring large samples, but issues concerning priors must then be considered as discussed above.

The analysis that we have described specifically deals with the case where the physiological response is assumed to be dichotomous. It may be that the latent factor has more than two levels (1, 16), in which case the analysis presented here would be misleading. It is important in all model-based analysis that care be taken to ensure that the model is theoretically justified. If theory indicates that there should be more than two classes, the hierarchical modeling approach that we have described can be readily extended. If researchers are unsure of the exact nature of the response but have a small set of candidates in mind, then it is also possible to consider a set of models using multimodel inference (10). Nevertheless, assessment of model fit is also an important step, as in all statistical analysis. For Bayesian inference, model assessment should be based on the posterior predictive checking (7), in which the observed data are compared with data generated under the model.

APPENDIX

WinBUGS Code

Two-part mixture of normal distributions with no covariate for $R$.

Model

# Likelihood
for(i in 1:14) {
  R[i]~dbern(p)
  mu[i]<-R[i]*beta
deltahep[i]~dnorm(mu[i],tau)
}

# Priors (set-1)
p~dbeta(1,1)
beta~dunif(0,10)
sd~dunif(0,100)
tau<-1/pow(sd,2)
}

Data

list(deltahep=c(1.476016,1.476016,2.066423,1.476016,2.77491,2.538748,1.889301,3.306276,0.059041,-0.2952,0.472325,0.295203,0.35424))

Two-part mixture of normal distributions with a covariate for $R$.

The covariate is the pre-race level of hepcidin.

Model

{
  # Likelihood
  for(i in 1:14) {
    R[i]~dbern(p[i])
    mu[i]<-R[i]*beta
deltahep[i]~dnorm(mu[i],tau)
  
1 When software such as WinBUGS is used, it is important that widely dispersed starting values are employed to ensure that posterior sampling occurs over the entire range of the posterior distribution.
logit(p[i]) <- b0 + b1*z[i]
}

# Priors
gamma0 ~ dlogis(0,1)
gamma1 ~ dlogis(0,1)
beta~dunif(0,10)
sd~dunif(0,100)
tau <- 1/pow(sd,2)

Data
list(deltahep = c(1.476016, 1.476016, 2.066423, 1.476016, 2.77491, 2.538748, 1.889301, 3.306276, 0, 0.059041, 0.2952, 0.472325, 0.295203, 0.35424),
    z = c(1.1071, 0.1312, -0.3854, -0.8447, -0.9021, -0.9021, -1.4762, -1.4762, 1.2793, 1.2793, 0.8775, 0.7627, 0.5331, 0.0164))

REFERENCES