Biometrics 61, 74–85
March 2005

Multiple Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data

Andrew Gelman,[1,*] Iven Van Mechelen,[2] Geert Verbeke,[3] Daniel F. Heitjan,[4] and Michel Meulders[2]

[1]Department of Statistics, Columbia University, New York 10027, U.S.A.
[2]Department of Psychology, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium
[3]Biostatistical Centre, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium
[4]Division of Biostatistics, University of Pennsylvania, Philadelphia 19104, U.S.A.
[*]*email:* gelman@stat.columbia.edu

Summary. In problems with missing or latent data, a standard approach is to first impute the unobserved data, then perform all statistical analyses on the *completed* dataset—corresponding to the observed data and imputed unobserved data—using standard procedures for complete-data inference. Here, we extend this approach to model checking by demonstrating the advantages of the use of completed-data model diagnostics on imputed completed datasets. The approach is set in the theoretical framework of Bayesian posterior predictive checks (but, as with missing-data imputation, our methods of missing-data model checking can also be interpreted as "predictive inference" in a non-Bayesian context). We consider the graphical diagnostics within this framework. Advantages of the completed-data approach include: (1) One can often check model fit in terms of quantities that are of key substantive interest in a natural way, which is not always possible using observed data alone. (2) In problems with missing data, checks may be devised that do not require to model the missingness or inclusion mechanism; the latter is useful for the analysis of ignorable but unknown data collection mechanisms, such as are often assumed in the analysis of sample surveys and observational studies. (3) In many problems with latent data, it is possible to check qualitative features of the model (for example, independence of two variables) that can be naturally formalized with the help of the latent data. We illustrate with several applied examples.

Key words: Bayesian model checking; Exploratory data analysis; Multiple imputation; Nonresponse; Posterior predictive checks; Realized discrepancies; Residuals.

1. Introduction

1.1 *Difficulties of Model Checking with Missing and Latent Data*

The fundamental approach of goodness-of-fit testing is to display or summarize the observed data, and compare this to what might have been expected under the model. If there are systematic discrepancies between the data summaries and their reference distribution under the assumed model, this implies a misfit of the model to the data. Model checks include analytical methods such as χ^2 and likelihood ratio tests, and graphical methods such as residual and quantile plots. In missing- and latent-data settings, two complications arise that can in practice often lead to models being checked in only a cursory fashion if at all.

The first complication comes because in missing-data situations the reference distribution of a data summary, whether analytical or graphical, is implicitly determined by the data that *could* have been seen under the model. As a result, comparing the data to what could have been observed requires a model for the missing-data mechanism—in order to obtain a reference distribution for *which* data points are observed—as well as a model for the data themselves. Modeling the process that generated the missing data can be difficult, and any requirement that this be done will drastically reduce the practicality of model checking procedures. As a result, model checking is generally applied either to complete-data segments of the problem or only approximately.

The second complication arises with the latent data (defined broadly to include, for example, group-level parameters in hierarchical models). Even if there is a full model for the observation process (and, hence, it is not a problem to simulate replications of the observed data), the latent data may be of scientific interest. As such, we may wish to construct tests using these latent categories or variables. As an example, one may think of regression diagnostics in hierarchical models involving residuals calculated on the basis of group-level parameters (that are considered as latent data). Unlike standard residuals that are difficult to interpret for hierarchical models (see Hodges, 1988), those based on latent data are independent.

The characterization of unobserved data as "missing" or "latent" is somewhat arbitrary; as is well known in the

context of EM and similar computational algorithms, latent and missing data have the same inferential standing as unknown quantities with a joint distribution under a probability model. For the purpose of this article, we distinguish based on the interpretation of the completed dataset: *missing data* have the same structure as observed data whereas *latent data* are structurally different. For example, Section 4.2 describes a situation in which children's ages are reported rounded to the nearest 1, 6, or 12 months (see the top graph in Figure 6). We consider the children's true ages as latent data, in that if we were given the completed dataset (all the true ages and all the reported ages), then we would wish to distinguish between the true and reported ages—they play different roles in the model. In fact, in this example, the type of rounding (whether to 1, 6, or 12 months) is another latent variable. How could missing data arise in this context? If some of the children in the study were missing some recorded covariates, or if age were not even reported for some children, these would be missing data—because, once these data were imputed, they would be structurally indistinguishable from the observed data.

1.2 *Predictive Checks with Unobserved Data*

In this article, we propose to resolve difficulties of model checking in missing- and latent-data settings using the framework of Bayesian posterior predictive checks (Rubin, 1984; Gelman, Meng, and Stern, 1996). The general idea of predictive assessment is to evaluate any model based on its predictions given currently available data (Dawid, 1984). Predictive criteria can be used as a formal approach for the evaluation and selection of models (Geisser and Eddy, 1979; Seillier-Moiseiwitsch, Sweeting, and Dawid, 1992). Here we focus on graphical and exploratory comparisons (as in Buja et al., 1988; Buja, Cook, and Swayne, 1999; Gelman, 2004), in addition to numerical summaries based on test statistics.

Gelman et al. (1996) define posterior predictive checks as comparisons of observed data y_{obs} to replicated datasets $y_{\text{obs}}^{\text{rep}}$ that have been simulated from the posterior predictive distribution of the model with parameters θ. In this article, we extend posterior predictive checking within the context of missing or latent data by including unobserved data in the model checks. Model checking then will be applied to *completed data*, which will typically require multiple imputations of the unobserved data.

The approach of including unobserved data in model checks will be shown to yield various advantages. The situation is similar to that of the EM algorithm (Dempster, Laird, and Rubin, 1977), data augmentation (Tanner and Wong, 1987), and multiple imputation (Rubin, 1987, 1996). The EM and data augmentation algorithms take advantage of explicitly acknowledging unobserved data in finding posterior modes and simulation draws. The multiple imputation approach similarly accounts for uncertainties in missing data for Bayesian inference. The approach proposed in this article then completes this idea for model checking. In general, a key advantage of completed-data model checks is that they can be directly understandable in ways that observed-data checks are not, allowing, for example, graphical model checks (analogous to residual plots) that are interpretable without need for formal computation of reference distributions. We shall illustrate this with several instances of missing- and latent-data problems from a wide range of application areas, with various statistical models, and a variety of graphical displays.

Despite the simplicity of the approach, we have seen it only rarely in the statistical literature (with exceptions including the analysis of realized residuals in linear models and censored-data models, Chaloner and Brant, 1998, and Chaloner, 1991; and latent continuous responses in discrete-data regressions, Albert and Chib, 1995). We attribute this to an incomplete conceptual foundation. We hope that this article, by placing completed-data diagnostics in a general framework (in which observed-data test statistics are a special case), and illustrating in a variety of applications, will motivate their further use.

This article is organized as follows. Section 2 defines the basic notation and ideas underlying our recommended approach, first for missing and then for latent data. Sections 3 and 4 present several examples from applied work by ourselves and others, and Section 5 discusses some of the lessons we have learned from these applied examples.

2. Notation, Underlying Ideas, and Implementation

We set up our completed-data model checking using the theoretical framework of Little and Rubin (1987) and Gelman et al. (2003, Chapter 7) for Bayesian inference with missing data. The two relevant tasks are defining the predictive distribution for replicated data, and choosing the completed-data summaries to display. We discuss the theoretical issues in detail in Section 2.1 for the case of missing data and then in Section 2.2 briefly consider the latent data setting. We present our approach in algorithmic form in Section 2.3.

2.1 *Missing Data*

2.1.1 *Bayesian notation using inclusion indicators.* We use the term *missing data* for potentially observed data that, unintentionally or by design, have been left unobserved. Consider observed data y_{obs} and missing data y_{mis}, which together form a "completed" dataset $y_{\text{com}} = (y_{\text{obs}}, y_{\text{mis}})$. If y were fully observed, we would perform inference for the parameter vector θ defined by the data model $p(y \mid \theta)$ and possibly a prior distribution $p(\theta)$. Instead, we must condition on the available information: the observed data y_{obs} and the *inclusion indicator* vector I, which describes which units of y are observed and which are not. (For simple scenarios of missingness, we would label $I_i = 1$ for observed data i and 0 for missing data. More generally, I could have more than two possible values in settings with partially informative missing-data patterns such as censoring, truncation, and rounding.) The model is completed by a probabilistic "inclusion model," $p(I \mid y_{\text{com}}, \phi)$, with a prior distribution $p(\phi \mid \theta)$ on the parameters ϕ of the inclusion model.

Bayesian analysis then works with the joint posterior distribution $p(\theta, \phi, y_{\text{mis}} \mid y_{\text{obs}}, I) \propto p(\theta)\, p(\phi \mid \theta) p(y_{\text{com}} \mid \theta) p(I \mid y_{\text{com}}, \phi)$. It is necessary to formally include I in the model because, in general, the pattern of which data are observed and which are unobserved can be informative about the parameters of interest in the model. In addition, all these probability distributions are implicitly conditional on any fully observed covariates.

An important special case occurs if $p(\theta, y_{\text{mis}} \mid y_{\text{obs}}, I) = p(\theta, y_{\text{mis}} \mid y_{\text{obs}}) \propto p(\theta) p(y_{\text{com}} \mid \theta)$, in which case the inclusion

model is *ignorable* (Rubin, 1976). A key issue in using ignorable models is that they do not require a model $p(I \mid y_{\text{com}}, \phi)$ or a functional form for $p(\phi \mid \theta)$. Two jointly sufficient conditions for ignorability are "missingness at random"—that the probability of the missing-data pattern depends only on observed data—and "distinct parameters"—that ϕ and θ are independent in their prior distribution. In practice, most statistical analyses with missing data either assume ignorability (after including enough covariates in the model so that the assumption is reasonable; for example, including demographic variables in a sample survey and making the assumption that nonresponse depends only on these covariates) or set up specific nonignorable models. As we shall discuss in Section 2.1.3 and in the example in Section 3.1, under an ignorable model one can simulate replications of the completed data y_{com} without ever having to simulate or model the missing-data mechanism.

2.1.2 *Posterior predictive replications in case of missing data.* Replicating the complete data is relatively simple, requiring knowledge only of the complete-data model and parameters, whereas replicating the observed data also requires a model for the missingness mechanism. Thus, from the standpoint of replications, observed datasets—which are characterized by (y_{com}, I)—are more complicated than completed datasets y_{com}.

To simulate replicated datasets for model checking, one can start with the observed data and observed inclusion pattern (y_{obs}, I), then estimate the parameter vector θ simultaneously with the missing data y_{mis}—this is the *data augmentation* paradigm of Dempster et al. (1977) and Tanner and Wong (1987). In simulation-based inference, the result is a set of "multiple imputations" $l = 1, \ldots, L$ of the completed data y^l_{com} along with the corresponding draws of the parameters (θ^l, ϕ^l). The completed datasets can be compared to their expected distribution under the model, or to properties of the reference distribution such as independence, zero mean, or smoothness.

In general, a replicated experiment can lead to a different missing-data pattern, and so the reference distribution for y_{obs} must be determined from the reference distribution of y_{com} along with that of the inclusion pattern I.

2.1.3 *Test variables in the presence of missing data.* In predictive model checking, test variables can be thought of as data displays or summaries, and a key issue is how to construct graphical summaries to reveal important (and often unanticipated) model flaws. This is the problem of exploratory data analysis (Tukey, 1977), here in a modeling context. The best way to understand these choices is to look at practical examples, as we do in Sections 3 and 4. We set up a general notation here.

With missing data, the most general form of a test variable is $T(y_{\text{com}}, I, \theta, \phi)$, the corresponding posterior predictive replication being $T(y^{\text{rep}}_{\text{com}}, I^{\text{rep}}, \theta, \phi)$. Since y_{obs} is a deterministic function of y_{com} and I, this formulation includes observed-data tests as a special case. In general, we imagine replicating y_{com} and possibly replicating I, but the latter only if the test quantity depends on the pattern of missing data. As we discuss here and in the examples, it often makes sense to choose a test variable that depends only on y_{com} and not on I at all.

Although test variables of the form $T(y_{\text{obs}})$ are easier to compute for any given dataset, we would like to consider test variables of the form $T(y_{\text{com}})$, for three reasons. First, the substantive interest typically lies in the complete-data model (what we would do if we observed all the data), so a test variable based on the completed data should be easier to understand substantively. This is important, considering that "practical significance" is as important as "statistical significance" in model checking. Second, as noted at the end of the previous section, the posterior predictive distribution for $y^{\text{rep}}_{\text{com}}$ depends only on the complete-data model (and, of course, the posterior distribution for θ), whereas the posterior predictive distribution for $y^{\text{rep}}_{\text{obs}}$ can also depend on the distribution for the inclusion variable (because the observed units need not be the same in the observed and replicated data). As a result, test statistics of the form $T(y_{\text{com}})$ can be checked using fewer assumptions than are required to test $T(y_{\text{obs}})$. This is particularly important when using ignorable models such as are often assumed in the analysis of observational studies (see Gelman et al., 2003, Section 7.7). Third, in many cases the reference distribution of the replicated test variable, $T(y^{\text{rep}}_{\text{com}})$, has a particularly simple form, often involving independence among variables. As a result, the model can be checked informally using just the simulated realized values, $T(y_{\text{com}})$, with an implicit comparison to a known reference distribution.

2.2 *Latent Data*

Latent data can be defined as the structurally unobserved variables that play a key role in the model for the observed data. Consider observed data y_{obs} that are modeled in terms of latent data y_{lat}, with "completed" dataset $(y_{\text{obs}}, y_{\text{lat}})$. Latent-data problems may be considered as a special case of the general missing-data case, characterized by a structurally modeled inclusion variable I. Bayesian analysis then uses the joint posterior distribution $p(\theta, y_{\text{lat}} \mid y_{\text{obs}}) \propto p(\theta)\, p(y_{\text{lat}} \mid \theta) \times p(y_{\text{obs}} \mid y_{\text{lat}}, \theta)$.

In the latent-data context, I is structurally fixed, and so there are no inclusion-model parameters ϕ. Hence, we then have two main possibilities to define the posterior predictive replications: (a) keeping y_{lat} fixed and varying y_{obs} (i.e., setting $y^{\text{rep}}_{\text{lat}} = y_{\text{lat}}$ and drawing $y^{\text{rep}}_{\text{obs}}$ from $p(y_{\text{obs}} \mid y_{\text{lat}}, \theta)$, and (b) varying both y_{lat} and y_{obs} (i.e., drawing $(y^{\text{rep}}_{\text{lat}}, y^{\text{rep}}_{\text{obs}})$ from $p(y_{\text{lat}}, y_{\text{obs}} \mid \theta) = p(y_{\text{lat}} \mid \theta)\, p(y_{\text{obs}} \mid y_{\text{lat}}, \theta)$).

In the latent-data context the most general form of a test summary is $T(y_{\text{obs}}, y_{\text{lat}}, \theta)$, the corresponding posterior predictive replication being $T(y^{\text{rep}}_{\text{obs}}, y^{\text{rep}}_{\text{lat}}, \theta)$. This formulation includes observed-data test summaries as a special case. In general, we recommend examining the test summaries that check in a natural way, key features of the model under consideration. In many latent-data models such summaries will depend on y_{lat} as well as y_{obs}.

Many datasets fit with latent-data models also have missing data. One can then put the inclusion indicators I into the model and proceed as in Section 2.1, with the additional feature that latent data are present. Test variables can be defined from the completed observable data y_{com}, which includes the imputations of the missing and latent data.

2.3 *Implementation*

The most general implementation of the completed-data model checks proceeds in three steps:

1. Perform inference jointly for the parameters θ (and, if necessary, ϕ) and the missing and latent data y_{mis}, y_{lat}, thus obtaining a set of L imputed datasets y_{com}. Inference for the model parameters can be represented by a point estimate or, more generally, by L draws from the posterior distribution.
2. Construct a test variable—in the context of this article, often a graph—that is a function of the completed data, y_{com} and possibly the inclusion indicators I and the parameters θ. The L imputations induce L possibilities for the test variable, and these can be displayed as multiple imputations (as in the second or third row of Figure 6).
3. Construct the reference distribution of the test variable, which can be done analytically (as with some χ^2 tests), or using the complete-data model given a point estimate of the parameters, or given posterior simulations of the parameters, or using other approaches such as cross-validation or bootstrapping to summarize inferential uncertainties. In any case, the result is a distribution, or a set of simulated replications, of the test variable assuming the fitted model. Depending on the details of the problem, the replications can be displayed graphically, for example as in the overlain lines in Figures 2 and 3.

For the observed-data model checks—or more generally, for any test variables that depend on I as well as y_{com}—the third step requires replication of the inclusion indicators as well as the complete data, as discussed in Section 2.1.3.

In practice, it is often convenient to simplify step 2 above. Datasets typically have internal replication, and often a single random imputation conveys the look of a graphical test variable, without the need for displaying several random draws. The bottom row of plots in Figure 4, for example, displays a single completed dataset. For simplicity, we often work with a single imputation if the data have enough structure. A related strategy is to create the diagnostic plot several times and, if the multiply imputed completed datasets look similar, to display just one of them. When summarizing with a numerical test statistic, one can use the entire distribution to compute p-values, as we illustrate in Section 4.1.

We can often simplify step 3—the computation and display of the reference distribution—by comparing the graphical test variable to an implicit reference distribution. For example, residual plots are compared to the null hypothesis of zero mean and independence. (In a latent-data posterior predictive framework, unlike with point estimation, residuals are independent in their reference distribution.) We shall illustrate less structured implicit comparisons in Figures 4 and 6.

3. Applications with Missing Data

3.1 *Randomized Experiments with an Ignorable Dropout Model*

A common problem in studies of persons or animals is that subjects drop out in the middle of the experiment, creating a problem of missing data. After imputation, we can use the completed-data methods to check model fit, as we illustrate here.

Table 1
Summary of the number of observations taken at each occasion for the rat example, for each group separately and in total

	Number of observations			
Age (days)	Control	Low dose	High dose	Total
50	15	18	17	50
60	13	17	16	46
70	13	15	15	43
80	10	15	13	38
90	7	12	10	29
100	4	10	10	24
110	4	8	10	22

Verbeke and Lesaffre (1999) analyzed the longitudinal data from a randomized experiment, the aim of which was to study the effect of inhibiting testosterone production on the craniofacial growth of male Wistar rats. A total of 50 rats were randomly assigned to either a control group or one of the two treatment groups where treatment consisted of a low or high dose of the drug Decapeptyl, which is an inhibitor for testosterone production in rats. The treatment started at the age of 45 days, and measurements were taken at 50 days and every 10 days thereafter. The responses of interest are distances (in pixels) between well-defined points on X-ray pictures of the skull of each rat, taken after the rat has been anesthetized. See Verdonck et al. (1998) for a detailed description of the experiment.

For the purpose of this article, we consider one of the measurements that can be used to characterize the height of the skull. The individual profiles are shown in Figure 1 and show a high degree of dropout. Indeed, many rats do not survive anaesthesia and therefore drop out before the end of the experiment. Table 1 shows the number of rats observed at each occasion. Of the 50 rats randomized at the start of the experiment, only 22 survived all seven measurements. Verbeke and Lesaffre (1999) studied the effect of the dropout on the efficiency of the final testing procedures, and derived alternative designs with less risk of huge losses of efficiency when dropout would occur. They modeled the jth measurement y_{ij} for the ith rat, $j = 1, \ldots, n_i$, $i = 1, \ldots, N$, as

$$y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + b_i + \varepsilon_{ij}, & \text{if low dose} \\ \beta_0 + \beta_2 t_{ij} + b_i + \varepsilon_{ij}, & \text{if high dose} \\ \beta_0 + \beta_3 t_{ij} + b_i + \varepsilon_{ij}, & \text{if control dose,} \end{cases} \tag{1}$$

where the transformation $t_{ij} = \log(1 + (\text{Age}_{ij} - 45)/10)$ is used to linearize the subject-specific profiles. The parameter β_0 then represents the average response at the time of treatment, and β_1, β_2, and β_3 represent the average slopes for the low dose, high dose, and control groups. The assumption of a common average intercept is justified by the randomization of the rats. For each subject i, the parameter b_i fits the deviation of its intercept from the average value in the population, and the ε_{ij}'s denote the residual components; it is assumed that they all are independently and normally distributed with mean 0 and standard deviations σ_b and σ_ϵ, respectively.

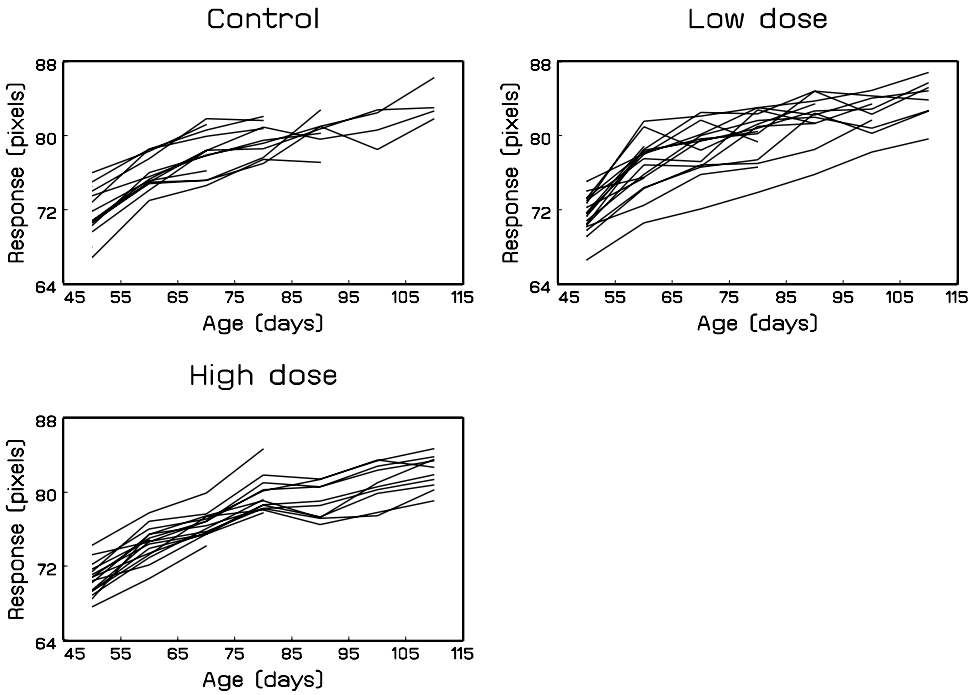

Figure 1. Individual profiles for rats in each of the three treatment groups separately, for the ignorable dropout example in Section 3.1.

Inspection of Figure 1 suggests a specific model violation: in the high-dose condition, the residual variance seems to be smaller than in the two other conditions, at least before age 75 days. Such a result could be interesting in understanding the effects of the treatment. However, it is hard to interpret this graph because, even under an ignorable model, dropout can depend on previous measurements. For example, a lack of extreme measurements at high-time values could be explained by dropout rather than by underlying data.

The assumption of ignorability can, by definition, never be formally checked without making strong assumptions about possible associations between dropout and the missing outcomes, and it is important to study the sensitivity of the conclusions to the underlying assumptions. This dataset has previously been extensively analyzed (see Verbeke et al., 2001), and based on conversations with the clinicians involved in this experiment, there seems to be no clinical evidence that missingness might depend on unobserved outcomes.

In this example, we can safely assume ignorability of the inclusion mechanism; we therefore use (1) to impute the missing data (based on mixed-model estimates for the parameters b_i). Next we calculate, for each age, the standard deviation across rats of the y_{com} values. This standard deviation captures both the between-rat variance in intercept b_i and the residual variance σ_ϵ. (Because we calculate the test summary separately for each simulation draw, this standard deviation is *not* inflated by estimation uncertainty in the posterior distribution.)

The results of a single randomly imputed completed dataset—the observed data supplemented with a random draw of the missing data from the posterior distribution—appear in Figure 2, along with the standard deviations of 20 replicated datasets (again based on the mixed-model estimates). This figure supports the impression that the residual variance in the high-dose condition is somewhat smaller than assumed under the model whereas the reverse seems to hold for the low-dose condition. The pattern is suggestive but not statistically significant, in that the replications show that such a pattern is possible under the model.

This finding inspired us to try out a model expansion with the condition-dependent residual variances σ_{ϵ_1}, σ_{ϵ_2}, and σ_{ϵ_3}. Such a model expansion can be justified on substantive grounds as it formalizes dose-dependent irregularities in growth speed. A likelihood ratio statistic revealed that the expanded model tends to be preferable over the original model (1), LR = 5.4, df = 2, $p = 0.067$, whereas for the expanded model AIC = 943.0 and for model (1) AIC = 944.4.

Figure 3 checks the expanded model for the replication of the 20 datasets as well as for the imputation of the missing outcomes. When compared to Figure 2, the completed standard deviation lines are clearly more in the center of the reference distribution of replicated data, especially toward the end of the study (where most of the missingness occurs). Note also the much smaller completed standard deviation at 60 days in the control group (compared to Figure 2), even though an imputation is needed at that time point for two rats only. However, one of these rats had an exceptionally small initial value (at 50 days). The imputation is now based on a smaller residual variance, hence a larger within-subject correlation, implying that the imputed value at age 60 days for this rat will tend to be smaller as well. Finally, the control group is also the smallest, containing only 15 rats.

These results show that our graphical approach to checking fit is useful in that it helps in finding out relevant directions

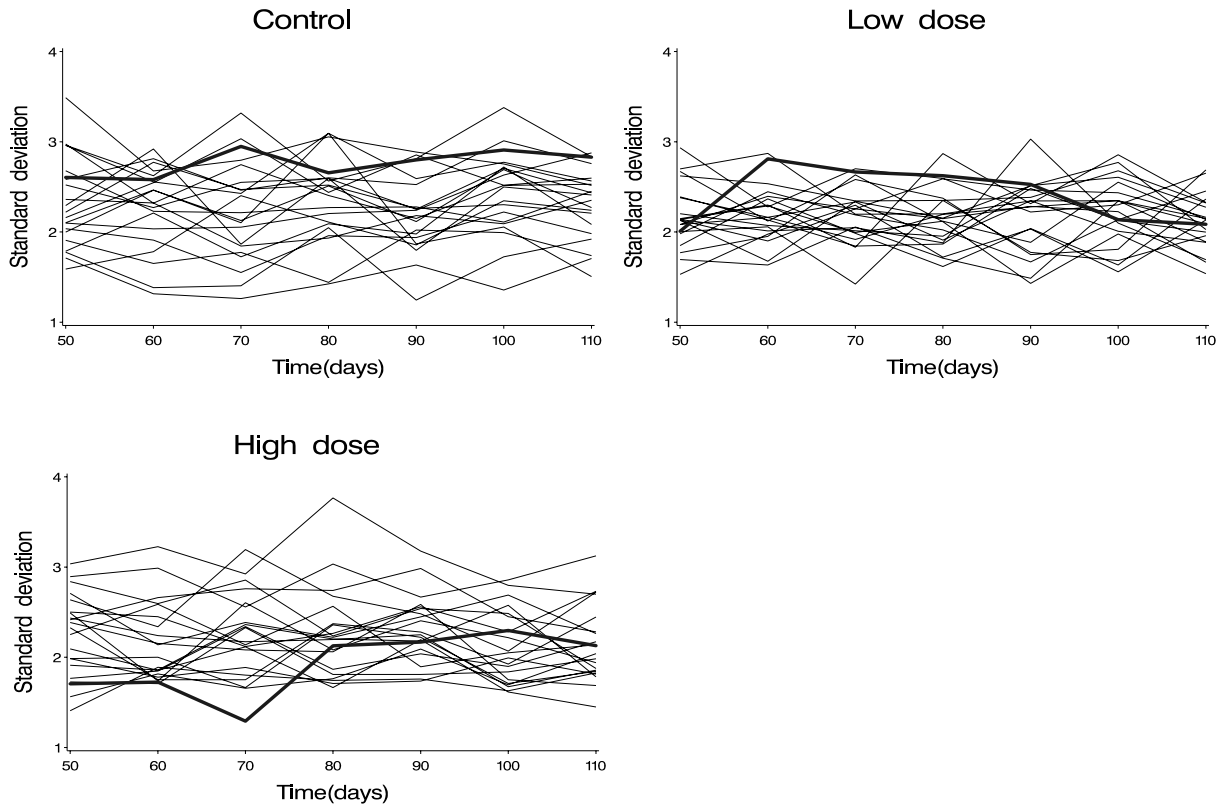

Figure 2. Standard deviations from completed dataset (in bold) compared to the standard deviations from 20 replicated datasets (assuming equal variances for the three groups), plotted for each treatment group separately, for the ignorable dropout example in Section 3.1.

for specifying alternative models. If desired, candidate models that are generated in this way can be compared using numerical criteria (e.g., AIC). In the rat example, in this way, a potentially meaningful model improvement was obtained, suggested by the results of the graphical check.

3.2 *Clinical Trials with Nonignorable Dropout*

The previous example illustrated the common setting in which missing data are imputed using an ignorable model. In other settings, however, dropout is affected by outcomes under study that have not been fully recorded, and so it often makes sense to use nonignorable models (for example, in a study of pain-relief drugs, a subject may drop out if he or she continues to feel the pain). As a result, the analysis cannot simply be done on the observed data alone (Diggle and Kenward, 1994). Methods based on the Bayesian modeling of dropouts can be thought of as multiple imputation approaches in which (a) the measurements that would have occurred are imputed, and then (b) a completed-data analysis is performed. A key intermediate stage here is the completed dataset, which we can plot to see whether any strange patterns appear. We illustrate with an example from Sheiner, Beal, and Dunne (1997).

The top row of Figure 4 shows the distribution of recorded pain measurements over time for patients who were randomly assigned to be given one of three doses of a new pain-relief drug immediately following a dental operation. In this top row of plots, the width of the bar at each time represents the proportion of participants still in the trial. Patients were allowed to drop out at any time by requesting to be switched to a pain reliever that is known to be effective. The data show heavy dropout, especially among the controls. In addition, there seems to be a pattern of decreasing pain over time at all doses—but it is not clear how this is affected by the dropout process.

Sheiner et al. (1997) fit to these data a model with three parts. Internally for each subject is a pharmacokinetic differential equation model of the time course of the concentration of the drug in different compartments of the body. This model implicitly includes an impulse–response function of internal concentrations to administered doses of the drug. At the next level, the pain-relief data were fit by an ordered multinomial logistic model with probabilities determined by a nonlinear function of the internal concentration of the drug. Finally, missingness was modeled nonignorably, with the probability of dropping out depending on the pain level at the time (which is unobserved under dropout).

Once this model has been fit to data, it can be used to make predictions under alternative input conditions, as demonstrated by Sheiner et al. (1997), who determined a more effective dosing regimen that is estimated to give a consistently high level of pain relief with a low total dose. In addition, the model yields estimated uncertainty distributions for the underlying full time series of pain scores that would have occurred for each patient in the absence of dropout. We show here (following Gelman and Bois, 1997) how these imputed

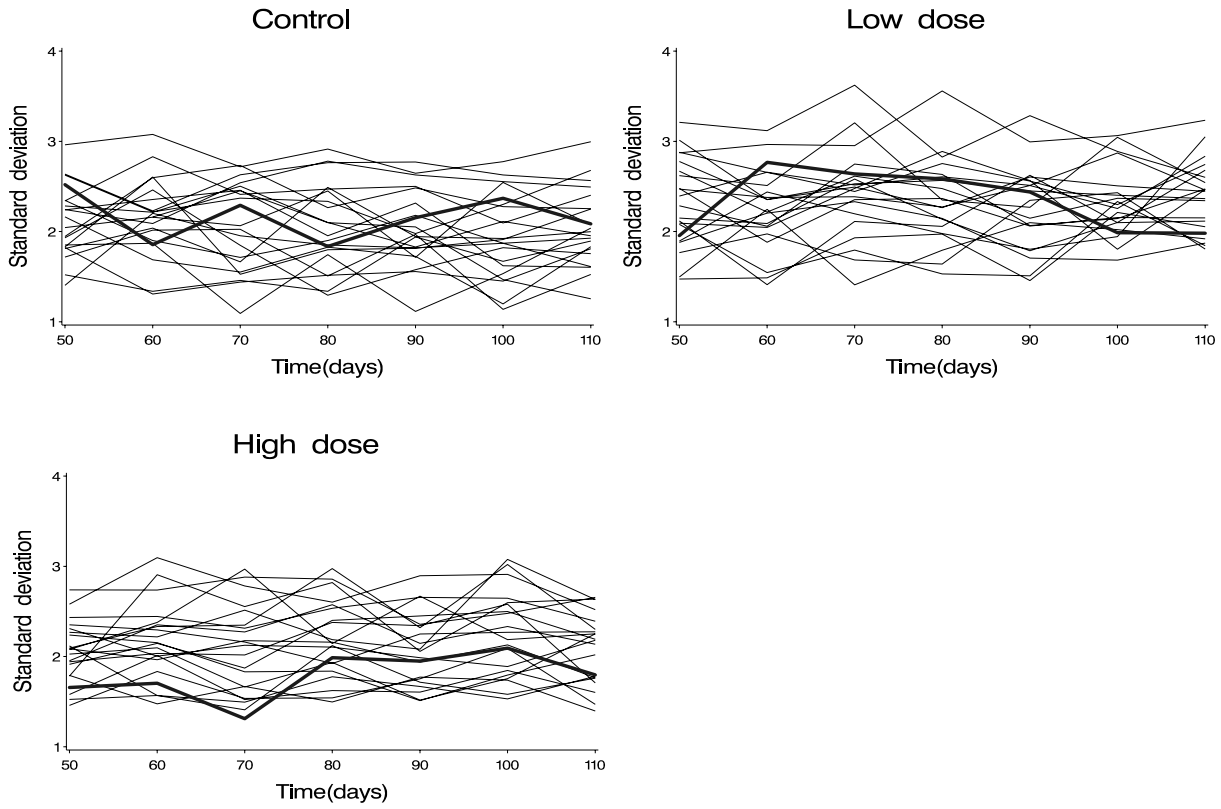

Figure 3. Predictive checks for the expanded model with group-specific residual variances. Compare to the checks for the simpler model in Figure 2.

pain scores can be used to summarize the estimated underlying patterns in the data.

The bottom row of Figure 4 shows the graphs similar to the top row, but of the completed dataset with imputations for the dropouts. (Here, a simple deterministic scheme was used for the imputations, but the method could be used with multiple imputations, leading to several sets of graphs corresponding to the different imputations.) For all doses, the completed data show immediate pain relief followed by some increasing pain. These plots show the dose–response relation far more clearly than did the observed-data plots in the top row.

Plotting the completed dataset is interesting here even if it does not reveal model flaws: the completed dataset is much easier to understand and interpret than the plot of observed data alone, and substantive hypotheses are more directly interpretable in terms of the completed data. These plots can be seen as a model check, not compared to a posterior predictive distribution but rather to whatever substantive knowledge is available about pain relief.

4. Applications with Latent Data

4.1 *Latent Psychiatric Classifications*

Psychiatric symptom judgments of patients by psychiatrists and clinical psychologists may be based on implicit classifications of the patients by the clinicians in some implicit syndrome taxonomy that is shared by the clinicians (Van Mechelen and De Boeck, 1989). According to a clinician, a symptom then will be present in some patient if there is at least one implicit syndrome that applies to that patient and that implies the symptom in question. Maris, De Boeck, and Van Mechelen (1996) have formalized this idea in a model that includes probabilistic links between symptoms and latent syndromes on the one hand, and between patients and latent syndromes on the other hand. In particular, let $(y_{\text{obs}})_{ijk}$ equal 1 if patient i has symptom j according to clinician k, and $(y_{\text{obs}})_{ijk}$ equal 0 otherwise. The assumed model then implies latent variables for the patients $(y_{\text{lat,p}})_{ijkl}$ and latent variables for the symptoms $(y_{\text{lat,s}})_{ijkl}$, each pertaining to $l = 1, \ldots, L$ latent syndromes

$$(y_{\text{lat,p}})_{ijkl} = \begin{cases} 1 & \text{if, when patient } i \text{ is judged on symptom} \\ & j \text{ by clinician } k, \text{this patient is considered} \\ & \text{to suffer from latent syndrome } l \\ 0 & \text{otherwise,} \end{cases}$$

$$(y_{\text{lat,s}})_{ijkl} = \begin{cases} 1 & \text{if, when patient } i \text{ is judged on symptom} \\ & j \text{ by clinician } k, \text{this symptom is considered} \\ & \text{to be implied by latent syndrome } l \\ 0 & \text{otherwise.} \end{cases}$$

The model further assumes that

$$(y_{\text{lat,p}})_{ijkl} \sim \text{Bern}(\theta_{p,il}), \quad (y_{\text{lat,s}})_{ijkl} \sim \text{Bern}(\theta_{s,jl}),$$

all latent variables being independent. As stated above, clinician k will then judge symptom j to be present in patient i if there is at least one syndrome l for which (a) patient i is

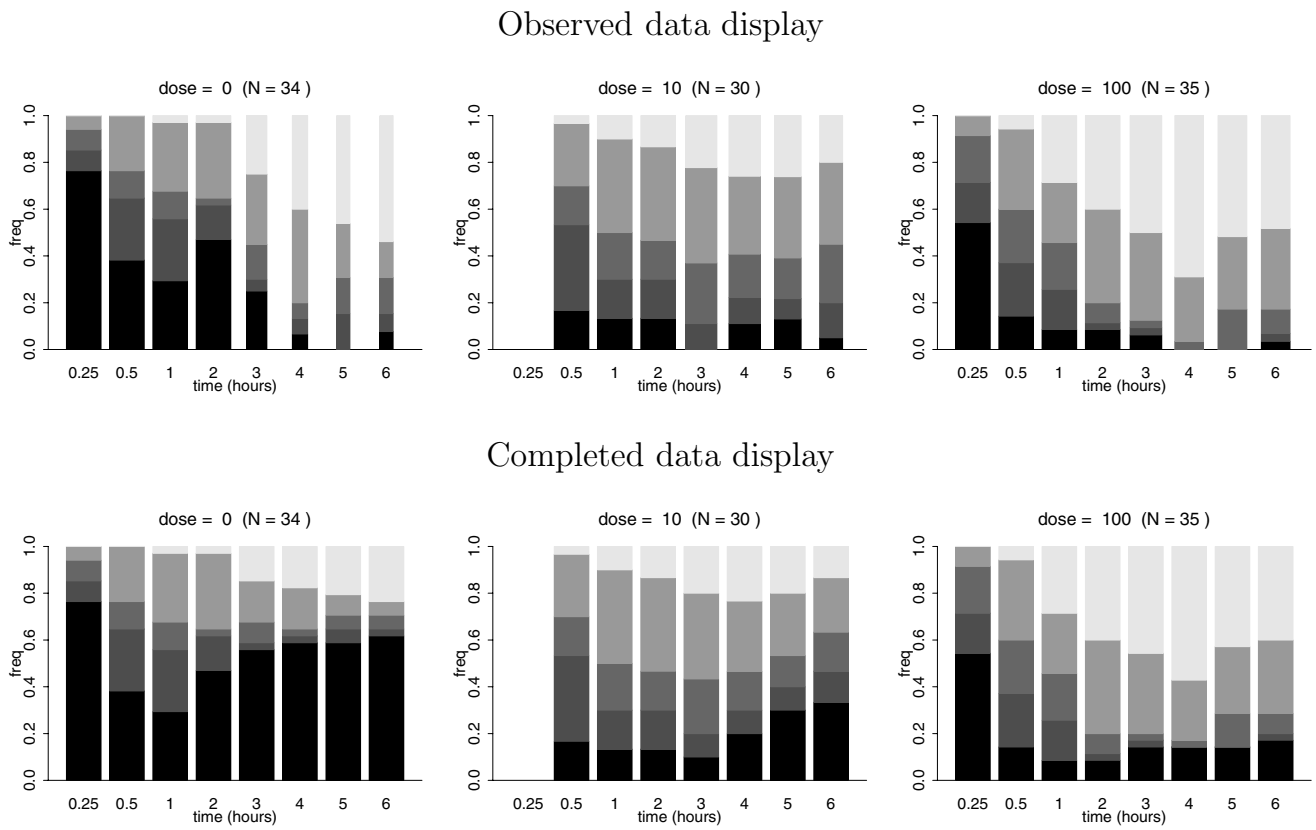

Figure 4. Summary of pain-relief responses over time under different doses from the clinical trial with nonignorable dropout discussed in Section 3.2. In each summary bar, the shadings from bottom to top indicate "no pain relief" and intermediate levels up to "complete pain relief." The graphs in the top row include only the persons who have not dropped out (with the width of the histogram bars proportional to the number of subjects remaining at each time point). The graphs in the bottom row include all persons, with imputed responses for the dropouts. As discussed in Section 3.2, the bottom row of plots—which are based on completed datasets—are much more directly interpretable than the observed-data plots on the top row. From Sheiner, Beal, and Dunne (1997) and Gelman and Bois (1997).

judged by clinician k to suffer from it, *and* (b) symptom j is judged by clinician k to be implied by it. Stated formally

$$(y_{\text{obs}})_{ijk} = 1 \quad \text{if there exists an } l \text{ for which}$$
$$(y_{\text{lat,p}})_{ijkl} = 1 \quad \textit{and} \quad (y_{\text{lat,s}})_{ijkl} = 1.$$

When fitting the model to symptom judgments of patients by several clinicians, the model assumptions could be violated if there are systematic differences between clinicians in the links between symptoms and latent syndromes. Natural test variables to check this assumption can be defined making use of the latent Bernoulli variables $y_{\text{lat,s}}$.

We illustrate with data from Van Mechelen and De Boeck (1990) on 23 psychiatric symptom judgments for 30 patients by 15 clinicians. As test variables we calculate, for each symptom j and for each latent syndrome l, the variance across clinicians of the summed realizations of the corresponding symptom–syndrome link variable:

$$T_{jl} = \frac{1}{K}\sum_k \left[\sum_i (y_{\text{lat,s}})_{ijkl}\right]^2 - \left[\frac{1}{K}\sum_k \sum_i (y_{\text{lat,s}})_{ijkl}\right]^2.$$

We further summarize the fit for each syndrome and symptom by posterior predictive p-values; for any test variable $T_{jl}(y_{\text{lat,s}})$, the p-value is $\Pr(T_{jl}(y^{\text{rep}}_{\text{lat,s}}) > T_{jl}(y_{\text{lat,s}}))$, and can be computed using the set of M multiple imputations of the parameters and completed dataset (Meng, 1994b; Gelman et al., 1996). Figure 5a shows the histogram of the posterior predictive p-values for the between-clinician variance for the link between the first syndrome and each of 23 symptoms, and Figure 5b shows the corresponding histogram for the third latent syndrome (which could be identified as an implicit schizophrenia syndrome). For the third, unlike for the first, latent syndrome, the variation in several symptom–syndrome links across clinicians is greater in the data than assumed under the model. This can be further clarified by plots such as Figure 5c, which shows a plot of 2000 pairwise comparisons of $T_{jl}(y^{\text{rep}}_{\text{lat,s}})$ and $T_{jl}(y_{\text{lat,s}})$ for the symptom "inappropriate affect" and the latent schizophrenia diagnosis. This example illustrates how model checks can be formed using latent data only.

4.2 *Rounded and Heaped Data*

We next illustrate with an example of imputed continuous latent data. Heitjan and Rubin (1990) analyzed a survey of

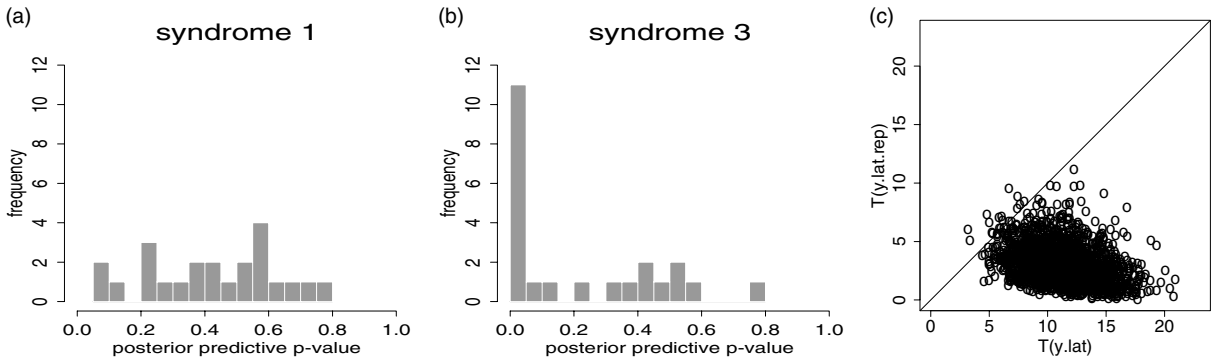

Figure 5. (a, b) Histograms of 23 posterior predictive p-values for (a) first and (b) third latent syndrome in the latent psychiatric classification example. The relevant test variables are the variance across clinicians of the summed symptom–latent syndrome links. The extreme p-values in the histogram on the right correspond to symptoms for which there is significantly more variation than expected by the data, based on the completed-data classification into the third latent syndrome. (c) Plot of 2000 comparisons of $T(y_{\text{lat}}^{\text{rep}})$ versus $T(y_{\text{lat}})$ for the symptom "inappropriate affect" and the latent schizophrenia diagnosis in this example. Data from Van Mechelen and De Boeck (1990).

children's health in Africa in which the exact ages of the children are not known—only "reported ages" given by the parents were available, along with the anthropometric data including height and weight. The original purpose of the survey was to combine height and weight with recorded age to produce tables classifying the children by nutritional status: Being thin for one's height suggests current malnutrition, and being short for one's age suggests a history of malnutrition. Standard curves for these variables are based on data from the United States, where children's ages are typically known with great accuracy.

The top histogram in Figure 6 shows the reported ages for the children in the sample. A striking feature of the data is that many of the ages were evidently reported as truncated or rounded (to the nearest 12 months, for example). Thus there is serious concern that many, perhaps most, of the ages are imprecise, and that using reported ages as the truth may lead to wholesale misclassification of the nutritional status.

Because the level of coarsening evidently depends on the unobserved true age, Heitjan and Rubin modeled the age reporting using nonignorable models, considering two approaches identified with implicit and explicit models of the age-reporting process. The implicit model took ages divisible by 12 months and randomly imputed them uniformly in the interval of the reported age ±6 months; and took ages divisible by 6 but not by 12, randomly imputing them ±3 months. The notion is that if reported age equals a full year, it is because the subject rounded to the nearest year, and if reported age equals a half year, it is because the subject rounded to the nearest half year. (Such a model would not be valid for coarsened age data from the United States, where the practice is generally to truncate age to the next lowest year or half year; in Africa, rounding is thought to be a more plausible model.) In a second class of explicit models, the authors predicted age from available anthropometric variables, assuming constraints on the age consistent with a process of rounding to either the nearest year, half year, or month, with the probability of each of these rounding procedures estimated from data (see Heitjan and Rubin, 1990, for details). In these models it was judged that a linear model for age on the square root scale was reasonable.

To assess fit, one can examine the histograms of imputed ages. A model that inappropriately corrects for the reporting process will yield implausible histograms of exact ages. For example, the top histogram in Figure 6, the reported data, can be viewed as an imputation of the set of exact ages under a very simple model of zero-reporting error. The middle row of histograms in Figure 6 shows three draws of the imputed exact ages under the implicit uniform model, and the bottom row in Figure 6 displays the histograms of the imputed ages under the linear prediction model.

Because the problem with these data is judged to be "heaping" at years and half years, one might wish to base diagnostics on the fractions of subjects whose ages are divisible by 6. If the age distribution is roughly uniform, this fraction should be around 1/6. In the original data, the fraction of subjects with reported ages divisible by 6 is 83%, with 95% confidence interval [79%, 88%]; clearly these data do not fit such a model. There was concern that with the uniform model (the middle row of Figure 6), and perhaps to a lesser extent with the linear model (the bottom row of Figure 6), there could be a tendency to smooth too many subjects away from multiples of 6 months. A certain amount of this behavior is evident in the histograms. Nevertheless the effect is not strong, as confidence intervals for the fraction of ages divisible by 6 (based on five imputations) are [8.3%, 23.4%] for the uniform model and [9.0%, 21.0%] for the linear model, both comfortably covering the null value of 16.7%.

The actual modeling and imputation procedures used to fit the linear prediction models are quite complex, involving a nonignorable selection model for determining the level of coarseness used in rounding the data and accounting for the fact that a subject with reported age divisible by 12 may have rounded to the nearest year, or half year, or even to the nearest month. Despite the rather arduous modeling that

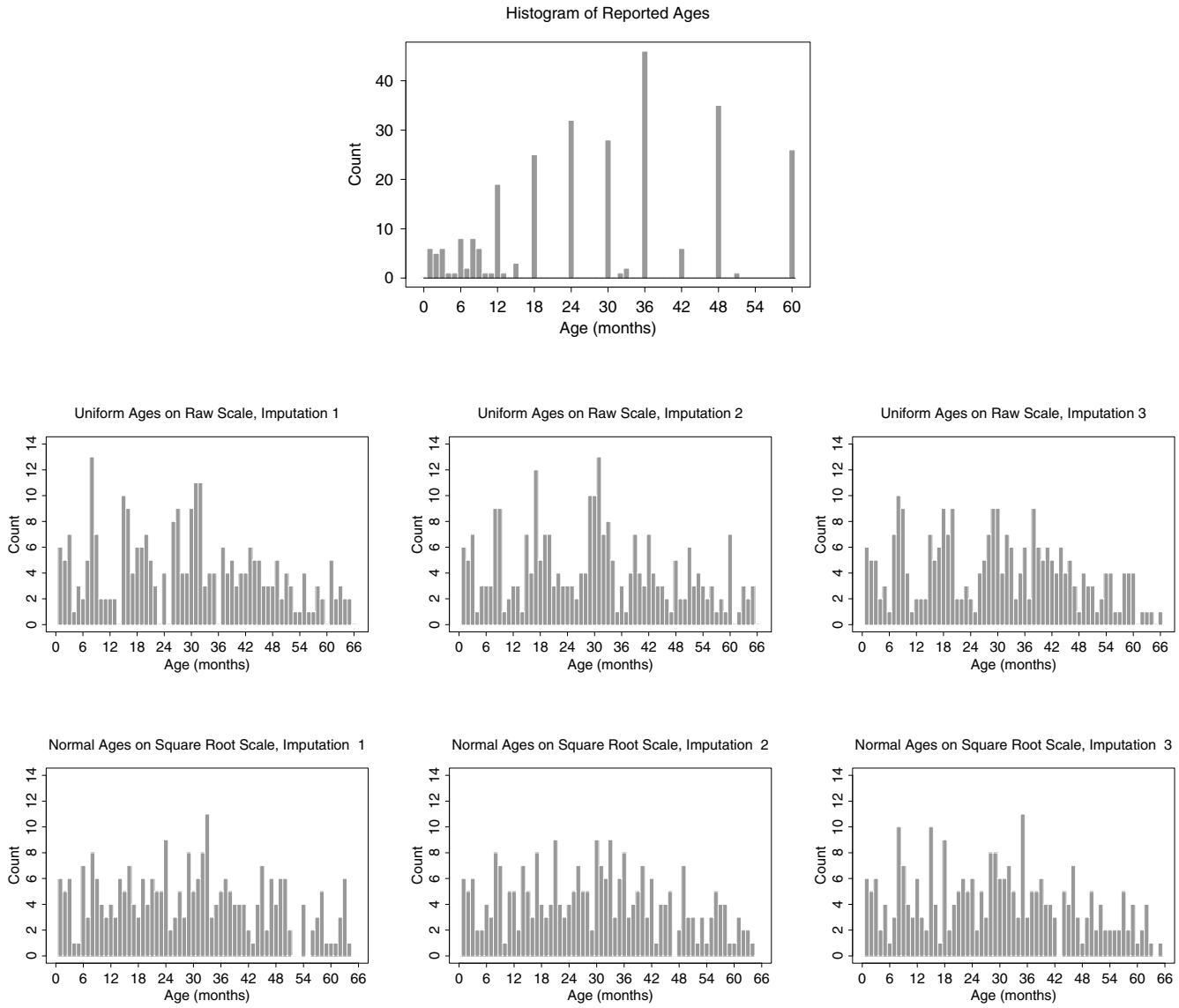

Figure 6. Top row: Histogram of the distribution of recorded ages for a sample of children, from the heaped-data example of Section 4.2. The uneven look of the histogram is presumably due to rounding of reported ages. Middle and bottom rows: Histograms of three draws from the posterior distribution of estimated true ages under each of two candidate imputation methods. The comparison to the posterior predictive distribution (a sample from a smooth distribution of ages) is implicit. Adapted from Heitjan and Rubin (1990).

was done in this example, we were able to check the fit of the models quite easily using completed-data replications.

5. Discussion

5.1 *Potential for Integrating Missing-Data Imputation and Diagnostics into Fully Model-Based Inference*

Statistics is moving toward more elaborate analyses of more complicated data structures, which inevitably feature missing and latent data. As our models become more complicated, it is important to develop methods to check their fit. A general feature of our approach is the separation of the data analysis into two steps: (1) model fitting (including the creation of imputations for the missing and latent data), and (2) model checking using the complete data (and possibly also the observed inclusion pattern). The test summaries used as model checks need not refer to the missing-data structure at all. This is similar to the multiple imputation context in which the data analyst need not be knowledgeable about the missing-data model (see Rubin, 1987; Meng, 1994a).

The main idea of this article—defining reference distributions based on multiply imputed completed datasets—is applicable not only to posterior predictive tests but also to other methods of Bayesian model checking and sensitivity analysis, such as model averaging, model expansion, and crossvalidation (see Gelfand, Dey, and Chang, 1992; Draper, 1995; Kass and Raftery, 1995). We also recall the distinctions between "practical" and "statistical" significance: A model may be useful even if it clearly does not fit some aspect of the

data (as indicated, for example, by a posterior predictive p-value) and, conversely, fit of the model in one aspect does not guarantee that it is acceptable for other purposes.

5.2 *Distinct Advantages of the Proposed Approach*

The benefits from the approach described in this article show up in three ways. First, the proposed approach yields diagnostics that are easily interpretable. For example, Figure 4 shows how a simple summary display of completed data (bottom row of plots) is much easier to interpret than the raw data (top row of plots) for the purpose of understanding of the time patterns of pain relief and, of comparing to any implicit hypotheses about these patterns. For another example, the time plots in Figures 2 and 3 would be more difficult to interpret outside the completed-data imputation framework.

Second, the proposed approach enables one to account for uncertainty in a way that allows important model checks to be performed visually. For example, in the plots in Figures 2 and 3, each of the thin lines summarizes inference for a single random imputation of the completed data, with the spread among the lines indicating inferential uncertainty. Predictive uncertainty can also be summarized using p-values, as in Figure 5.

Third, the proposed completed-data diagnostics give us a better theoretical understanding of the potential and limitations of our modeling assumptions. For example, Figure 6 compares completed data to implicit assumptions of smoothness of the underlying age distribution. In the spirit of exploratory data analysis, this test can be performed visually without requiring an explicit model for the smoothness.

We conclude by noting that, once a model has been fitted and multiple imputations have been created, the computations of completed-data model checking are typically straightforward—requiring direct simulation and graphical display but not heavy computations such as integrations or Markov chain simulations. Completed-data diagnostic displays avoid the data-collection artifacts that are common with observed-data plots (see, for example, Figure 4), and we have found them helpful in understanding models and data in a variety of examples.

Acknowledgements

We thank Phillip N. Price and several reviewers for helpful comments. This work was supported in part by fellowship F/96/9 and research grant OT/96/10 of the Research Council of Katholieke Universiteit Leuven, grant IAP P5/24 of the Belgian Federal Government, and grants SBR-9708424, SES-9987748, and SES-0318115 and Young Investigator Award DMS-9796129 of the U.S. National Science Foundation.

Résumé

Dans les problèmes avec données manquantes ou latentes, une approche standard consiste à imputer les données non observées, puis à effectuer toutes les analyses statistiques sur le jeu de données complétées—correspondant aux données observées et aux données non observées imputées—en utilisant les procédures standard pour l'inférence sur données complètes. Ici nous étendons cette approche à la vérification de modèle en montrant les avantages qu'il y a à utiliser des diagnostics de modèles à données complétées sur des jeux de données complétées par imputation. Cette approche se place dans le cadre théorique de vérification bayésienne prédictive a posteriori (mais, de même qu'avec l'attribution des données manquantes, nos méthodes de vérification de modèles pour données manquantes peuvent aussi être interprétées comme de l'inférence prédictive dans un contexte non bayésien). Nous envisageons des diagnostics graphiques dans ce cadre.

On peut citer pour l'approche par données complétées les avantages suivants : 1) On peut souvent vérifier l'ajustement au modèle en considérant de manière naturelle des quantités-clé qui ont un intérêt de fond, ce qui n'est pas toujours possible avec les données observées seules. 2) Dans les problèmes avec données manquantes, on peut imaginer des vérifications qui ne supposent pas de modéliser le manque ou le mécanisme d'inclusion; ce dernier aspect est utile pour l'analyse des mécanismes de collecte de données ignorables mais inconnus, tels qu'on les suppose souvent dans l'analyse des enquêtes par échantillonnage ou des études observationnelles. 3) Dans de nombreux problèmes avec données latentes, il est possible de vérifier des aspects qualitatifs du modèle (par exemple l'indépendance de deux variables) qui peuvent être formalisés de façon naturelle à l'aide de variables latentes. Nous illustrons cela au moyen de plusieurs exemples d'application.

References

Albert, J. and Chib, S. (1995). Bayesian residual analysis for binary regression models. *Biometrika* **82,** 747–759.

Buja, A., Asimov, D., Hurley, C., and McDonald, J. A. (1988). Elements of a viewing pipeline for data analysis. In *Dynamic Graphics for Statistics*, W. S. Cleveland and M. E. McGill (eds), 277–308. Belmont, California: Wadsworth.

Buja, A., Cook, D., and Swayne, D. (1999). Inference for data visualization. Talk given at Joint Statistical Meetings. `www.research.att.com/~andreas/#dataviz`

Chaloner, K. (1991). Bayesian residual analysis in the presence of censoring. *Biometrika* **78,** 637–644.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75,** 651–659.

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Society A* **147,** 278–292.

Dawid, A. P. (1992). Prequential data analysis. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, M. Ghosh and P. K. Pathak (eds), 113–126. Hayward, California: Institute of Mathematical Statistics.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39,** 1–38.

Diggle, P. and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis. *Applied Statistics* **43,** 49–93.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B* **57,** 45–97.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions

with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4,* J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds), 147–167. New York: Oxford University Press.

Gelman, A. (2004). Exploratory data analysis for complex models (with discussion). *Journal of Computational and Graphical Statistics* **13,** 755–787.

Gelman, A. and Bois, F. Y. (1997). Discussion of "Analysis of non-randomly censored ordered categorical longitudinal data from analgesic trials," by L. B. Sheiner, S. L. Beal, and A. Dunne. *Journal of the American Statistical Association* **92,** 1248–1250.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd edition. London: CRC Press.

Gelman, A., King, G., and Liu, C. (1998). Multiple imputation for multiple surveys (with discussion). *Journal of the American Statistical Association* **93,** 846–874.

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.

Heitjan, D. F. and Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association* **85,** 304–314.

Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). *Journal of the Royal Statistical Society B* **60,** 497–536.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association* **90,** 773–795.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis* **48,** 198–206.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association* **79,** 393–398.

Maris, E., De Boeck, P., and Van Mechelen, I. (1996). Probability matrix decomposition models. *Psychometrika* **61,** 7–29.

Meng, X. L. (1994a). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* **9,** 538–573.

Meng, X. L. (1994b). Posterior predictive p-values. *Annals of Statistics* **22,** 1142–1160.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91,** 473–489.

Seillier-Moiseiwitsch, F., Sweeting, T. J., and Dawid, A. P. (1992). Prequential tests of model fit. *Scandinavian Journal of Statistics* **19,** 45–60.

Sheiner, L. B., Beal, S. L., and Dunne, A. (1997). Analysis of non-randomly censored ordered categorical longitudinal data from analgesic trials (with discussion). *Journal of the American Statistical Association* **92,** 1235–1255.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82,** 528–540.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Van Mechelen, I. and De Boeck, P. (1989). Implicit taxonomy in psychiatric diagnosis. *Journal of Social and Clinical Psychology* **8,** 276–287.

Van Mechelen, I. and De Boeck, P. (1990). Projection of a binary criterion into a model of hierarchical classes. *Psychometrika* **55,** 677–694.

Verbeke, G. and Lesaffre, E. (1999). The effect of drop-out on the efficiency of longitudinal experiments. *Applied Statistics* **48,** 363–375.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001). Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics* **57,** 7–14.

Verdonck, A., De Ridder, L., Verbeke, G., Bourguignon, J. P., Carels, C., Kuhn, R., Darras, V., and de Zegher, F. (1998). Comparative effects of neonatal and prepubertal castration on craniofacial growth in rats. *Archives of Oral Biology* **43,** 861–871.

Received October 2003. *Revised April* 2004.
Accepted April 2004.