Methods for Using Data Abstracted from Medical Charts to Impute Longitudinal Missing Data in a Clinical Trial

Paul L. Hebert, PhD1,*†, Leslie T. Taylor, PhD1, Jason J. Wang, PhD2, Margo A. Bergman, PhD1

1Northwest Center for Outcomes Research in Older Adults, VA Puget Sound Health Care System, Department of Veterans Affairs, Seattle, WA, USA; 2Department of Health and Mental Hygiene, Primary Care Information Project, New York City, New York, NY, USA

ABSTRACT

Objective: To describe a method for imputing missing follow-up blood pressure data in a clinical hypertension trial using blood pressures abstracted from medical charts. Methods: We tested a two-step method. In the first, a longitudinal mixed-effects model was estimated on blood pressures abstracted from medical charts. In the second, the patient-specific fitted values from this model at follow-up were used to impute blood pressures missing at follow-up in the trial. Simulations that imposed alternative missing data mechanisms on observed trial data were used to compare this approach to imputation approaches that do not incorporate data from charts. Results: For data that are missing at random, incorporating the fitted values from chart-based longitudinal models leads to estimates of the trial-based blood pressures that are unbiased and have lower mean squared deviation than do blood pressures imputed without the chart-based data. For data that are missing not at random, incorporating fitted values ameliorates but does not eliminate the inherent missing data bias. Conclusions: Incorporating chart data into an imputation algorithm via the use of longitudinal mixed-effects model is an efficient way to impute longitudinal data that are missing from a randomized trial.

Keywords: blood pressure, clinical trials, imputation, longitudinal data, missing data.

Copyright © 2011, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

Introduction

When a substantial percent of patients in a randomized trial fail to show up for their follow-up visit, the resulting missing outcome data can seriously confound the interpretation of the results of the trial [1–3]. Statistical techniques for handling missing data, such as single or multiple imputation (MI), can yield unbiased estimates of treatment effectiveness in the presence of missing data [2–11], provided that the missing data conform to a missing at random (MAR) mechanism. The MAR assumption holds that the probability that some outcome data is missing can depend on other observed variables, but not on the value of the outcome itself, conditional on the observed variables. That is, after controlling for other observed variables, the value of the outcome variable is not a predictor of missingness. MAR is a generally untestable assumption [12], and whether it holds or not depends in part on the other observed variables available to the researcher. The more pertinent information a researcher has about the potentially missing outcome value, the more likely the MAR assumption is to hold.

Just such pertinent information may be available in patient medical charts, and utilizing this information may improve our ability to validly impute missing trial data. For example, a patient who fails to report for a follow-up blood pressure reading in a hypertension trial may nevertheless continue to keep appointments with his or her clinician who records the patient’s blood pressure and other signs and symptoms in the medical chart in the course of routine medical care. It seems reasonable that these chart-based blood pressure readings should be useful for imputing the missing trial-based blood pressure readings, but incorporating the chart-based blood pressures into an imputation algorithm is complicated by two factors. First, the trial protocol specifies that the blood pressures be taken at specific times during the trial, and while the patient may have clinic visits before and after these times, she or he may not have a visit during the specific window of time when the trial blood pressure was scheduled to take place. Second, the blood pressure recorded in the medical chart may be systematically different than the blood pressure that would have been recorded by trial personnel. Trial protocols, personnel, and equipment for recording blood pressure are standardized, whereas clinic procedures and equipment can vary from one provider to the next. In addition, so-called white coat hypertension [13] may cause blood pressure readings in clinicians’ offices to be systematically higher than those taken by trial personnel because patients’ blood pressures respond to the anxiety associated with seeing a clinician.

The goal of this article is to describe a novel method for overcoming these two challenges to using chart-based information to impute data missing from a clinical trial, and to assess via simulation the utility of this method compared to standard methods of imputing missing trial data. We apply this method to the imputa-
tion of missing blood pressure data from a hypertension trial. The method involves first specifying a mixed effects longitudinal model of blood pressure measurements taken from medical charts for each patient. From this model we recover the fitted value (both the fixed and random components) of patients’ predicted blood pressures for the days the patients were scheduled for follow-up blood pressure measurements in the trial. We then use these data in either a single- or multiple-imputation framework to impute missing trial blood pressure measurements.

The article is organized into two parts. In the first part we describe the method and apply it to data from the trial. In the second part we conduct simulations to assess the utility of this new method compared to other commonly used imputation methods.

Data and Methods

Data

The data come from a randomized trial of nurse-led disease management for patients with uncontrolled hypertension in east and central Harlem in New York City. In this trial, 416 patients were randomized to receive usual care, a home blood pressure monitor, or a home blood pressure monitor plus follow-up counseling by a trained nurse. The outcomes were blood pressure at 9 months and 18 months, which were taken in person by trained study personnel with standardized equipment and procedures. These systolic and diastolic readings are referred to as trial-based blood pressures. Patients also consented to have their blood pressure and other medical information abstracted from their medical charts from 1 year before and 18 months after enrollment in the trial. These readings are referred to as chart-based blood pressures.

Methods: Combining chart- and trial-based blood pressures Incorporating chart-based blood pressures into an imputation algorithm requires several steps. In the first step we estimated a mixed effect linear model of chart-based blood pressure as a function of time as shown in equation (1)

\[ Y_{it} = \beta_0 + \beta_1 t + \beta_2 t^2 + u_{it} + e_{it} \]  

Where \( Y_{it} \) is systolic blood pressure for patient i recorded from a medical chart at time t, which is days since enrollment in the trial; negative values represent pre-enrollment times. Time and its square appear as fixed effects in equation (1) to capture the overall trend in chart-based blood pressures over time.

Patient-specific random intercepts \( u_{i0} \) and random slopes \( u_{it} \) capture the idiosyncratic mean and trajectory of blood pressure for each patient. More complex models that involved higher powers of time and random slopes for the \( t^2 \) term were also investigated, but the specification of equation (1) was chosen based on the Bayesian Information Criterion. We assumed an unstructured covariance matrix for the random effects to allow for a correlation between the random slopes and intercepts, and estimated the parameters using restricted maximum likelihood.

In the second step, we calculated the fitted value of equation (1) at \( t = 274 \) and \( t = 548 \), which correspond to 9 months and 18 months after trial entry, when patients in the trial were scheduled for a follow-up blood pressure measurement. These fitted values included estimates for the fixed components as well as predictions of the random components \( u_{i0} \) and \( u_{it} \). These fitted values are shrinkage estimators of what the systolic blood pressure would have been for a specific patient if that patient had an outpatient visit on the day she or he was scheduled for a follow-up blood pressure reading from study personnel. Shrinkage estimators are useful in this context because the estimates for patients with few chart-based blood pressure measurements with which to estimate an individual specific trend are “shrunk” toward the population mean estimated trend [14].

In the third step, we tested if these fitted values are useful for predicting trial-based blood pressures at 9- and 18-month follow-up. We estimated a base model (equation 2a) that is a linear regression of trial-based blood pressure at follow-up month \( m (m = 9,18) \) as a function of baseline characteristics \( X_i \), baseline trial-based systolic blood pressure \( Y_{im} \), and a dummy variable \( I_{18} (m) \) that is equal to one if month = 18 and zero otherwise. Note that equation (2a) has no data in common with equation (1). The former uses chart-based blood pressure as the dependent variable and contains no covariates other than continuous measures of time. The latter uses trial-based blood pressure, includes patient baseline characteristics, and represents time by a dummy variable.

We then estimated a comparison model (equation 2b) that included the fitted values \( \hat{Y}_{im} \) from equation (1) as an additional explanatory variable, and recorded the additional variance explained by this model in terms of adjusted \( R^2 \). We used robust standard errors clustered on the patient to account for the maximum of two observations per patient.

\[ Y_{im} = \delta_0 + \delta_1 X_i + \delta_2 Y_{im} + \delta_3 I_{18} (m) + v_{im} \]  

(2a)

\[ Y_{im} = \delta_0 + \delta_1 X_i + \delta_2 \hat{Y}_{im} + \delta_3 I_{18} (m) + \delta_4 v_{im} \]  

(2b)

Results: Combining chart- and trial-based blood pressures Table 1 shows that 416 patients enrolled in the trial at baseline and 152 had missing blood pressure readings at either 9 or 18 months after enrollment. Women and Hispanics were less likely to be missing at follow-up, and systolic blood pressure at baseline was somewhat higher among patients who would eventually miss a follow-up reading. These 416 patients had 3345 blood pressure readings recorded in medical charts during outpatient visits during the 2 years surrounding the trial, with a large number occurring before and after the trial. Slightly lower blood pressure readings were recorded at routine outpatient visits than during visits with trial personnel.

Table 2 shows a comparison of longitudinal models for trial-based blood pressure, which correspond to equations (2a) and (2b) above. Each patient who completed a follow-up interview is included in this model; a patient who completed both the 9- and 18-month interview contributes two observations to the models. The first column of Table 2 shows coefficients and \( P \) values from a linear regression where the dependent variable was systolic blood pressure and explanatory variables included baseline characteristics of trial participants. The second column shows the same model with the addition of the fitted values from the chart-based systolic blood pressure model (this model is available from the author) as an explanatory variable. The addition of the fitted values adds significantly to the explanatory power of the model, as is demonstrated by the substantial increase in the adjusted \( R^2 \), from 0.102 without the chart-based fitted values to 0.264 with them. The significant decrease in the coefficient on baseline blood pressure in model (2b) compared to (2a) likely reflects that the chart-based fitted values is picking up some of the variance explained by baseline trial-based blood pressure. A similar benefit of including fitted values from the chart-based model is shown for diastolic blood pressure in columns three and four.

Methods: Simulation to compare imputation methods that incorporate chart-based data with standard imputation methods Next, we ran simulations to explore whether using these chart-based fitted values in a single or multiple imputation algorithm improved the quality of the imputation of trial blood pressures. The strategy was to force two missing data mechanisms on the trial-based blood pressure data—a MAR mechanism and a missing not at random (MNAR) mechanism—and then compare how various imputation methods performed under these scenarios. To do this we first estimated a logistic model where the dependent variable was a time-varying indicator for whether the patient’s trial-based blood pres-
sure data was missing at month 9 or 18, as shown in equation (3a).

\[ \Pr(\text{Missing}_{m} = 1) = f(\theta_0 + \alpha_i X_i + \alpha_{im} Y_{im} + \alpha_m 1_{im(m)}) \]  

(3a)

\[ \hat{P}_{im} = f(\hat{\theta}_0 + \hat{\alpha}_i X_i + \hat{\alpha}_{im} Y_{im} + \hat{\alpha}_m 1_{im(m)} + \hat{\theta}_m) \]  

(3b)

where \( f() \) is the logistic function, \( \alpha \) are the estimated coefficients from (3a) and \( Y_{im} \) are observed systolic trial-based blood pressures at month m for patient i transformed to a standard normal variable. By setting \( \theta \) equal to different values we can affect an MAR or MNAR pattern of missing data on the observed data. Setting \( \theta = 0 \) affects an MAR missing data mechanisms while \( \theta \) further from zero implies data that are MNAR. For example, \( \theta = \log(1.5) \) has the effect of increasing the odds of being missing at follow-up 1.5-fold for every one standard deviation increase over the mean in observed study blood pressure at follow-up. Because the standard deviation of systolic blood pressure was 16.8 and the probability of missing was set at 25%, this implied that a patient whose observed blood pressure was 16.8 mm Hg higher than the mean at 9 months had a 33% probability (i.e., 1.5 times higher odds) of being assigned “missing” at 9 months. We conducted simulations with values for \( \theta \) of 0, \( \log(1.25) \), and \( \log(1.5) \).

For each simulation, we drew a random variable from a Bernoulli distribution, and replaced the observed systolic blood pressure for patient i at month m with a missing value if the Bernoulli random variable was equal to one. In each simulation 25% of the observed blood pressures were converted to missing values. This rate of missing was similar to that observed in our trial and other hypertension trials [15], and is the attrition rate at which pathologies in imputation methodologies have been shown to occur [10]. In sensitivity analyses, we also used missing data rates of 10% and 40% (available from the author).

We then created six estimates of the missing blood pressures: two were derived from ad hoc imputation methods and four from regression-based methods. These are described in Table 3. The two ad hoc methods were last data carry forward (LDCF) and next and last (NAL). NAL imputes the missing 9-month blood pressure with the average of the baseline and 18-month provided the 18-month data are available; otherwise it uses LDCF. We used two regression methods (XB and X + Chl) in which the predicted values of equations (2a) and (2b), respectively, were used to impute the missing blood pressure. The two equations differ in that the latter includes the fitted value of the chart-based blood pressure as an additional covariate. In both cases, we added a

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total</th>
<th>Complete case</th>
<th>Any missing</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>416</td>
<td>264</td>
<td>152</td>
<td></td>
</tr>
<tr>
<td>Mean systolic blood pressure at baseline, mm Hg (sd)</td>
<td>153 (16.8)</td>
<td>151 (16.7)</td>
<td>156 (16.8)</td>
<td>0.015</td>
</tr>
<tr>
<td>Mean diastolic blood pressure at baseline, mm Hg (sd)</td>
<td>86.0 (13.4)</td>
<td>85.0 (13.0)</td>
<td>87.9 (13.8)</td>
<td>0.028</td>
</tr>
<tr>
<td>Age, mean (sd)</td>
<td>60.8 (11.6)</td>
<td>60.9 (11.3)</td>
<td>60.5 (12.1)</td>
<td>0.731</td>
</tr>
<tr>
<td>Race, % (n)</td>
<td>59.1 (246)</td>
<td>63.3 (167)</td>
<td>52.0 (79)</td>
<td>0.063*</td>
</tr>
<tr>
<td>Interviewed in Spanish, % (n)</td>
<td>30.8 (129)</td>
<td>26.1 (69)</td>
<td>38.8 (59)</td>
<td>0.007</td>
</tr>
<tr>
<td>Education, % (n)</td>
<td>51.4 (214)</td>
<td>52.7 (139)</td>
<td>49.3 (75)</td>
<td>0.515</td>
</tr>
<tr>
<td>Black non-Hispanic</td>
<td>17.5 (73)</td>
<td>16.7 (44)</td>
<td>19.1 (29)</td>
<td>0.533</td>
</tr>
<tr>
<td>Hispanic</td>
<td>36.5 (152)</td>
<td>33.3 (88)</td>
<td>42.1 (64)</td>
<td>0.421</td>
</tr>
<tr>
<td>Coronary artery disease, % (n)</td>
<td>19.5 (81)</td>
<td>17.8 (47)</td>
<td>22.4 (34)</td>
<td>0.257</td>
</tr>
<tr>
<td>Diabetes, % (n)</td>
<td>51.4 (214)</td>
<td>52.7 (139)</td>
<td>49.3 (75)</td>
<td>0.515</td>
</tr>
<tr>
<td>Depression, % (n)</td>
<td>17.5 (73)</td>
<td>16.7 (44)</td>
<td>19.1 (29)</td>
<td>0.533</td>
</tr>
<tr>
<td>Psychiatric diseases, % (n)</td>
<td>12.7 (53)</td>
<td>12.5 (33)</td>
<td>13.2 (20)</td>
<td>0.846</td>
</tr>
<tr>
<td>Renal disease, % (n)</td>
<td>17.1 (71)</td>
<td>19.3 (51)</td>
<td>13.2 (20)</td>
<td>0.108</td>
</tr>
<tr>
<td>Body mass index, % (n)</td>
<td>36.5 (152)</td>
<td>33.3 (88)</td>
<td>42.1 (64)</td>
<td>0.421</td>
</tr>
<tr>
<td>Inadequate health literacy, % (n)</td>
<td>43.5 (181)</td>
<td>39.0 (103)</td>
<td>51.3 (78)</td>
<td>0.015</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>70.9 (295)</td>
<td>74.2 (196)</td>
<td>65.1 (99)</td>
<td>0.049</td>
</tr>
<tr>
<td>Female, % (n)</td>
<td>70.9 (295)</td>
<td>74.2 (196)</td>
<td>65.1 (99)</td>
<td>0.049</td>
</tr>
<tr>
<td>Education, % (n)</td>
<td>51.4 (214)</td>
<td>52.7 (139)</td>
<td>49.3 (75)</td>
<td>0.515</td>
</tr>
<tr>
<td>Interviewed in Spanish, % (n)</td>
<td>30.8 (129)</td>
<td>26.1 (69)</td>
<td>38.8 (59)</td>
<td>0.007</td>
</tr>
<tr>
<td>Education, % (n)</td>
<td>51.4 (214)</td>
<td>52.7 (139)</td>
<td>49.3 (75)</td>
<td>0.515</td>
</tr>
<tr>
<td>Interviewed in Spanish, % (n)</td>
<td>30.8 (129)</td>
<td>26.1 (69)</td>
<td>38.8 (59)</td>
<td>0.007</td>
</tr>
<tr>
<td>Blood pressure readings from charts</td>
<td>86.0 (13.4)</td>
<td>85.0 (13.0)</td>
<td>87.9 (13.8)</td>
<td>0.028</td>
</tr>
<tr>
<td>Systolic, mean (sd)</td>
<td>79.8 (12.6)</td>
<td>79.2 (12.6)</td>
<td>80.9 (12.7)</td>
<td>0.187</td>
</tr>
<tr>
<td>Pre-enrollment</td>
<td>3345</td>
<td>2232</td>
<td>1113</td>
<td>117</td>
</tr>
<tr>
<td>0-9 mo</td>
<td>1725</td>
<td>1139</td>
<td>586</td>
<td></td>
</tr>
<tr>
<td>9-18 mo</td>
<td>860</td>
<td>594</td>
<td>266</td>
<td></td>
</tr>
<tr>
<td>&gt;18 mo</td>
<td>446</td>
<td>302</td>
<td>144</td>
<td></td>
</tr>
</tbody>
</table>
normal mean-zero random disturbance term with variance equal to the estimated residual variances from (2a) and (2b), respectively. Finally, we employed two methods based on a multiple imputation algorithm: the first (MI) used the explanatory variables in (2a) as explanatory variables and the second (MI + Cht) used the variables in (2b). For both, we created 10 imputations for each

<table>
<thead>
<tr>
<th>Imputation method</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad hoc</td>
<td>LDCF</td>
<td>Last data carry forward</td>
</tr>
<tr>
<td>Ad hoc</td>
<td>NAL</td>
<td>Next and last. The average of the last observed value before the missing value, and the next observed value. If no next value exists, then LDCF is used.</td>
</tr>
<tr>
<td>Regression based single imputation</td>
<td>XB</td>
<td>Predicted value from a regression of trial-based blood pressures on baseline patient characteristics, blood pressure at baseline, and a dummy variable for month. Additional variance is added in the form of a mean zero normal disturbance with variance equal to the estimated variance of the regression error.</td>
</tr>
<tr>
<td>Regression based single imputation</td>
<td>XB + Cht</td>
<td>Similar to XB but the regression includes as an additional covariate the fitted values from a linear mixed model of chart-based blood pressure as a function of time and time squared, with patient specific random components to the intercept and time parameters.</td>
</tr>
<tr>
<td>Multiple imputation</td>
<td>MI</td>
<td>Ten imputations based on the regression model used to predict XB</td>
</tr>
<tr>
<td>Multiple imputation</td>
<td>MI + Cht</td>
<td>Ten imputations based on the regression model used to predict XB + Cht, which include the chart-based fitted values as a covariate</td>
</tr>
</tbody>
</table>
missing blood pressure value. We used the multiple imputation algorithm “ice” in Stata 10 (2007, Stata Corp, College Station, TX) to perform the imputations. The “ice” algorithm uses imputation by chained equations, which has been shown to work well when compared to the more classic imputation approach using imputations from Bayesian posterior distributions [16].

We compared the properties of the six models using bias (bias = \( \frac{1}{n} \sum (Y_{im} - \hat{Y}_{im}) \)) and the root mean squared deviation

\[
\text{RMSD} = \sqrt{\frac{1}{n} \sum (Y_{im} - \hat{Y}_{im})^2},
\]

where n is the number of individuals with “missing” and imputed blood pressures, \( Y_{im} \) is true systolic blood pressure for individual i at follow-up month m, and \( \hat{Y}_{im} \) is systolic blood pressure imputed by one of the six methods. We also calculated the proportion of variance (PV = var(\( \hat{Y} \))/var(Y)), where var(\( \hat{Y} \)) was calculated using Rubin’s rule for the multiply imputed estimates, and coverage, which is the proportion of times the confidence interval for the mean of the imputed blood pressures included the mean of the true blood pressures. Each simulation involved 1000 replications and was conducted using Stata 10.

Results: Simulation to Compare Imputation Methods that Incorporate Chart-Based Data with Standard Imputation Methods

Figure 1 shows the simulation results for imputing systolic blood pressure from 1000 replications where the missing data rate was set at 25%. Results for diastolic blood pressure (not shown) were similar to those for systolic; results for missing data rates of 10% and 40% were similar to those for 25% and are available from the author. The logistic model of missingness at follow-up that was used to estimate the coefficients in equation (3a) and effect

the MAR and MNAR scenarios described below is also available from the author.

Results when the data are missing at random In the simulations that assumed a MAR missing data mechanism, the regression and multiple imputation methods generated imputations that were close on average to the true blood pressure measures, regardless of whether fitted values from charts were used to generate the imputation. However, incorporation of the chart-based fitted values led to lower RMSD for either the single imputation (RMSD for XB = 29 vs. XB + Cht = 26) or the multiple imputation algorithm (RMSD for MI = 30 vs. MI + Cht = 27). The estimates derived from multiple imputation were overdispersed (proportionate variance for MI = 2.7; MI + Cht = 2.4), but the MI and MI + Cht algorithms achieved coverages that were closest to nominal (Table 4).

The two ad hoc imputation methods substantially overestimated the true blood pressure (LDCF bias = 9.8 mm Hg; NAL bias = 7.6 mm Hg), which should be expected given the downward trend over time in blood pressure in the trial. The proportionate variance shows that the ad hoc estimates were somewhat under-dispersed compared to true values (proportionate variance for LDCF = 0.81; NAL = 0.77). The substantial bias and under-dispersion caused very poor coverage rates (Table 4).

Results when the data are missing not at random. When the missing data are MNAR, the model-based estimates systematically imputed blood pressures that were biased downward as expected (Fig. 1); however, the versions of these methods that incorporate the chart-based fitted values were less biased (e.g., for \( \theta = \log(1.5) \) bias XB = −8.2; XB + Cht = −6.6), had lower root mean squared
methods under MNAR not because these techniques were sup-
proportionate variances that were marginally closer to unity.

missing follow-up blood pressures were higher on average than
pressures were more likely to be “missing” at follow-up. Thus, the
MNAR scenario. We assumed patients with high follow-up blood
pressures are likely to be “missing” in the MNAR scenario, these ad hoc methods did not perform
overestimate the follow-up blood pressure if the missing data are
creased over time so using the ad hoc method of imputing a miss-
data we used than the methods. In the trial, blood pressures de-
tated to the ad hoc approaches, but this may be more a function of the
superiority of the new methods. For simulation scenarios in which
new method, but the interpretation still suggests the marginal
for comparisons of ad hoc methods (such as LDCF) to the proposed
the missing data bias. Thus, choosing between an ad hoc method
that has been recommended by some other studies [9] and the
new methods developed here should best be made with a good
understanding of the trends in the observed data over time and a
reasonable conceptual model of why data are missing.

Our incorporation of chart-based data is similar to other data
enhancement techniques [5,17] in which data from administrative
databases were used to replace or enhance clinical databases. In
these studies, a potentially missing indicator for a risk factor for a
given study subject was coded as present if either administrative
or clinical data sources indicated that the risk factor was present
during the calendar year. Our study differs from these analyses in
that the timing of the measurement was critical; simply averaging
chart-based blood pressure over a year would not capture the
downward trend in blood pressure observed in the trial.
Other studies that explored imputation of longitudinal missing data [3,9]
have found better results by using a patient’s own data to impute
missing observations rather than using the cross sectional mean
of data from other patients. Our results are similar: using a pa-
ient’s own data abstracted from clinical charts proved useful for
predicting a patient’s missing trial-based data.

The mechanism by which the chart-based method improved
regression-based single or multiple imputation is fairly straight-
forward. The chart-based multilevel model included patient spe-
cific intercepts and patient-specific time trends, so that a predic-
tion of chart-based blood pressure for a particular patient on a
particular day could be derived regardless of whether or not a
patient had an outpatient visit on that day. These predicted chart-
based blood pressures were highly correlated with observed trial-
based blood pressures taken on that day. This is evidenced by a
regression of trial-based blood pressures on chart-based fitted val-
ues (equation (2b)) in which the coefficient on the chart-based
values was 0.845 and highly significant (P < 0.001). This suggests
that for a patient who had a chart-based fitted value of blood pres-
ure 10 mm Hg higher than average at follow-up, we would expect
an 8.45 mm Hg higher than average blood pressure if the patient
kept the follow-up appointment with trial personnel on that day.
The inclusion of chart-based fitted values in the regression re-
sulted in a greater than 2.5-fold increase in the regression’s ad-
justed R². Because of this high correlation, the inclusion of these
fitted values led to better estimates of the missing trial-based
blood pressures.

In addition, because the chart-based fitted values are estimates
of what the missing blood pressure would have been had it been
measured in the trial, their inclusion in an imputation algorithm
helped to reduce the missing data bias in MNAR situations where
the probability of being missing was a function of the missing
blood pressure value. Other studies have found that when data are
MNAR traditional multiple imputation techniques cannot over-

<table>
<thead>
<tr>
<th>Missing data mechanism</th>
<th>LDCF</th>
<th>NAL</th>
<th>XB</th>
<th>XB + Cht</th>
<th>MI</th>
<th>MI + Cht</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing at random: θ = 0</td>
<td>0.00</td>
<td>0.01</td>
<td>0.83</td>
<td>0.87</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>Missing not at random: θ = 1.25</td>
<td>0.02</td>
<td>0.19</td>
<td>0.36</td>
<td>0.47</td>
<td>0.67</td>
<td>0.76</td>
</tr>
<tr>
<td>Missing not at random: θ = 1.50</td>
<td>0.18</td>
<td>0.62</td>
<td>0.05</td>
<td>0.09</td>
<td>0.18</td>
<td>0.29</td>
</tr>
</tbody>
</table>

LDCF, last data carry forward; MI, multiple imputation; MI + Cht, multiple imputation including chart-based data; NAL, next and last; XB, single imputation from regression; XB + Cht, single imputation from regression including chart-based data.

Discussion

We described a new method for using longitudinal data abstracted
from patients’ charts to impute data missing from a clinical trial, and
tested the method using data from a recently completed hyper-
tension trial. We tested a two-step process in which we first
estimated a longitudinal mixed model of the blood pressures ab-
stracted from patients charts as a function of time and patient-
specific random components to the intercept and time param-
eters, and recovered estimates of the patient-specific fitted values
from this model on the day the patient was scheduled for a fol-
low-up visit at 9 and 18 months after entry into the trial. This fitted
value is an estimate, according to data recorded in the patient
chart, of the patient’s blood pressure on the day he or she was
scheduled to return for a follow-up trial-based blood pressure
reading. In the second step we used these fitted values to impute
the missing trial-based blood pressures.

In simulations we found that this two-step process performed
marginally better than regression-based single or multiple impu-
tation methods that did not incorporate the chart-based fitted val-
ues. Incorporating fitted values from charts into a regression-
based single or multiple imputation algorithm yielded estimates
with lower bias, lower mean squared deviation, and a propor-
tionate variance closer to unity than did other estimates, even in sim-
ulations where a MNAR missing data mechanism was imposed.
In these latter simulations, incorporating the fitted values from
chart-based models ameliorated but did not eliminate the bias
that is inherent when data are MNAR.

The interpretation of the simulation results is more complex for
comparisons of ad hoc methods (such as LDCF) to the proposed
new method, but the interpretation still suggests the marginal
superiority of the new methods. For simulation scenarios in which
the missing data are MAR, the novel method was clearly superior
to the ad hoc approaches, but this may be more a function of the
data we used than the methods. In the trial, blood pressures de-
creased over time so using the ad hoc method of imputing a miss-
ing follow-up blood pressure with a baseline value would naturally
overestimate the follow-up blood pressure if the missing data are
MAR. In the MNAR scenario, these ad hoc methods did not perform
as badly, but this was mostly a function of how we designed the
MNAR scenario. We assumed patients with high follow-up blood
pressures were more likely to be “missing” at follow-up. Thus, the
missing follow-up blood pressures were higher on average than the
observed follow-up blood pressures. Because baseline blood
pressures were also higher on average than observed follow-up
blood pressures, using baseline blood pressures to impute missing

deviations (e.g., for θ = log(1.5) RMSD XB = 31; XB + Cht = 28), and propor-
tionate variances that were marginally closer to unity.

The ad hoc methods were less biased under MNAR than under
MAR but continued to impute blood pressure estimates that were
too high, increasingly underdispersed, and had poor coverage
properties. As discussed below, the bias was smaller for the ad hoc
methods under MNAR not because these techniques were supe-
rier but because of the way the MNAR scenarios were designed.
Chart-based blood pressure measurements can be effectively used to impute missing blood pressure data in a clinical trial. A two-step process of estimating a longitudinal mixed model on the chart data, and using the fitted values from this model as explanatory variables in a regression-based imputation algorithm led to imputed blood pressure values with lower bias and smaller root mean squared deviations than did imputation algorithms that did not incorporate the chart-based blood pressures. This technique might be useful for imputing a variety of longitudinal data that are missing from a clinical trial but collected during routine clinical visits outside the trial.

Source of financial support: This work was funded by a grant from the National Institutes of Health, National Center for Minority Health and Health Disparities (1R01MD00270-01). The views expressed herein are those of the authors and do not necessarily represent the views of the National Institutes of Health and other affiliated institutions.

Conclusions

Chart-based blood pressure measurements can be effectively used to impute missing blood pressure data in a clinical trial. A two-step process of estimating a longitudinal mixed model on the chart data, and using the fitted values from this model as explanatory variables in a regression-based imputation algorithm led to imputed blood pressure values with lower bias and smaller root mean squared deviations than did imputation algorithms that did not incorporate the chart-based blood pressures. This technique might be useful for imputing a variety of longitudinal data that are missing from a clinical trial but collected during routine clinical visits outside the trial.

Source of financial support: This work was funded by a grant from the National Institutes of Health, National Center for Minority Health and Health Disparities (1R01MD00270-01). The views expressed herein are those of the authors and do not necessarily represent the views of the National Institutes of Health and other affiliated institutions.

R E F E R E N C E S