Research Article

Statistics in Medicine

Received 1 July 2014, Accepted 26 May 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6562

Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis

Juned Siddique,[a*†] Jerome P. Reiter,[b] Ahnalee Brincks,[c] Robert D. Gibbons,[d] Catherine M. Crespi[e] and C. Hendricks Brown[f]

There are many advantages to individual participant data meta-analysis for combining data from multiple studies. These advantages include greater power to detect effects, increased sample heterogeneity, and the ability to perform more sophisticated analyses than meta-analyses that rely on published results. However, a fundamental challenge is that it is unlikely that variables of interest are measured the same way in all of the studies to be combined. We propose that this situation can be viewed as a missing data problem in which some outcomes are entirely missing within some trials and use multiple imputation to fill in missing measurements. We apply our method to five longitudinal adolescent depression trials where four studies used one depression measure and the fifth study used a different depression measure. None of the five studies contained both depression measures. We describe a multiple imputation approach for filling in missing depression measures that makes use of external calibration studies in which both depression measures were used. We discuss some practical issues in developing the imputation model including taking into account treatment group and study. We present diagnostics for checking the fit of the imputation model and investigate whether external information is appropriately incorporated into the imputed values. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: multiple imputation; data synthesis; individual participant data meta-analysis; external data calibration; data fusion; posterior predictive checking

1. Introduction

Meta-analysis has been used as a tool for synthesizing scientific literature for more than 30 years [1,2] and is the primary technique used in evidence-based medicine, particularly by the Cochrane Collaboration [3], as it can greatly increase power to detect effects. Cross-study synthesis becomes necessary as a field matures, and findings from individual studies must be combined to approach questions that cannot be answered through studying individual trials in isolation.

Traditionally, meta-analysis involves combining published summary findings across similar studies by placing each of these summaries on a common metric, such as a standardized mean difference or log odds ratio. A single overall measure of impact is then obtained which represents the overall effect across different trials [4–6]. There can be major shortcomings from meta-analyses that are based solely on published results, either from differences in the ways that the individual trials were analyzed or from the use of different outcome measures. Because there are no raw data available, meta-analysis can only

[a] *Department of Preventive Medicine, Northwestern University, Chicago, IL, U.S.A.*
[b] *Department of Statistical Science, Duke University, Durham, NC, U.S.A.*
[c] *Department of Public Health Science, University of Miami, Miami, FL, U.S.A.*
[d] *Departments of Medicine and Public Health Sciences, University of Chicago, Chicago, IL, U.S.A.*
[e] *Department of Biostatistics, University of California Los Angeles, Los Angeles, CA, U.S.A.*
[f] *Department of Psychiatry and Behavioral Sciences, Northwestern University, Chicago, IL, U.S.A.*
**Correspondence to: Juned Siddique, Department of Preventive Medicine, Northwestern University, Chicago, IL, USA.*
[†] *E-mail: siddique@northwestern.edu*

Copyright © 2015 John Wiley & Sons, Ltd.

synthesize from reported findings, restricting topics of inquiry to those questions that have already been addressed within individual studies. For more sophisticated analyses, such as mediation, moderation, growth modeling, and subgroup analysis, published reports of these analyses are rarely available for more than a small proportion of the studies, and even when they are, different modeling and reporting decisions across studies often lead to analytic incompatibility.

In response to limitations of traditional meta-analysis, an increasingly popular approach is individual participant data (IPD) meta-analysis in which the raw individual-level data for each study are obtained and used for synthesis [7]. With the raw data in hand, an analyst can adjust for patient-level covariates and take into account repeated measures, missing values, and differential follow-up times.

However, IPD meta-analysis can introduce challenges of its own. In particular, a common situation is when different outcome measures are used to assess the same construct in different studies [8]. The term *harmonization* has been coined to describe the procedure of placing variables on the same scale in order to permit pooling of data from a large number of studies [9, 10].

In this article, we describe a multiple imputation approach for harmonizing measures across longitudinal intervention trials when there is no overlap in outcome measures within trials. We seek a method that will permit pooling of IPD from studies that use different outcome measures for the same construct, which could greatly increase power for treatment effect analysis. We extend existing methods for harmonization by addressing harmonization in a longitudinal setting where different studies have different follow-up times and the relationships between outcome measures may change over time. Our method makes use of external calibration studies in which IPD are available on both outcomes, but the calibration studies are not to be included in final analyses. In addition to providing information on the relationship between outcome measures, the calibration studies facilitate the use of imputation diagnostics to assess the quality of imputations.

This article is organized as follows. In Section 2, we describe the example that motivated this work, a study of five randomized clinical trials investigating the effectiveness of fluoxetine for the treatment of depression among adolescents. We also discuss existing methods for harmonization. In Section 3, we describe our imputation model and diagnostics for checking the quality of imputations when variables are missing for all participants within a study. Section 4 presents the results of applying our methods to the fluoxetine data, and Section 5 offers discussion and areas for future work.

2. Motivation and background

2.1. *Motivating example*

The motivating example for this work are IPD from five randomized controlled trials (RCTs) studying the effectiveness of fluoxetine for the treatment of depression among adolescents. Two different measures of depressive symptoms were used in the trials. Four of the trials used the Children's Depression Rating Scale (CDRS), and one trial used the Hamilton Depression Rating Scale (HDRS). None of the five trials used *both* the CDRS and the HDRS. The HDRS employs 17 items to assess depressive symptoms, and scores range from 0 to 50 [11]. The CDRS also contains 17 items but was designed to assess depressive symptoms specifically in children ages 6–12 years [12]. Scores on the CDRS range from 17 to 113. Our objective is to combine information from all studies to estimate the effect of fluoxetine on the HDRS.

These five studies consist of all RCTs of fluoxetine in adolescents that included 30 or more patients, all of whom had a diagnosis of major depressive disorder. All five studies were double-blind, placebo-controlled RCTs. Data for four of the five fluoxetine trials were obtained from Eli Lilly and Co. Data from the fifth fluoxetine trial came from the Treatment for Adolescents With Depression Study (TADS) [13] and were obtained from the National Institute of Mental Health and do not include participants from a third TADS arm who received cognitive behavioral therapy. Only summary scores were available in all the trials; item-level data for neither the CDRS nor the HDRS were provided.

The top portion of Table I displays descriptive statistics by study for the HDRS and CDRS at baseline as well as age, gender, study duration, and sample size, and indicates which studies used which depression measures. Resche-Rigon *et al.* [14] refer to missing values that are missing for every observation in a study as *systematically missing*. Across the five fluoxetine studies, ages ranged from 7–18 years and trial size ranged from 40 to 221 participants. The length of the trials were from 3 to 12 weeks with anywhere from 3 to 9 assessments during the course of a study.

In this setting with no overlap between the CDRS and the HDRS, the partial correlation between the two depression measures after conditioning on background covariates is inestimable [15]. Therefore, in

Copyright © 2015 John Wiley & Sons, Ltd.

Table I. Descriptive statistics and missing data patterns of fluoxetine and venlafaxine trials at baseline. Not available (NA) indicate trials where the depression measure was not used.

Trial	CDRS (SD)	HDRS (SD)	Age (range)	Male (%)	Duration (weeks)	No. assess.	*n*
TADS fluoxetine	56 (10.8)	NA	13 (8–18)	51	9	7	221
Eli Lilly fluoxetine 1	44 (12.1)	NA	11 (7–18)	72	3	3	219
Eli Lilly fluoxetine 2	58 (10.3)	NA	13 (7–18)	54	8	9	172
Eli Lilly fluoxetine 3	55 (12.8)	NA	15 (12–18)	47	12	3	96
Eli Lilly fluoxetine 4	NA	22 (3.5)	16 (12–17)	45	6	7	40
			Calibration trials				
Wyeth venlafaxine 1	54 (8.7)	18 (5.1)	12 (7–17)	51	8	8	167
Wyeth venlafaxine 2	58 (9.2)	16 (4.8)	12 (7–18)	58	8	8	191

CDRS, Children’s Depression Rating Scale; HDRS, Hamilton Depression Rating Scale; TADS, Treatment for Adolescents With Depression Study; SD, standard deviation.

Table II. Partial correlation (controlling for age and gender) between CDRS and HDRS in calibration samples. Baseline (week 0) includes all participants. Subsequent estimates are based only on control participants.

Week	Study 1	Study 2	Overall
0	0.71	0.51	0.57
1	0.74	0.59	0.67
2	0.78	0.71	0.73
3	0.81	0.74	0.76
4	0.83	0.79	0.81
6	0.86	0.77	0.81
8	0.88	0.84	0.85

CDRS, Children’s Depression Rating Scale; HDRS, Hamilton Depression Rating Scale.

order to estimate the association between the two depression measures, data from two additional placebo-controlled adolescent depression trials were obtained, which used a different medication (venlafaxine), but included summary scores for both the CDRS and the HDRS on every participant [16]. Because these trials use a different treatment medication, we are not interested in including them in our final analysis. However, we do want to use the data from these two trials to obtain information on the relationship between the CDRS and HDRS in order to harmonize the CDRS and HDRS data in the fluoxetine trials. For this reason, we refer to these two trials as ‘external calibration trials’.

The last two rows of Table I provide information on the calibration trials, which were both 8 weeks in duration. An interesting and important feature of the external calibration trials is the changing nature of the partial correlation of the CDRS and the HDRS as a function of time. Table II displays the partial correlation (controlling for age and gender) of the CDRS and the HDRS as a function of study week. These partial correlations were estimated by using two separate linear regression models to regress CDRS on age and gender and HDRS on age and gender. The correlation between the error terms from these two regression models provides the partial correlation. This procedure was performed separately by time point. As can be seen, the partial correlation increases over the course of the study with an overall correlation of 0.57 at baseline and a correlation of 0.85 by the end of the study. One possible explanation for this increasing correlation is that the distribution of both measures is left-truncated at baseline because of the fact that participants were required to have an elevated depression score in order to meet inclusion criteria.

2.2. *Existing data harmonization methods*

There are a number of existing methods for data harmonization, which make use of the fact that even if different studies use different outcomes, they are attempting to measure the same construct. These methods fall roughly into three general classes: (i) linear or z-transformations to create a common metric

Copyright © 2015 John Wiley & Sons, Ltd.

across data sets; (ii) latent variable methods that identify an underlying latent construct across all studies [17–21]; and (iii) multiple imputation methods that treat unobserved measures as missing data and replace the missing values with plausible values [14, 22].

Transforming original units into standard deviation units via a z-transformation is a relatively simple solution and does not require common items across studies. A drawback is that this method does not take into account differences in variability across studies and assumes the underlying constructs are the same and measured equally well across studies [9]. Also, the transformed variable may not have the same scientific interpretation and familiarity as the original units with which substantive researchers are better able to interpret results and compare them with other findings in the literature. Latent variable methods solve some of the challenges associated with using z-scores but require strong assumptions about latent structure that can be difficult to check, especially when there are few or no variables of interest measured simultaneously in the data sets to be harmonized. Both approaches, as currently implemented, struggle with repeated observations, and typically assume that repeated observations on the same participant are independent.

Multiple imputation [23] is a natural approach for handling missing data because of unmeasured outcomes and has a number of advantages over existing methods. With multiple imputation, missing values are replaced with two or more plausible values to create two or more completed data sets. Analyses are then conducted separately on each data set, and final estimates are obtained by combining the results from each of the imputed data sets using rules that account for within-imputation and between-imputation variability. See [24] for a review.

In the context of harmonization for IPD meta-analysis, once unmeasured variables have been imputed, analyses and their subsequent inferences are based on existing scales of interest rather than a z-score or a latent variable. In addition, after the data set has been filled in, it can be shared with other investigators and be used for numerous analyses using complete data methods. In fact, once a variable has been multiply imputed, it may be used as an outcome in one analysis, and as a covariate in another analysis.

While multiple imputation was originally conceived for use in single surveys with missing data, the method has been extended to numerous areas. Most relevant to data harmonization in IPD meta-analysis is *data fusion* where two (or more) data sets from the same population (but different samples) are to be combined, and some variables only appear in one data set or the other, but no data set contains all variables of interest [25–27].

For example, in our motivating example of five RCTs where the HDRS and the CDRS are systematically missing, multiple imputation can be used to complete the concatenated data set, by imputing the HDRS in four of the studies and the CDRS in the fifth study. However, in this setting in which the HDRS and CDRS are never jointly observed (as they are in the five fluoxetine trials), the maximum likelihood estimate of the partial correlation between the HDRS and the CDRS, conditioning on background covariates is inestimable [15]. When estimating these parameters and generating imputations in a Bayesian framework, the posterior estimates of the partial correlations will be equal to the prior correlation [28]. Thus, the typical non-informative priors used in Bayesian imputation models would result in partial independence between the HDRS and CDRS. Because the two measures are presumed to be measuring a similar construct, this is unlikely to be a plausible assumption.

Rässler [26] proposed a multivariate normal imputation model for the data fusion setting where imputations are generated by positing a value for the partial correlation between two outcomes that are never jointly observed. Rässler [26] suggests three ways of specifying this correlation: (i) from its prior based on a distributional assumption (e.g., uniform over some range); (ii) using several arbitrary values; and (iii) using values estimated from an external data set with information on the joint distribution between the two variables of interest.

Gelman *et al.* [22] developed an imputation model to impute missing data from several sample surveys where not all the surveys asked the same questions. Here, a separate imputation model is fit for each survey, but with parameters across the surveys linked using a hierarchical model so that imputations for questions not asked in a survey are determined by data from the other surveys in the population as well as by available responses to other questions in that survey. An advantage of this approach is that covariates may be included in the regression model at both the individual and survey levels.

3. Methods

Our approach for harmonizing the depression data across the multiple trials is in the spirit of Gelman *et al.* [22] and Rässler [26] where the uncollected depression measures are considered missing data and missing

Copyright © 2015 John Wiley & Sons, Ltd.

observations are multiply imputed. We extend these existing approaches to the case of longitudinal data, and we develop methods that make use of external calibration data. Not only do the calibration data provide information on the relationship between the two depression measures, but they also facilitate the use of imputation diagnostics to assess the quality of the imputations.

We begin by concatenating the calibration data sets with the fluoxetine data sets. We then generate multiple imputations for the missing CDRS and HDRS data using an imputation model that estimates the relationship between these two measures based on information from the calibration data. Once the missing data have been imputed, we perform diagnostics to check whether the imputed data are consistent with the observed data. Finally, we remove the calibration data and perform post-imputation analyses on the five fluoxetine trials using the multiply imputed completed data sets.

3.1. *Imputation model*

Our imputation method is based on a multivariate random-effects model that jointly models the two depression measures over time [29]. Using notation similar to that of [30], let y_{ij} and w_{ij} be the HDRS and CDRS scores, respectively, for participant i measured at time j where $j = 1, \dots n_i$. Our model is

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{gender}_i + \beta_3 time_{ij} + \beta_4 \left(time_{ij} \times T_i\right) + \eta_{y0i} + \eta_{y1i} time_{ij} + \varepsilon_{yij} \\ w_{ij} &= \alpha_0 + \alpha_1 \text{age}_i + \alpha_2 \text{gender}_i + \alpha_3 time_{ij} + \alpha_4 \left(time_{ij} \times T_i\right) + \eta_{w0i} + \eta_{w1i} time_{ij} + \varepsilon_{wij} \end{aligned} \tag{3.1}$$

where T_i is a variable that indicates whether participant i received fluoxetine or placebo and $time_{ij}$ is the number of days since baseline, log transformed to approximate linearity and avoid the use of an additional parameter to estimate a quadratic term [31]. The β and α parameters represent fixed effects, while η_{y0i} and η_{y1i} are random intercept and slope terms for y, and η_{w0i} and η_{w1i} are random intercept and slope terms for w. We assume these random effects $\eta_i = (\eta_{y0i}, \eta_{y1i}, \eta_{w0i}, \eta_{w1i})$ have a multivariate normal distribution $\eta_i \sim N_4(\mathbf{0}, \Delta)$, where Δ is a 4×4 variance–covariance matrix.

The residual errors $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{yij}, \varepsilon_{wij})$ are also correlated and follow a bivariate normal distribution $\boldsymbol{\varepsilon}_{ij} \sim N_2(0, \Sigma)$. Residual errors $\boldsymbol{\varepsilon}_{ij}$ are independent across participants and time and are independent of the random effects.

This model allows the covariance between the CDRS and HDRS to vary as a function of time as was observed in the external calibration data. We have

$$\begin{aligned} Cov(y_{ij}, w_{ij}) &= Cov\left(\eta_{y0i} + \eta_{y1i} time_{ij} + \varepsilon_{yij}, \eta_{w0i} + \eta_{w1i} time_{ij} + \varepsilon_{wij}\right) \\ &= d_{13} + time_{ij} d_{14} + time_{ij} d_{23} + time_{ij}^2 d_{24} + \sigma_{yw} \end{aligned}$$

where d_{ij} is the ith row and jth column of the random effects variance–covariance matrix and σ_{yw} is the covariance of the residual errors. In addition to modeling the correlation of the two measures (and its change) over time, the model also estimates the within-subject correlation on the same measure over time, that is, the intra-class correlation.

Note that the imputation model described in Equation (3.1) does not include a term for trial. A fixed effect for trial is not estimable in this model because for some trials all outcomes are missing. There is little information available to estimate random trial effects because the number of trials is small and estimating random effects at the trial level for both the HDRS and the CDRS requires us to estimate the correlation of these two measures at the trial level. With only two calibration trials, we do not have enough degrees of freedom to estimate the study-level correlation between the CDRS and the HDRS. Further complicating our efforts is the fact that the direction of the correlation is negative, which is unlikely to be the true direction of this relationship and is more likely due to the small sample size.

3.2. *Model assumptions*

We highlight four assumptions in the imputation model described in Section 3.1.

Assumption 1

We assume the missing depression data are missing at random (MAR) such that the probability of missingness depends only on observed data included in the model and not on unobserved data. In our setting, the MAR assumption implies that an investigator's decision not to use a depression measure is not related to unobserved variables, in particular the unmeasured depression score itself. The choice of whether to

Copyright © 2015 John Wiley & Sons, Ltd.

use the HDRS or the CDRS is primarily based on age, where the CDRS was designed to be more relevant to younger children than the HDRS. Because we are able to condition on age in our imputation models and the calibration data are based on an age range comparable with the age ranges in the fluoxetine trials, we are comfortable with this assumption.

Assumption 2

The second assumption relates to the partial correlation between the CDRS and the HDRS. As before, let y_{ij} and w_{ij} indicate the HDRS and CDRS scores, respectively, for participant i at time j. The variable T_i is the treatment assignment for participant i, and $X_i = (\text{age}_i, \text{gender}_i)$ are the age and gender for participant i. Treatment assignment, age, and gender are observed on all participants. Then, at any two time points j and k.

$$\text{Corr}\left(y_{ij}, w_{ik} | X_i = x_i, T_i = t_i\right) = \text{Corr}\left(y_{ij}, w_{ik} | X_i = x_i\right), \tag{3.2}$$

that is, the partial correlation between the HDRS and CDRS does not depend on treatment group. We make this assumption because the external calibration trials do not use fluoxetine and thus do not provide information on the correlation between measures for fluoxetine participants. We restrict our external calibration trials to placebo participants (and all participants at baseline) and make the assumption in Equation (3.2) that the partial correlation is the same for placebo and fluoxetine participants.

Assumption 3

The third assumption is that the partial correlation between the CDRS and the HDRS does not differ by trial. Let y_{ijl} and w_{ijl} indicate the HDRS and CDRS scores, respectively, for participant i at time j in trial l. For two participants i and i' in trials l and m,

$$\text{Corr}\left(y_{ijl}, w_{ikl} | X_i = x_i\right) = \text{Corr}\left(y_{i'jm}, w_{i'km} | X_{i'} = x_i\right). \tag{3.3}$$

Assumption 4

Our fourth assumption concerns independence of observations within the same trial in our imputation model. As before, let y_{ijl} and w_{ijl} indicate the HDRS and CDRS scores, respectively, for participant i at time j in trial l. Then we make the following assumptions. At any two time points j and $k, i \neq i'$, and conditioning on auxiliary variables $\text{age}_i, \text{gender}_i$,

$$\text{Corr}\left(y_{ijl}, y_{i'kl}\right) = 0 \tag{3.4}$$

$$\text{Corr}\left(w_{ijl}, w_{i'kl}\right) = 0 \tag{3.5}$$

$$\text{Corr}\left(y_{ijl}, w_{i'kl}\right) = 0. \tag{3.6}$$

3.3. Estimation and imputation

We used a Bayesian approach to estimate the parameters of our imputation model and to generate imputations. As noted by [32], imputations generated from a Bayesian model are *proper* in the sense that they incorporate both model parameter uncertainty as well as uncertainty due to missingness. We placed non-informative priors ($N(0, 10^3)$) on the regression coefficients of the fixed effects (β, α) in Equation (3.1). For both the random effects covariance matrix and the residual error covariance matrix, we use inverse-Wishart priors with 6 and 4 degrees of freedom, respectively, and a scale parameter equal to the identity matrix.

Markov chain Monte Carlo (MCMC) via the software OpenBUGS [33], was used to draw model parameters and generate imputations. We diagnosed convergence of the MCMC algorithm using statistics developed by [34] and concluded that the Markov chain converged after 20,000 iterations. OpenBUGS code and MCMC diagnostics for some of the imputation model parameters are included in the supplementary materials. Imputations were generated by running an additional 50,000 iterations on a single chain and drawing a set of parameters from every 500 iterations. This high thinning value was chosen to ensure that imputed values were generated from approximately independent draws of the imputation model parameters [32].

After imputing the missing outcome measures, we removed the calibration data and analyzed only data from the fluoxetine trials. In this setting, when some of the records used to estimate the imputation model are not used for analysis, the usual multiple imputation combining rules variance estimator has

Copyright © 2015 John Wiley & Sons, Ltd.

positive bias because the conditioning used by the imputer and the analyst are not matched [35]. To address this, we used Reiter's [35] method of two-stage multiple imputation and its associated combining rules, which has better performance in this context. First, m values of the parameters in the imputation model are sampled, and then, nested within each set of parameter draws, n imputations are generated for each missing value. This results in mn completed data sets. Analyses are performed separately on each imputed data set, and the results are combined using the two-stage imputation rules described in [35], which are provided in Appendix A. We used $m = 100$ sets of parameter draws with $n = 2$ imputations nested within each parameter draw for our application.

3.4. *Diagnostics*

In our setting, where the amount of missing data is considerable, and where we are imputing values for every participant within a trial, it is particularly important to check the imputation model and the quality of its imputations. We use two general approaches for checking our imputation model: graphical comparisons of the observed data versus the imputed data [22, 36] and posterior predictive checks using numerical summaries based on test statistics [37]. In both cases, we focus on diagnostics that capture important features of the data that are relevant to our target analyses.

Graphical exploration consisted of comparing observed (imputed) HDRS values with imputed (observed) CDRS values as a function of time. To do this, we created side-by-side scatterplots of depression scores over time. In one panel, the data are observed, in the other, the data are every participant's average imputed values at each time point. While the scale of the two depression measures is different, we expect to observe similar patterns over time. If not, this suggests a problem with the fit of our imputation model.

The numerical approach follows the posterior predictive checking and re-imputation method of He and Zaslavsky [37] and used the following strategy. First, we duplicated the data from the two external calibration trials and deleted the HDRS values in this duplicated version of the calibration data. Next, we concatenated these duplicated data sets with the original seven data sets (the five fluoxetine trials and the two original calibration trials) resulting in a data set containing nine trials. Finally, we generated imputations using the methods described in Section 3.1, imputing missing values in the fluoxetine studies and the missing HDRS values that were deleted in the duplicated calibration data . Let Y be the observed HDRS calibration data and Y^{imp} the imputed version of Y. To compare observed data with imputed data, we use a *test statistic*, $T(Y, \theta)$, some scalar function of the data and possibly imputation model parameters. Posterior predictive checking consists of comparing $T(Y, \theta)$ to the distribution of $T(Y^{imp}, \theta)$. Lack of fit of the imputed data to the observed data can be quantified by the *posterior predictive p-value* (*ppp*) [38, 39]; the probability that the imputed data are more extreme than the observed data, as measured by the test statistic, that is,

$$ppp = 2 \times \min\left(Pr(T(Y^{imp}, \theta) > T(Y, \theta)|Y), Pr(T(Y^{imp}, \theta) < T(Y, \theta)|Y)\right). \tag{3.7}$$

In practice, we can calculate *ppp* by simulation. For each imputed data set $v, v = 1, \dots, V$, we calculate $T(Y^{imp,v}, \theta^v)$. The estimated two-sided *ppp* is the proportion of the V test statistics that equal or exceed the test statistic based on the observed data, that is,

$$ppp = \frac{2}{V} \times \min\left(\sum_{v=1}^{V} I(T(Y^{imp,v}, \theta^v) > T(Y, \theta^v)), \sum_{v=1}^{V} I(T(Y^{imp,v}, \theta^v) < T(Y, \theta^v))\right).$$

A small *ppp* suggests that the proposed imputation model is not adequate to support the targeted post-imputation analysis [37]. We investigated three sets of test statistics that capture important relationships linked to our substantive analyses. These test statistics are: (i) the partial correlation (controlling for age and gender) between the HDRS and CDRS at weeks 0, 4, 6, and 8; (ii) the HDRS means at weeks 0, 4, 6, and 8; and (iii) the fixed intercept and slope terms from a random intercept and slope regression model regressing HDRS on (log) days since baseline.

4. Results based on application to adolescent data

We begin this section by presenting the results of the diagnostics to ensure that our imputations are reasonable and are replicating important relationships relevant to our target analyses. We then analyze the fluoxetine data first using the CDRS as the outcome, then the HDRS.

Copyright © 2015 John Wiley & Sons, Ltd.

4.1. *Results from diagnostics*

Figure 1 presents scatter plots of CDRS and HDRS data versus time (days since baseline) for each of the trials in Table I. The first row of the panel are the two calibration data sets in which both the CDRS and the HDRS are observed. The next two rows of the panel are the fluoxetine trials that only used the CDRS. For these trials, we have imputed 200 HDRS values for every observation in the trial and displayed the average of these 200 imputations. The last row is the fluoxetine trial that used the HDRS. Here, we imputed CDRS values for every participant in the trial. As can be seen, within each trial, the imputed values appear to preserve the relationship that is seen between the observed depression score and time.

Figure 2 displays the results of posterior predictive checks (based on 200 imputed data sets) on the partial correlation between the CDRS and the HDRS in the duplicated calibration data at weeks 0, 4, 6, and 8. At all time points, the correlation based on imputed values is close to the observed calibration study correlation (indicated by the vertical line), and our imputation model is clearly capturing the increasing correlation over time. Posterior predictive *p*-values are large at all weeks; thus, we find no evidence that the model generates unreasonable imputations.

Figure 3 displays the results of posterior predictive checks on the HDRS means in the duplicated calibration data at weeks 0, 4, 6, and 8. At all time points, the imputed mean is close to the observed calibration study mean (indicated by the vertical line). Posterior predictive *p*-values are large at all weeks. As with the partial correlations, none of these *p*-values suggest that our imputations are unreasonable. Because the observed correlation between the CDRS and HDRS was lowest at baseline, it is not surprising that our imputations are less accurate at baseline than at other time points.

Figure 4 displays the results of posterior predictive checks of the fixed intercept and slope coefficients of a random intercept and slope regression model of HDRS score as a function of *log*(number of days since baseline + 1) in the calibration data. As with the other posterior predictive checks, the parameters based on imputed values are close to observed values and posterior predictive *p*-values do not suggest an imputation model that is failing to capture important relationships.

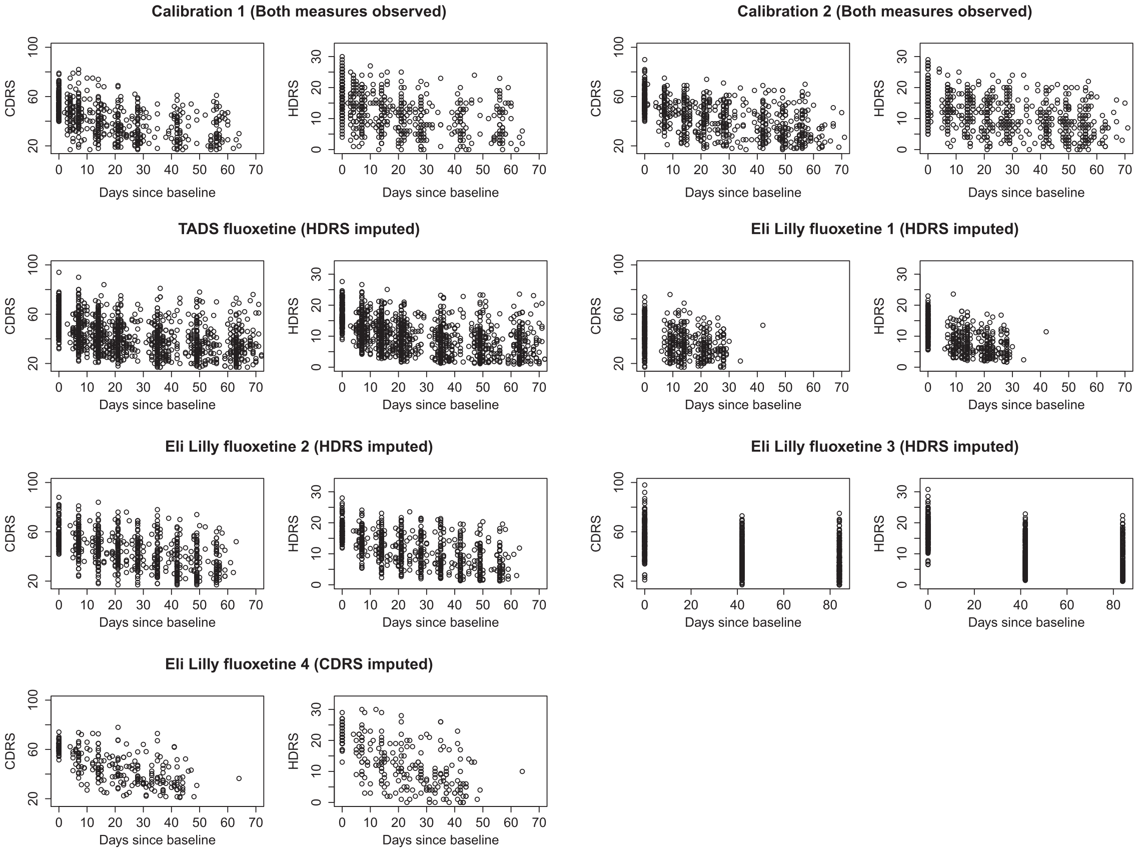

Figure 1. Plot of Children's Depression Rating Scale (CDRS) scores versus Hamilton Depression Rating Scale (HDRS) scores by study. Imputed values are the mean of 200 imputations. TADS, Treatment for Adolescents With Depression Study.

Copyright © 2015 John Wiley & Sons, Ltd.

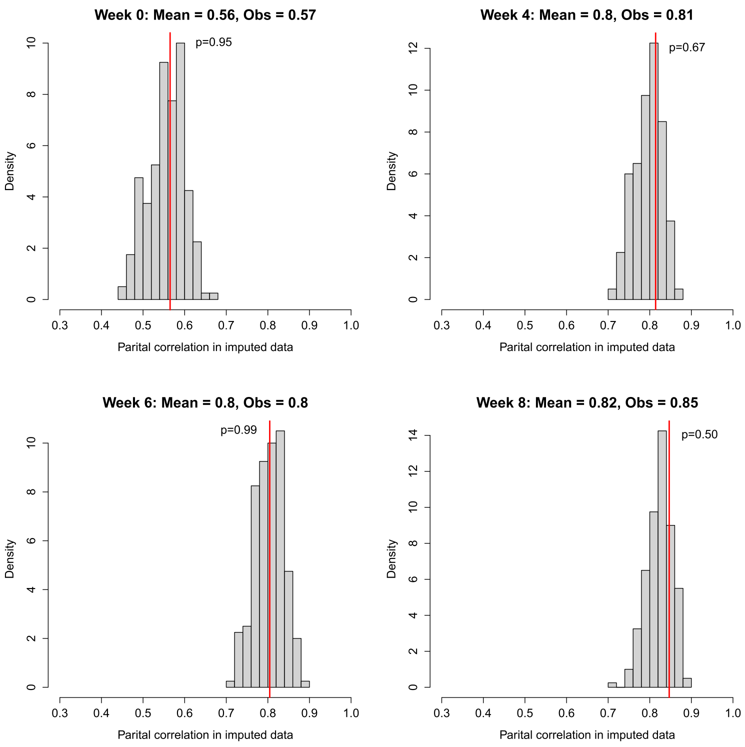

Figure 2. Posterior predictive checks of the partial correlation (controlling for gender and age) between the Hamilton Depression Rating Scale and the Children's Depression Rating Scale in the external calibration data by week. The histogram displays the distribution of partial correlations based on imputed values from 200 imputed data sets. The vertical lines indicate the observed partial correlation. The posterior predictive *p*-values shown in each histogram are two-sided.

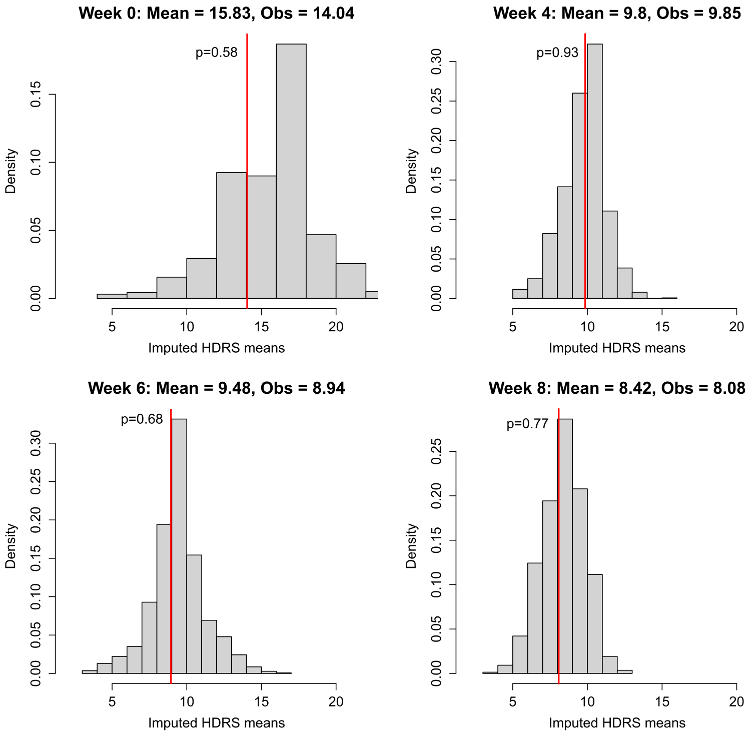

Figure 3. Posterior predictive checks of weekly Hamilton Depression Rating Scale (HDRS) means in the external calibration data. The histogram displays the distribution of imputed means from 200 imputed data sets. The vertical lines indicate the observed means. Posterior predictive *p*-values are two-sided.

Copyright © 2015 John Wiley & Sons, Ltd.

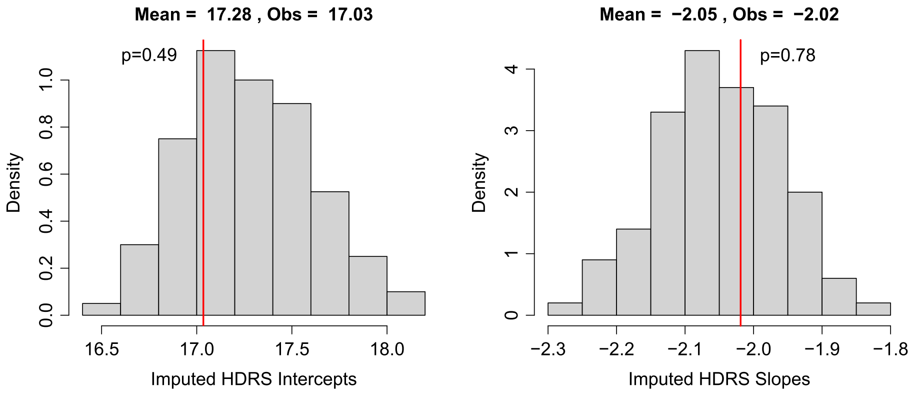

Figure 4. Posterior predictive checks of regression coefficients from a random intercept and slope regression model estimating the effect of time on Hamilton Depression Rating Scale (HDRS) score in the external calibration data. The panel on the left displays the intercept in the model, the panel on the right the slope. The histograms display the distribution of the regression coefficients based on imputed values from 200 imputed data sets. The vertical lines indicate the observed regression coefficients. Posterior predictive p-values are two-sided. Time is *log*(number of days since baseline + 1).

Table III. Descriptive statistics and missing data patterns of fluoxetine and venlafaxine trials at baseline. Values with an asterisk next to them indicate that they are based on multiple imputations.

	CDRS	HDRS	Age		Duration	No.	
Trial	(SD)	(SD)	(range)	Male (%)	(weeks)	assess.	*n*
TADS fluoxetine	56 (10.8)	17 (5.5)*	13 (8–18)	51	9	7	221
Eli Lilly fluoxetine 1	44 (12.1)	14 (5.6)*	11 (7–18)	72	3	3	219
Eli Lilly fluoxetine 2	58 (10.3)	18 (5.4)*	13 (7–18)	54	8	9	172
Eli Lilly fluoxetine 3	55 (12.8)	17 (5.9)*	15 (12–18)	47	12	3	96
Eli Lilly fluoxetine 4	62 (10.4)*	22 (3.5)	16 (12–17)	45	6	7	40
			Calibration trials				
Wyeth venlafaxine 1	54 (8.7)	18 (5.1)	12 (7–17)	51	8	8	167
Wyeth venlafaxine 2	58 (9.2)	16 (4.8)	12 (7–18)	58	8	8	191

CDRS, Children's Depression Rating Scale; HDRS, Hamilton Depression Rating Scale; TADS, Treatment for Adolescents With Depression Study; SD, standard deviation.
*Based on imputed values.

4.2. *Post-imputation analysis of fluoxetine data*

Based on the findings from the imputation diagnostics, we proceeded to discard the calibration data and analyze the five fluoxetine trials. Table III is the same as Table I earlier, but here, we have filled in the missing cells with those based on imputed values. Looking at Table III, those trials with low baseline CDRS scores tended to have low baseline HDRS scores. The Eli Lilly fluoxetine trial 1 had the lowest mean baseline CDRS score of all the trials, including the calibration trials and also has the lowest mean baseline HDRS score. Conversely, the Eli Lilly fluoxetine trial 4 had a high mean baseline HDRS score as compared with the calibration trials, and this is also reflected in its imputed mean baseline CDRS score.

We then analyzed the CDRS scores (both observed and imputed) as a function of treatment and time using the following random-effects regression model. Let CDRS_{ijl} be the CDRS score for participant i at occasion $j, j = 1, \ldots, n_i$ in trial $l, l = 1, \ldots, N$; and let $time_{ij}$ be the time since baseline and T_i a variable indicating whether participant i received fluoxetine or placebo. Then our model is

$$\text{CDRS}_{ijl} = \beta_0 + \beta_1 time_{ij} + \beta_2 \left(time_{ij} \times T_i\right) + b_{0l} + b_{0i} + b_{1i} time_{ij} + \varepsilon_{ijl}. \tag{4.1}$$

As in our imputation model, time has been log transformed. The term b_{0l} is a trial-specific random intercept term that follows a normal distribution and takes into account between-trial variability. The terms b_{0i} and b_{1i} are random intercept and slope terms at the participant level, respectively, and follow a bivariate normal distribution independent of the trial-specific random intercept. The error term ε_{ijl} also follows a

Copyright © 2015 John Wiley & Sons, Ltd.

normal distribution and is independent of the random effects. The model was fit using the lmer function from the R [40] package lme4 [41].

In this model, inference focuses on the regression coefficient β_2, the time by treatment interaction. This term represents the difference in slopes between treatment and control groups. Our trials are short in duration, and this parameterization fits our data well. A more flexible approach to model the effect of treatment would be to treat time as a nominal variable by specifying indicator variables for each time period and interacting these indicator variables with treatment. However, because each study has different follow-up times, we are unable to fit this more flexible model in this particular application.

The top half of Table IV presents the results of our analysis using only the observed CDRS scores as well as using both observed and imputed CDRS scores. As noted earlier, post-imputation analyses were based on the two-stage imputation combining rules of Reiter [35]. Not surprisingly, there is very little difference in parameter estimates and their standard errors between the two CDRS analyses. This is because only one trial did not use the CDRS, and this was the smallest trial in our data set ($n = 40$). Thus the observed CDRS analysis consists of 708 of the 748 participants in our analysis. Focusing on the treatment by time interaction in Table IV, the treatment effect is significant in both CDRS analyses.

The bottom half of Table IV presents the observed-only analysis and the post-imputation analysis of HDRS scores in the fluoxetine trials using the same model in Equation (4.1) but with the HDRS as the outcome. Here, the observed-only results are based on a single trial ($n = 40$), whereas the post-imputation analysis are based on imputing HDRS scores for 708 out of the 748 (95%) participants in our analysis. Because the observed-only HDRS analysis is based on a single trial, the trial-level random intercept term is omitted. There are substantial differences between the observed-only and post-imputation HDRS analyses. In particular, the intercept in the observed analysis is much higher than that of the post-imputation analysis, reflecting the high baseline depression scores in fluoxetine trial 4. Also, the time by treatment interaction term is not significant in the observed-only analysis, but it is significant in the imputed analysis. It appears that in trial 4, both treatment and placebo participants improved a great deal and by similar amount, possibly because of their high depression scores at baseline. When including the other four trials in the analysis, a treatment effect is detected as the placebo participants in these trials did not improve to the same degree as the placebo participants in trial 4.

Table IV. Observed-only and post-imputation analyses of Children's Depression Rating Scale (CDRS) and Hamilton Depression Rating Scale (HDRS) scores in fluoxetine trials. The observed-only HDRS analysis is based on a single trial does not include a random effect at the trial level. All other models include a trial-level random effect.

		Observed				Imputed			
Outcome	Parameter	Est	SE	*t*-value	*p*-value	Est	SE	*t*-value	*p*-value
CDRS	Intercept	54.00	2.56	21.12	<0.001	55.20	2.26	24.39	<0.001
	Time	−3.79	0.18	−21.34	<0.001	−3.97	0.17	−22.77	<0.001
	Tx∗Time	−1.06	0.21	−4.92	<0.001	−0.94	0.21	−4.50	<0.001
	SD(b_{0l})	5.03				4.98			
	SD(b_{0i})	9.91				9.84			
	SD(b_{1i})	2.52				2.58			
	Corr(b_{0i}, b_{1i})	−0.46				−0.45			
	SD(ε_{ijl})	7.23				7.30			
HDRS	Intercept	22.59	0.67	33.48	<0.001	17.19	0.77	22.3	<0.001
	Time	−3.34	0.42	−7.87	<0.001	−1.62	0.04	−45.25	<0.001
	Tx∗Time	−0.57	0.55	−1.03	0.308	−1.22	0.18	−6.95	<0.001
	SD(b_{0l})	NA				1.8			
	SD(b_{0i})	2.11				4.74			
	SD(b_{1i})	1.46				1.23			
	Corr(b_{0i}, b_{1i})	0.13				−0.58			
	SD(ε_{ijl})	3.97				3.24			

SE, standard error; Tx, treatment; NA, not available; SD(b_{0l}), standard deviation of random trial-level intercepts; SD(b_{0i}), standard deviation of random subject-level intercepts; SD(b_{1i}), standard deviation of random subject-level slopes; Corr(b_{0i}, b_{1i}), correlation of random intercepts and slopes; SD(ε_{ijl}), standard deviation of residual error.

Copyright © 2015 John Wiley & Sons, Ltd.

Another meaningful difference between the two HDRS analyses in Table IV is the reduction in standard errors because of the inclusion of the additional data. There is a 67% reduction in the standard error of the time by treatment interaction based on the post-imputation analysis. As a result, the magnitude of the *t*-value of the treatment effect on the HDRS, −6.95, is similar to the magnitude of the *t*-value of the treatment effect on the CDRS which is −4.50.

5. Discussion

We have described a multiple imputation approach for harmonizing outcomes across multiple longitudinal trials. In our motivating example, we sought to harmonize two depression measures where there were no studies that used both measures. To do this, we made use of external calibration data in order to estimate key relationships and generate more accurate imputations. Besides providing information on the relationship between the two depression measures, the calibration data facilitate the use of diagnostics to address how well imputed values preserve important relationships related to the target analyses. A benefit of the multiple imputation approach is that once the missing data are filled in, analyses can proceed using complete data methods. This makes the data accessible by a wide variety of researchers, many of whom will not have the background knowledge or the technical expertise to harmonize the data themselves.

Our imputation model is a multivariate linear mixed-effects model and explicitly models the effect of time and treatment on our target variables. This is important so that our imputation model is congenial with our analysis objective [42], which is to investigate the difference in change over time between treatment and placebo groups. With differential follow-up times across study, attempting to impute missing data at each time point using standard software such as SAS Proc MI [43] or MICE [44] would result in many variables to be imputed and parameters to be estimated. Similarly, an imputation model based on a covariance pattern model rather than random effects would be difficult to fit and unlikely to be parsimonious.

An additional benefit of modeling the effect of time is that if a user of our method wishes to harmonize the assessment time points, they simply need to add a new row for each participant in their data set where the outcome is missing, and the time value is set to its desired value. While in our example, the missing values are outcomes in the analysis, our imputation model is flexible and makes no distinction between outcomes or time-varying covariates.

We make several assumptions in developing our imputation model. Most notable is the assumption of no between-study variability. Including random effects at the trial level in our imputation model is dependent upon being able to estimate the correlation between the HDRS and the CDRS at the trial level. With only two calibration trials, this correlation is not estimable. Further complicating our efforts is the fact that the direction of the correlation is negative. Thus, our imputation model assumes that observations on different participants in the same trial are independent. Our analysis model, however, did include a trial-level random effect term, and it is interesting to compare the trial-level intra-class correlations between the imputed CDRS analysis and the imputed HDRS analysis. For the CDRS analysis in which data from a single trial is imputed, the baseline ICC is 0.14. For the HDRS analysis in which data from four of five trials is imputed, the baseline ICC is 0.09. So although our imputation model ignores the effect of trial, between-trial variability does not seem to be severely attenuated in post-imputation analyses.

There are several limitations to our approach. Our imputation model is a linear mixed-effects model that assumes the outcomes are normally distributed, although they are actually bounded above and below, which can result in imputed values that are out of range. Methods have been developed for handling bounded continuous outcomes [45]. For our data, where the HDRS ranges from 0 to 50 and the CDRS ranges from 17 to 113, the data were relatively symmetric and using a Gaussian-based imputation model resulted in few out-of-range imputations. Only 2.7% of the imputed CDRS values were out of range, and only 5.2% of imputed HDRS values were out of range. For these observations, we redrew our imputations until imputed values were in range. Some research has shown that when imputing highly skewed limited-range variables, it is best to allow imputed values to remain out of range [46]. For our purposes, in which the normality assumption produced few out-of-range values, this is less of a concern. In addition, the results of our diagnostic checks suggest that our imputations are preserving important features of the data. In settings where outcomes are highly skewed and/or very limited in range, it may be appropriate to allow out-of-range imputed values or use more flexible imputation models that can handle a combination of continuous, ordinal, and binary data [47].

Our method requires calibration data because there is no overlap in depression measures within any of the trials of interest. If overlap does exist, then calibration data are not necessary—although they may

Copyright © 2015 John Wiley & Sons, Ltd.

still be useful. To investigate the utility of a calibration data set in our setting, we imputed the missing HDRS and CDRS values without using the calibration data and imputed each variable separately using a univariate mixed-effects imputation model [29] as implemented in the R [40] package pan [48]. We then analyzed the imputed data using the analysis model in Equation (4.1). The results are displayed in the Supporting Information. The results based on imputed data without the calibration sample are similar to the observed-only analyses and suggest that there is little benefit to imputing missing outcomes when the analysis model and imputation model condition on the same set of observed covariates.

Our calibration data were from trials of venlafaxine and as a result do not provide information on the partial correlation between the HDRS and CDRS among participants randomized to fluoxetine. Thus, we make the assumption that this partial correlation does not depend on the treatment group, and we restrict our external calibration trials to only placebo participants (and all participants at baseline). To explore whether the use of an antidepressant affects the partial correlation between the HDRS and CDRS, we compared the partial correlation between the HDRS and CDRS in the calibration trials, stratifying by time, treatment group, and study (included in the Supporting Information). Post baseline, the partial correlations in the treatment and control groups are similar at each time point. These results suggest that Assumption 2 is a reasonable one, in the sense that the use of an antidepressant does not seem to affect the partial correlation between the HDRS and CDRS. These results also suggest that we would see little change in our results if we were to include both treatment and control groups in the calibration data.

Our imputation model is complex and the random-effect covariance matrix itself includes 10 parameters. While it would be possible to fit a more parsimonious model, such as a model with a shared intercept and response specific random slopes, the tradeoff for this parsimony is less flexibility in estimating the correlation between the CDRS and HDRS. Because the primary goal of our model is generating imputations rather than inference, we see little advantage to parsimony when our calibration data allow us estimate the parameters from a more complicated model.

Directions for future work include examining harmonization across a larger number of trials and trials with time-varying variables and developing a three-level imputation model where repeated observations are nested within participants who are nested within trials. Our application had an insufficient number of trials to support such models. These models may make use of shared parameters [49] and/or use informative priors for situations when calibration data do not exist or are inadequate to model a larger set of parameters.

Increasingly, researchers are collecting data from multiple studies in order to synthesize findings and perform more sophisticated analyses. These projects will continue to grow as US funding agencies encourage data sharing [50,51], and more journals require the release of data to accompany papers. Methods that harmonize variables across data sets and facilitate analyses by many researchers are increasingly important in order to make full and efficient use of synthesized data and take advantage of the potential of IPD meta-analysis to address new questions not answerable by a single study.

Appendix A: Nested multiple imputation combining rules

In the following, we describe the nested multiple imputation approach of [35]. We use notation that follows closely to that of Shen [52].

Let Q be the quantity of interest. Assume with complete data, inference about Q would be based on the large sample statement that

$$(Q - \hat{Q}) \sim N(0, U),$$

where $\hat{Q}$ is a complete-data statistic estimating Q and U is a complete-data statistic providing the variance of $Q - \hat{Q}$. The $M \times N$ imputations are used to construct $M \times N$ completed data sets, where the estimate and variance of Q from the single imputed data set is denoted by $(\hat{Q}^{(m,n)}, U^{(m,n)})$ where $m = 1, 2, \ldots, M$ and $n = 1, 2, \ldots, N$. The superscript (m, n) represents the nth imputed data set under the set of parameters m. Let $\bar{Q}$ be the overall average of all $M \times N$ point estimates

$$\bar{Q} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \hat{Q}^{(m,n)}, \qquad (A.1)$$

and let $\bar{Q}_m$ be the average of the mth model,

$$\bar{Q}_m = \frac{1}{N}\sum_{n=1}^{N}\hat{Q}^{(m,n)}. \tag{A.2}$$

Three sources of variability contribute to the uncertainty in Q. These three sources of variability are: $\bar{U}$, the overall average of the associated variance estimates

$$\bar{U} = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}U^{(m,n)}, \tag{A.3}$$

B, the between-model variance

$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\bar{Q}_m - \bar{Q})^2, \tag{A.4}$$

and W, the within-model variance

$$W = \frac{1}{M(N-1)}\sum_{m=1}^{M}\sum_{n=1}^{N}(\hat{Q}^{(m,n)} - \bar{Q}_m)^2. \tag{A.5}$$

The quantity

$$T = \bar{U} + \left(1 + \frac{1}{M}\right)B - \left(1 + \frac{1}{N}\right)W \tag{A.6}$$

estimates the total variance of $(Q - \bar{Q})$. Interval estimates and significance levels for scalar Q are based on a Student's t-reference distribution

$$T^{-\frac{1}{2}}(Q - \bar{Q}) \sim t_v, \tag{A.7}$$

where v, the degrees of freedom, follows from

$$v^{-1} = \left[\frac{\left(1+\frac{1}{M}\right)B}{T}\right]^2 \frac{1}{M-1} + \left[\frac{\left(1+\frac{1}{N}\right)W}{T}\right]^2 \frac{1}{M(N-1)}. \tag{A.8}$$

It is possible that $T < 0$, particularly for small M and N. In this situation, analysts can use $\tilde{T} = (1+1/M)B$. Here inferences are based on a t-distribution with $M - 1$ degrees of freedom. Generally, negative values of T can be avoided by making M and N large.

Acknowledgements

The authors wish to thank Robert Weiss and Lei Liu for many helpful discussions. This work was supported by NIH grants R01 MH040859 (Drs Siddique, Brown, and Brincks), K07 CA154862-01 (Dr Siddique), R01 MH080122 (Drs Brown and Gibbons), and UL1TR000124 (Dr Crespi). Dr Reiter was supported by NSF grant NSF-SES-11-31897. The data were supplied by the National Institute of Mental Health (Treatment for Adolescents With Depression Study), Wyeth, and Eli Lilly and Co.

References

1. Glass GV. Primary, secondary, and meta-analysis of research. *Educational Researcher* 1976; **5**:3–8.
2. Glass TA, McAtee MJ. Behavioral science at the crossroads in public health: extending horizons, envisioning the future. *Social Science & Medicine* 2006; **62**:1650–1671.
3. Higgins JPT, Green S (eds). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*. The Cochrane Collaboration, 2011. Available from: http://www.cochrane-handbook.org.
4. Hedges LV, Olkin I. *Statistical Methods for Meta-analysis*. Academic Press: Orlando, FL, 1985.

Copyright © 2015 John Wiley & Sons, Ltd.

5. Cook TD. *Meta-Analysis for Explanation: A Casebook*. Russell Sage Foundation: New York, NY, 1994.
6. Wilson DB, Lipsey MW. The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychological Methods* 2001; **6**:413–429.
7. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ: British Medical Journal* 2010; **340**:521–525.
8. Lu G, Kounali D, Ades A. Simultaneous multioutcome synthesis and mapping of treatment effects to a common scale. *Value in Health* 2014; **17**:280–287.
9. Griffith L, van den Heuvel E, Fortier I, Hofer S, Raina P, Sohel N, Payette H, Wolfson C, Belleville S. Harmonization of cognitive measures in individual participant data and aggregate data meta-analysis. Technical Report, Agency for Healthcare Research and Quality Rockville, MD, 2013.
10. Hussong AM, Curran PJ, Bauer DJ. Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology* 2013; **9**:61–89.
11. Hamilton M. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry* 1960; **23**:56–62.
12. Poznanski EO, Freeman LN, Mokros HB. Children's depression rating scale–revised (September 1984). *Psychopharmacology Bulletin* 1985; **21**:979–989.
13. March J, Silva S, Petrycki S, Curry J, Wells K, Fairbank J, Burns B, Domino M, McNulty S, Vitiello B, Severe J, the Treatment for Adolescents With Depression Study (TADS) Team. Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents with Depression Study (TADS) randomized controlled trial. *JAMA: The Journal of the American Medical Association* 2004; **292**:807–820.
14. Resche-Rigon M, White IR, Bartlett JW, Peters SAE, Thompson SG. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine* 2013; **32**:4890–4905.
15. Rubin DB, Thayer D. Relating tests given to different samples. *Psychometrika* 1978; **43**:3–10.
16. Emslie GJ, Findling RL, Yeung PP, Kunz NR, Li Y. Venlafaxine ER for the treatment of pediatric subjects with depression: results of two placebo-controlled trials. *Journal of the American Academy of Child & Adolescent Psychiatry* 2007; **46**: 479–488.
17. Curran PJ, Hussong AM, Cai L, Huang W, Chassin L, Sher KJ, Zucker RA. Pooling data from multiple longitudinal studies:the role of item response theory in integrative data analysis. *Developmental Psychology* 2008; **44**:365–380.
18. Flora DB, Curran PJ, Hussong AM, Edwards MC. Incorporating measurement nonequivalence in a cross-study latent growth curve analysis. *Structural Equation Modeling* 2008; **15**:676–704.
19. Curran PJ. The seemingly quixotic pursuit of a cumulative psychological science: introduction to the special issue. *Psychological Methods* 2009; **14**:77–80.
20. Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological Methods* 2009; **14**:81–100.
21. Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychological Methods* 2009; **14**:101–125.
22. Gelman A, King G, Liu C. Not asked and not answered: multiple imputation for multiple surveys. *Journal of the American Statistical Association* 1998; **93**:846–857.
23. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons: New York, 1987.
24. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. *Statistics in Medicine* 2007; **26**:3057–3077.
25. Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* 1986; **4**:87–94.
26. Rässler S. A non-iterative Bayesian approach to statistical matching. *Statistica Neerlandica* 2003; **57**:58–74.
27. Reiter JP. Bayesian finite population imputation for data fusion. *Statistica Sinica* 2012; **22**:795–811.
28. Kadane JB. Some statistical problems in merging data files. *Journal of Official Statistics* 2001; **17**:423–433.
29. Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics* 2002; **11**:437–457.
30. Weiss RE. *Modeling Longitudinal Data*. Springer: New York, 2005.
31. Mosteller F, Tukey JW. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley: Reading, MA, 1977.
32. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC: New York, 1997.
33. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine* 2009; **28**:3049–3067.
34. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–472.
35. Reiter JP. Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika* 2008; **95**:933–946.
36. Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2008; **57**:273–291.
37. He Y, Zaslavsky AM. Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Statistics in Medicine* 2012; **31**:1–18.
38. Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 1996; **6**:733–760.
39. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis* 2nd edition. Chapman & Hall/CRC press: Boca Raton, FL, 2004.
40. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2012.
41. Bates D, Maechler M, Bolker B, Walker S. *lme4: Linear Mixed-Effects Models Using Eigen and s4*, 2013. http://CRAN.R-project.org/package=lme4 [Accessed on 7 January 2015], R package version 1.0-4.

Copyright © 2015 John Wiley & Sons, Ltd.

42. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**:538–558.
43. Yuan Y. Multiple imputation using SAS software. *Journal of Statistical Software* 2011; **45**:1–25.
44. Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. *Journal of Statistical Software* 2011; **45**:1–67.
45. Din SHM, Molas M, Luime J, Lesaffre E. Longitudinal profiles of bounded outcome scores as predictors for disease activity in rheumatoid arthritis patients: a joint modeling approach. *Journal of Applied Statistics* 2014; **41**:1627–1644.
46. Rodwell L, Lee KJ, Romaniuk H, Carlin JB. Comparison of methods for imputing limited-range variables: a simulation study. *BMC Medical Research Methodology* 2014; **14**:1–11.
47. Boscardin WJ, Zhang X, Belin TR. Modeling a mixture of ordinal and continuous repeated measures. *Journal of Statistical Computation and Simulation* 2008; **78**:873–886.
48. Zhao JH, Schafer JL. *pan: Multiple Imputation for Multivariate Panel or Clustered Data*, 2013. R package version 0.9.
49. Gueorguieva R. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling* 2001; **1**:177–193.
50. National Institutes of Health. Final NIH statement on sharing research data, 2003. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html [Accessed 3 March 2014].
51. National Science Foundation. Dissemination and sharing of research results, 2011. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4 [Accessed 3 March 2014].
52. Shen ZJ. Nested multiple imputation. *Ph.D. Thesis*, Department of Statistics, Harvard University, Cambridge, MA, 2000.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

Copyright © 2015 John Wiley & Sons, Ltd.