

HHS Public Access

Author manuscript *J Alzheimers Dis.* Author manuscript; available in PMC 2024 May 04.

Published in final edited form as:

J Alzheimers Dis. 2024; 99(1): 263–277. doi:10.3233/JAD-231047.

IMPROVING REGRESSION ANALYSIS WITH IMPUTATION IN A LONGITUDINAL STUDY OF ALZHEIMER'S DISEASE

Ganesh Chandrasekaran^a, Sharon X. Xie^{a,*}, the Alzheimer's Disease Neuroimaging Initiative

^aDepartment of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Background—Missing data is prevalent in the Alzheimer's Disease Neuroimaging Initiative (ADNI). It is common to deal with missingness by removing subjects with missing entries prior to statistical analysis; however, this can lead to significant efficiency loss and sometimes bias. It has yet to be demonstrated that the imputation approach to handling this issue can be valuable in some longitudinal regression settings.

Objective—The purpose of this study is to demonstrate the importance of imputation and how imputation is correctly done in ADNI by analyzing longitudinal Alzheimer's Disease Assessment Scale – Cognitive Subscale 13 (ADAS-Cog 13) scores and their association with baseline patient characteristics.

Methods—We studied 1063 subjects in ADNI with Mild Cognitive Impairment. Longitudinal ADAS-Cog 13 scores were modeled with a linear mixed-effects model with baseline clinical and demographic characteristics as predictors. The model estimates obtained without imputation were compared with those obtained after imputation with Multiple Imputation by Chained Equations (MICE). We justify application of MICE by investigating the missing data mechanism and model assumptions. We also assess robustness of the results to the choice of imputation method.

Results—The fixed-effects estimates of the linear mixed-effects model after imputation with MICE yield valid, tighter confidence intervals, thus improving the efficiency of the analysis when compared to the analysis done without imputation.

Conclusions—Our study demonstrates the importance of accounting for missing data in ADNI. When deciding to perform imputation, care should be taken in choosing the approach, as an invalid one can compromise the statistical analyses.

Keywords

Alzheimer's disease; biomarkers; missing data; imputation; longitudinal study

^{*}Correspondence to: Sharon X. Xie, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA. Fax: 215-573-4865; sxie@pennmedicine.upenn.edu. Author Contributions

Ganesh Chandrasekaran (Conceptualization; Methodology; Formal Analysis; Writing – Original Draft); Sharon Xie (Conceptualization; Methodology; Writing – Review & Editing; Supervision; Project Administration; Funding Acquisition)

Conflict of Interest/Disclosure Statement

The authors have no conflict of interest to report.

Introduction

Missing data is a prevalent problem in large-scale studies like the Alzheimer's Disease Neuroimaging Initiative (ADNI). Age, invasive procedures, and progression of clinical conditions can dissuade patients from consistently participating in the study and can result in incomplete data [1]. A common practice is to omit subjects with missing entries and perform analyses on the remaining observations. However, when there is considerable missingness in the data set, this can lead to significant loss of statistical efficiency as a result of reduced sample size. In cases where the missingness disproportionately affects certain groups in the study sample, this omission may even lead to biased results [2–4]. Given the large amount of missingness in the ADNI data set, it is important to consider accounting for missing data when performing statistical analyses.

Rubin in [5] describes three general mechanisms that cause missing data. The mechanisms causing missingness in ADNI are investigated in [1]. The simplest possibility, missing completely at random (MCAR), occurs when patient characteristics do not affect missingness; rather, information is missing purely at random. The authors in [1] found that missingness in the ADNI data set is not MCAR, but depends on patient characteristics and that the cause also varies across biomarkers and clinical groups. In this case, data can either be missing at random (MAR) or missing not-at-random (MNAR). Data is MAR if missingness is related to observed data, but not to any unobserved data. If missingness is related to unobserved data, it is MNAR. The two cases, MAR and MNAR, cannot be distinguished from the data alone; doing so requires external information [1,2]. In this paper, we plan to study the association of baseline clinical features with longitudinal scores from the Alzheimer's Disease Assessment Scale – Cognitive Subscale 13 (ADAS-Cog 13). We will use this as an example to illustrate the importance of accounting for missing data and demonstrate how to rigorously perform imputation in longitudinal models in ADNI.

The objective of our study is to illustrate the importance of accounting for MAR data in the longitudinal regression setting, an important regime for the study of Alzheimer's Disease (AD) considering the intricate temporal patterns involving biomarkers and disease progression [6–8]. We also demonstrate how accounting for missing data via imputation is done correctly in the longitudinal setting. We focus on a particular method, Multiple Imputation by Chained Equations (MICE), an iterative regression-based procedure commonly used for MAR data [9]. We show that imputation results in greater statistical efficiency, evidenced by tighter confidence intervals around estimates compared to the available case analysis (ACA), which makes use of only the available data. We justify the use of MICE in our data set by assessing model assumptions and performing sensitivity analyses.

Methods of dealing with missingness in ADNI have been considered in [6,10,11]. However, these studies differ in contribution from this current paper because they either do not consider imputation methods for non-MCAR data or are concerned with imputation for classification purposes only. Separately, imputation for MAR data in AD risk analysis has previously been considered in [12]. The authors use MICE as part of a sensitivity analysis

to determine whether their main findings changed with imputed data; in so doing, they acknowledge the value of considering missing data when attempting to perform reliable statistical analyses in AD studies. However, this study does not look at imputation in a longitudinal setting, which has its own particular challenges, as we do in this paper. Additionally, we provide detailed discussions of the various statistical considerations that need to be made for proper imputation, such as analysis model assumption diagnostics, imputation model diagnostics, and efficiency comparisons between methods, the majority of which are not mentioned in [12]. Overall, while some studies exist which have considered missingness in ADNI, our paper makes a novel contribution by focusing on the increasingly relevant longitudinal regime and providing a detailed account of the various statistical considerations that have to be made to guarantee reliable imputation results.

We focus specifically on the association of baseline clinical features with longitudinal scores from the Alzheimer's Disease Assessment Scale - Cognitive Subscale 13 (ADAS-Cog 13) - an exam widely used to assess cognitive dysfunction - among patients diagnosed with Mild Cognitive Impairment (MCI) [13]. We explore these associations in this study using the available case and using imputed data via MICE. We choose the ADAS-Cog 13 score, as opposed to other cognitive assessment scores available, because of its sensitivity for detecting progression of cognitive decline [14,15]. Furthermore, total ADAS-Cog 13 scores are not as susceptible to floor and ceiling effects as other assessment scores, making it more appropriate for use in a linear model [16,17]. Baseline clinical features of interest, shown in Table 1, are age, sex, years of education, reported family history of AD, baseline sum of boxes of Clinical Dementia Rating score (CDR-SB), levels of common cerebrospinal fluid (CSF) biomarkers of AD, and presence of the ϵ 4 allele of the apolipoprotein E gene (APOE ε4). We included baseline CDR-SB as a predictor to mitigate the effects of left-censorship on our model. Specifically, because the disease duration of MCI patients is unknown at the time of enrollment, and because CDR-SB serves as a reasonable surrogate for this unknown duration [18–20], its inclusion as a predictor will improve the analysis model as well as imputation of missing variables. There are three core CSF biomarkers of AD: 42-amino acid β -amyloid peptide (A β_{42}), phosphorylated tau protein (pTau), and total tau protein (tTau). In addition, biomarker ratios pTau/A β_{42} and tTau/ A β_{42} have been shown to be informative markers of disease state, with higher levels of both correlated with disease severity. Lower $A\beta_{42}$ levels and higher pTau, tTau, pTau/A β_{42} , and tTau/A β_{42} levels are consistently found amongst AD patients [6,7]. All five of these biomarkers are considered in our analyses. Presence of the APOE ε 4 allele predicts progression to AD [6,21] and therefore may also be associated with cognitive performance.

Methods

Study Data

The data used for the statistical analyses in this article were from ADNIMERGE.csv, FAMXHPAR.csv, and FHQ.csv, all downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database on October 28th, 2021. We studied 1063 subjects diagnosed with Mild Cognitive Impairment (MCI). Study protocols describe follow-ups every 6 months from the initial visit until month 24. After month 24, subjects are followed

up once every 12 months. Cognitive, imaging, genetic, and biochemical data are collected during each visit.

Before performing any analyses, levels of CSF biomarkers were clipped at their technical limits as in [22–24]. For example, $A\beta_{42}$ levels could only be detected within the range 200 pg/ml to 1700 pg/ml. For subjects with levels outside of this range, those levels were reported as <200 or >1700. We set such values to their respective limits of 200 and 1700. Analogously, we clipped pTau and tTau levels at their respective technical limits.

FAMXHPAR.csv and FHQ.csv were used to obtain information about patient-reported family history of AD. Family history was said to be present if the patient reported either of their parents having AD. If the patient reported that neither parent had AD, family history was said to be absent. Any other response (unsure or no entry) was considered a missing value to be imputed.

Statistical Analysis

We are interested in the ability of baseline covariates of MCI patients to predict their average ADAS-Cog 13 score across visits. To explore this, we developed five linear mixed-effects models – each involving one of the CSF biomarkers – based on scientific and statistical considerations. We refer to these models as the *analysis models*. Biomarker ratios were log-transformed throughout the analyses. We check all five of the models' underlying normality assumptions by graphical assessment of the distribution of estimated random intercepts. Then, the analysis models are fitted under two cases: with and without performing imputation beforehand. The results from these two different cases are then compared.

In order to fit the analysis models without imputation, we remove all subjects with any missing baseline measurements. We refer to this procedure as an *available case analysis* (ACA). We argue later in this section that the ACA is expected to produce valid inference – that is, that the fixed-effects and standard error estimates are all expected to be appropriate estimates of the true population quantities. However, since we are discarding potentially useful information, we may still suffer loss of statistical efficiency. As we expect both the ACA and imputation analysis to produce valid fixed-effects estimates, we can compare their statistical efficiencies based on estimated confidence intervals. A narrower confidence interval is indicative of a more efficient procedure.

Next, we fit the analysis models after imputing the missing data. We refer to this procedure as the *imputation analysis*, which we would like to compare with the ACA. The first step in designing an imputation procedure is to choose an imputation approach. We considered multilevel imputation with PAN to account for within-subject correlations. However, simulation studies suggest that imputation with PAN can be problematic under MAR data when missingness occurs in both the outcome and the predictors (Zhang P & Xie SX, unpublished manuscript). As a result, we decided to perform single-level imputation of our data with MICE, an iterative regression-based approach commonly used for MAR data.

For each of the variables that we will impute (we will refer to these as *target variables*), we must specify a model which MICE will use to generate imputations. We design these so-called *imputation models* for ADAS-Cog 13, each of the five CSF biomarkers, and patient-reported family history of AD, since these are the variables with the most substantial missingness, and consequently for which proper imputation is most crucial for producing valid results (see Tables 1 and 2 for information about missingness for each variable). The design of the imputation models is driven by three primary factors [3]: first, every predictor in the analysis model should be included in the imputation models. Second, any variable that predicts missingness of a target variable should be included in its imputation model. These variables can be identified by univariate tests of association. Third, variables that can explain considerable variance in the target variable should be included in its imputation model.

We performed MICE for each of the five analysis models separately. We considered 50 imputations based on recommendations from [25] with 40 maximum iterations of the MICE algorithm based on the convergence plots in Figure 1. Convergence is achieved when the lines – each tracking one of 50 imputations – stabilize around some mean and standard deviation region, which should be comparable to the means and standard deviations of the observed data set. Furthermore, the lines within a single graph should be intertwined, not showing any obvious trend. The results of the 50 imputations are pooled using Rubin's Rules [3].

Numerical data were imputed by Bayesian linear regression, and categorical data by logistic regression. We also considered imputation by predictive mean matching (pmm) for numerical data to assess robustness to different imputation procedures. Because the original CSF measurements were clipped, we clipped imputed CSF values at their technical limits after imputation. Results are shown for Bayesian linear regression imputation after clipping, although we verified that the results did not change significantly whether Bayesian linear regression or pmm was used, and whether or not we performed clipping.

We imputed the time-varying outcome, ADAS-Cog 13 scores, only at times for which the proportion of data missing was below a specified tolerance, since imputations become increasingly dependent on the correctness of the chosen imputation model when there is a larger percentage of missingness [4]. We chose tolerances of 50% (wherein we tolerate times for which up to 50% of ADAS-Cog 13 scores are missing), 65%, and 75%, and compared results across the three tolerances.

In the following subsections, we develop the analysis and imputation models we will use for this study.

Analysis Model

The analysis model is motivated by the scientific objective. We are interested in the ability of baseline characteristics of MCI patients to predict their average ADAS-Cog 13 across visits. This motivates a linear mixed-effects model as our analysis model, with longitudinal ADAS-Cog 13 as our outcome and time (in months), sex, age, years of education, self-reported family history of AD, APOE e4 allele presence, baseline CDR-SB score, and baseline

The mixed-effects model assumes a normal distribution on the random intercept. To assess adherence to this assumption, we plotted the histograms of the random intercepts for each of the five models when fitted with the available cases. The plots are shown in Figure 2 and they appear to fit a normal distribution well. It is important to note, however, that normality of the random intercept is a tricky assumption to verify, particularly due to the intangible nature of unobservable quantities like random effects [26]. Therefore, while it is still important to check these assumptions to detect egregious violations [27], the reader should keep in mind the limitations of such tests. It is encouraging, however, that fixed-effects and variance estimates for linear mixed-effects models seem reasonably robust to deviations from the normality assumption (though the same cannot be said of nonlinear models like the generalized linear mixed models) [26–29]. Altogether, these findings suggest that our model is still reasonable.

To be able to compare efficiencies of the ACA and the imputation analysis, we must establish validity of the ACA for these analysis models. In regression settings, including longitudinal regression, ACA on MAR data can still produce valid inference under certain conditions on the analysis model. Namely, if missingness of the covariates is independent of observed outcome values conditional on observed covariates, and if the ADAS-Cog 13 outcome is missing at random, the fixed-effects estimates of the ACA will be valid [30,31]. Therefore, to assess validity of the ACA, we first perform a multivariate logistic regression where the outcome is the missingness incidence of our target covariate and the baseline ADAS-Cog 13 score is a predictor, along with all the baseline covariates. We need not test association with ADAS-Cog 13 scores later than at baseline since our covariate missingness only occurs in baseline covariates.

Since we have five analysis models, we need to check that all of them will be valid by performing multivariate tests for each. Conditional on all other baseline covariates in the analysis models, ADAS-Cog 13 score at baseline was not associated with missingness in CSF measurement (odds ratio of 1.00 with 95% CI 0.98–1.03). For all five analysis models, we find no association between missingness of patient-reported family history and baseline ADAS-Cog 13 (odds ratio of 1.00 with 95% CI 0.98–1.03 for the three models with absolute CSF biomarkers, and odds ratio of 1.01 with 95% CI 0.98–1.04 for the two models with ratio biomarkers). These results suggest that the missingness of covariates is independent of observed outcome values conditional on observed covariates.

The second necessary condition to establish validity of the ACA is that the ADAS-Cog 13 outcome is MAR. Without additional information beyond observed data, it is inherently impossible to statistically determine whether the data is MAR or MNAR. This assessment needs to be made based on clinical expertise and knowledge of the data collection [2,3]. We have searched the literature related to ADNI study design and have not found scientific reasons against the MAR assumption of the outcome. Given the absence of scientific evidence against the MAR assumption, we use MAR as a starting point, which is often sensible to begin with [3]. Significant deviations from the MAR assumption may manifest in

diagnostic plots for the imputed data [3]. These diagnostic plots are used to explore possible evidence of deviations from the MAR assumption. They are discussed in the Results section, wherein we conclude that we have not found evidence against the MAR assumption for the longitudinal ADAS-Cog 13 score. Thus, for this study, we consider the outcome data to be MAR. Taken together, the assessment of the two necessary conditions suggests that the fixed-effects estimates of the ACA will produce valid inference.

Imputation Model

To perform imputation with MICE, we must first decide which variables to include in the imputation models. As per guidelines in [3], the imputation model for a target variable should include every variable in the analysis model and variables with explanatory power for that target variable. Furthermore, variables that predict missingness of the target variable should be included in its imputation model. In our case, [1] provides reason to believe that there are variables beyond those in our analysis model that should be included in the imputation model for CSF biomarkers because of association with their missingness. In particular, they found that family history of AD was associated with missingness of baseline CSF biomarker measurements. We tested this again with the larger data set that has become available since the mentioned article and found consistent results (odds of missingness of CSF at baseline decreases by a factor of 0.46 (95% CI: 0.33–0.66) with reported family history of AD). Our final imputation models for CSF biomarkers, patient-reported family history of AD, and ADAS-Cog 13 scores are as follows. The variable to the left of the tilde is the variable that needs to be imputed, while variables to the right are all covariates. Parentheses indicate time; for example, ADAS13(0) is the ADAS-Cog 13 score at month 0, i.e. baseline.

 $CSF_{bl} \sim AGE + GENDER + EDUCAT + APOE4 + CDRSB_{bl} + FamHist + ADAS13(0)$

 $FamHist \sim AGE + GENDER + EDUCAT + APOE4 + CDRSB_{bl} + CSF_{bl}$

 $ADAS13(t) \sim AGE + GENDER + EDUCAT + APOE4 + CDRSB_{bl} + FamHist + CSF_{bl} + ADAS13(t_1) + \dots + ADAS13(t_k)$

In the imputation model for ADAS-Cog 13 scores, we only impute scores collected at times *t* for which the missingness incidence of ADAS-Cog 13 scores was below the chosen tolerance (50%, 65%, or 75%). As predictors for this imputation model, we include ADAS-Cog 13 at times t_1 through t_k , which are all the times for which the missingness was below the tolerance, excluding time *t* itself (since a target variable should not be included as a predictor in its own imputation model).

It is worth noting that the times t_i through t_k may include times after t – that is, the imputation of the scores at time t may include scores in future times as predictors. However, this should not be taken as a statement about the influence of future measurements on past

measurements. Rather, it is simply a way to improve predictive accuracy of the imputation model on ADAS-Cog 13 scores. Indeed, as discussed earlier, variables were included in the above imputation models based on guidelines in [3], one of which is that variables with explanatory power for target variables should be included in their imputation. This improves predictive accuracy of the target variable, improving the quality of the imputed data sets. To that end, in our model ADAS-Cog 13 scores at time *t* are imputed with those at other times, past or future. Nevertheless, the imputation model could still be designed to include only those times prior to *t*, we verified that this valid alternative approach does not reduce the quality of the imputed data sets or change the qualitative results of this particular study.

In imputing CSF values with Bayesian linear regression, we assume a linear relationship between the CSF values and the predictors of our imputation model. We assess the validity of this model visually with the residual plots of the imputation models. As an example, consider the imputation model for tTau:

 $tTAU_{bl} \sim AGE + GENDER + EDUCAT + APOE4 + CDR SB_{bl} + FamHist + ADAS13(0)$

The top-left plot of Figure 3 shows the residuals from the above model fitted to the observed (i.e. not imputed) values in the tTau model. The residuals are mean-zero with similar variance across fitted values, indicating adherence to a linear relationship. We also checked the distribution of residuals from the imputation models using a histogram and a quantile-quantile plot (top-center and -right of Figure 3). The plots show a reasonable adherence to a normal distribution. We arrive at the same conclusion for the imputation models for the other four CSF biomarkers.

The same diagnostics can be applied to the imputation of the longitudinal outcome. We assessed the validity of the linear model of ADAS-Cog 13 scores with the same procedure we used above for the tTau imputation model. Since there are multiple imputation models for ADAS-Cog 13 (one for each time point), we performed the diagnostics for each one. The bottom row of Figure 3 shows the same plots as with the tTau imputation model for the ADAS-Cog 13 model at month 6. The plots also show a reasonable adherence to a normal distribution, as do those for the scores at all other time points, though these are not shown.

Results

Demographic Data

Table 1 provides baseline demographic information for the 1063 subjects in this study. The proportions of values missing are also provided. Family history of AD and baseline CSF measurements have the most significant missingness – nearly 50%.

Available Case Analysis

The results of the ACA are shown in Figure 4 for the $A\beta_{42}$, pTau, and tTau analysis models, and in Figure 5 for the pTau/A β_{42} and tTau/A β_{42} models, in the columns labeled ACA. Point estimates for predictors are shown along with their confidence intervals. The figures selectively show results for salient features (Age at baseline, APOE ε 4 allele presence, CSF

biomarker level at baseline, and years of education); the complete results can be found in Supplementary Tables 1 and 2. The ACA reduces our available data set to 301 subjects, less than one-third of the number of subjects in our complete data set. However, since we expect the ACA to produce valid inference, our fixed-effects estimates should still be reliable, and we can check that they align with clinical expectations. The effects of visit time and age are positive, which is in line with our expectation that cognitive performance will progressively decline on average over time and with age in the MCI group (higher ADAS-Cog 13 scores reflect greater cognitive impairment). Decreasing A β_{42} and increasing values of the other four biomarkers are also associated with poorer cognitive function [6,7], in line with our estimates. APOE e4 allele presence is associated with lower cognitive performance [21,32], although in the A β_{42} and ratio biomarker models the ACA estimate is not significant.

Imputation Analysis and Imputation Model Diagnostics

Results presented in the *MICE* columns of Figures 4 and 5 are for the case of 50% missingness tolerance and Bayesian linear regression with clipping of CSF values at their technical limits. From Table 2, we see that a tolerance of 50% would allow us to impute ADAS-Cog 13 scores at visit times 0, 6, 12, 24, and 36 months. Before we compare the results of the mixed-effects models between the available and imputed cases, we must first ensure that our imputed data sets look reasonable. Since CSF measurements, family history, and ADAS-Cog 13 scores are our major sources of missingness, we focus on assessing their imputed values.

To ensure that imputed values are reasonable, we look at boxplots of a subset of the 50 imputed data sets for the CSF values, which had significant missingness, alongside boxplots of the observed data. These plots are shown in Figure 6. Since CSF is clipped, we see the same upper and lower bounds for the distribution of the imputed data sets as for the observed data. As mentioned previously, whether or not we clipped did not have an impact on the qualitative results discussed in this section. Importantly, the median, range, and interquartile range of the imputed data sets are comparable to those of the observed data.

We also look at the proportions of subjects with and without patient-reported family history of AD. Figure 7 displays these proportions for the observed data (in black) and for the first 15 imputed data sets (the different shades of blue). In the observed data set, we see that about 60% of subjects reported family history of AD while the other 40% reported no family history of AD. The imputed data sets have proportions comparable to these, giving some evidence that they are reasonable imputations.

Since we have also imputed the time-varying outcome, we analyze the plots shown in Figure 8 of the behavior of imputed ADAS-Cog 13 scores for the tTau model. We use months 12 and 36 as illustrative examples. The top two boxplots in Figure 8 show that the median and interquartile range of the imputed data sets are comparable with those of the observed data for both months 12 and 36. However, the range of the imputed data sets exceeds the range of the observed data sets. Indeed, ADAS-Cog 13 scores are values between 0 and 85, but the imputed values are not restricted to this possible range. It is not always necessary that individual values be realistic, so long as sample estimates, like mean and standard deviation, are preserved in the imputed data sets [33]. Still, to assess

robustness we performed our analyses again using predictive mean matching for ADAS-Cog 13 scores. The results regarding estimates and confidence intervals are nearly identical to those obtained by using Bayesian linear regression, so our interpretations remain the same. The bottom two plots in Figure 8 show the behavior with predictive mean matching of the imputed ADAS-Cog 13 scores at 12 and 36 months, which now take on only values between 0 and 85. Though not shown in the figure, these are also nearly identical to results obtained when using yet another imputation method: Bayesian linear regression followed by clipping imputed values between 0 and 85.

Figure 8 also gives us some insight into the MAR assumption for the longitudinal outcome of ADAS-Cog 13 score. As mentioned previously, without additional information beyond observed data, it is inherently impossible to statistically determine whether the data is MAR or MNAR [2,3]. However, significant deviations from the MAR assumption may manifest in diagnostic plots for the imputed data, such as Figure 8 of our study, in which the differences between the distributions of the observed data (blue) and imputed data (red) could end up being very different [3]. It is encouraging, in our case, that those distributions appear very similar, as displayed in Figure 8. Thus we have not found evidence against the MAR assumption for the longitudinal ADAS-Cog 13 score.

To be able to compare ACA with MICE results, we must ensure that the estimates do not deviate far from those of the ACA, as this would indicate bias in our imputation model. We can get a rough sense of whether the estimates are reasonably close to each other by observing whether their confidence intervals overlap. All of our estimates are reasonably close to the ACA estimates by this standard, as can be seen in Figures 4 and 5, and further verified with Supplementary Tables 1 and 2.

We see that, for the $A\beta_{42}$ and $pTau/A\beta_{42}$ models, the confidence intervals surrounding the APOE e4 estimate has become tighter, and now indicates a significant effect. This is not surprising given that with imputation we are able to recover the remaining two-thirds of the subjects, drastically increasing our available data set. As mentioned earlier, this is in line with clinical expectations [21,32]. For the tTau/A β_{42} model, the estimate has become tighter after imputation, but not enough to indicate a significant effect. For the pTau and tTau models, APOE e4 is significant already in the ACA, with confidence intervals becoming even tighter after imputation. Also worth noting are the tighter confidence intervals around the Age estimates, which become significant after imputation for the analysis models with pTau/A β_{42} and tTau/A β_{42} , as well as the education estimates, which become significant across all five analysis models after imputation. There are some studies which are in line with this finding regarding education [34,35], while others do not find any significant association between ADAS-Cog scores and education level [14].

When performing these same analyses for tolerances 65%, and 75%, we draw the same conclusions described above for all of the results. Overall, these results demonstrate the increased statistical efficiency of the imputed data analysis over the ACA.

Discussion

In this study we illustrated the importance of imputation in a longitudinal analysis of the ADNI data set. We demonstrated that MICE provides improvement in statistical efficiency compared to the ACA while still producing valid inference. Of note, across the five models we see tighter confidence intervals around the effect of APOE e4 allele presence and education level. For all models, we see that the effect of education level becomes significant after imputation. These significant effects can be interpreted as a strong association between education level and average ADAS-Cog 13 scores across visits, controlling for other baseline features.

We verified the normality assumptions of the mixed-effects model by showing that random intercept estimates appear normally distributed. We assessed the validity of a linear relationship between baseline CSF measurements and other baseline features in our imputation model. Our residual plot shows that this relationship is reasonable, and the residuals from the imputation model appear normally distributed.

We assessed robustness of results to various tolerances of missingness: 50%, 65%, and 75%. Estimates and confidence intervals were similar across all tolerances, leading to the same interpretations in the Results section for all cases.

While we reported results for imputation of numerical variables with Bayesian linear regression, we found that results held when using predictive mean matching. In particular, they were robust to the choice between Bayesian linear regression and predictive mean matching for the ADAS-Cog 13 scores, which has a range of values between 0 and 85 that is not captured by Bayesian linear regression as it is in pmm. The results were also nearly identical if we performed Bayesian linear regression followed by clipping imputations between 0 and 85. This robustness is not surprising since, as discussed in the Results section, the purpose of imputation is not necessarily to draw realistic values, but to preserve sample-level properties like mean and standard deviation. Results were also nearly identical whether or not we clipped CSF values at their technical limits.

Our study demonstrates the importance of accounting for missing data in the ADNI data set. When deciding to perform imputation, care should be taken in choosing the approach, as an invalid one can compromise the statistical analyses. The mechanism of missing data generation, choice of imputation procedure and imputation methods, and assumptions of the imputation procedure should all be scrutinized. While we chose MICE as our imputation procedure for this study, it would be interesting to compare the results of the mixed-effects model under other appropriate methods, e.g. random forest imputation. Sometimes, the parametric assumptions of imputation models may be violated after model diagnostics. In those situations, other more robust and/or nonparametric models should be employed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Funding

This work is supported by NIH grants NS102324, AG072979, AG066597, and AG062418.

Data Availability Statement

The data supporting the findings of this study are available at adni.loni.usc.edu upon acceptance of application for ADNI data.

References

- Lo RY, Jagust WJ (2012) Predicting missing biomarker data in a longitudinal study of Alzheimer disease. Neurology 78, 1376–1382. [PubMed: 22491869]
- Horton NJ, Kleinman KP (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat 61, 79–90. [PubMed: 17401454]
- 3. van Buuren S (2021) Flexible Imputation of Missing Data, Second Edition 2nd Edition St, Chapman and Hall/CRC.
- 4. White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 30, 377–399. [PubMed: 21225900]
- 5. Rubin DB (1976) Inference and Missing Data. Biometrika 63, 581.
- 6. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Jack CR, Jagust W, Liu E, Morris JC, Petersen RC, Saykin AJ, Schmidt ME, Shaw L, Siuciak JA, Soares H, Toga AW, Trojanowski JQ (2012) The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement 8.
- Molinuevo JL, Ayton S, Batrla R, Bednar MM, Bittner T, Cummings J, Fagan AM, Hampel H, Mielke MM, Mikulskis A, O'Bryant S, Scheltens P, Sevigny J, Shaw LM, Soares HD, Tong G, Trojanowski JQ, Zetterberg H, Blennow K (2018) Current state of Alzheimer's fluid biomarkers. Acta Neuropathol 136.
- Toledo JB, Xie SX, Trojanowski JQ, Shaw LM (2013) Longitudinal change in CSF Tau and Aβ biomarkers for up to 48 months in ADNI. Acta Neuropathol 126, 659–670. [PubMed: 23812320]
- 9. Azur MJ, Stuart EA, Frangakis C, Leaf PJ (2011) Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 20, 40–49. [PubMed: 21499542]
- Campos S, Pizarro L, Valle C, Gray KR, Rueckert D, Allende H (2015) Evaluating imputation techniques for missing data in ADNI: A patient classification study. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 9423, 3–10.
- Xiang S, Yuan L, Fan W, Wang Y, Thompson PM, Ye J (2014) Bi-level multi-source learning for heterogeneous block-wise missing data. Neuroimage 102 Pt 1, 192–206. [PubMed: 23988272]
- Dodge HH, Zhu J, Woltjer R, Nelson PT, Bennett DA, Cairns NJ, Fardo DW, Kaye JA, Lyons DE, Mattek N, Schneider JA, Silbert LC, Xiong C, Yu L, Schmitt FA, Kryscio RJ, Abner EL (2017) Risk of incident clinical diagnosis of Alzheimer's disease-type dementia attributable to pathology-confirmed vascular disease. Alzheimers Dement 13, 613–623. [PubMed: 28017827]
- 13. Mohs RC, Knopman D, Petersen RC, Ferris SH, Ernesto C; Grundman M, Sano M, Bieliauskas L, Geldmacher D, Clark C, Thai LJ (1997) Development of Cognitive Instruments for Use in Clinical Trials of Antidementia Drugs: Additions to the Alzheimer's Disease Assessment Scale That Broadens Its Scope. Alzheimer Disease & Associated Disorders 11, 13–21. [PubMed: 9194962]

- 14. Pyo G, Elble RJ, Ala T, Markwell SJ (2006) The Characteristics of Patients With Uncertainty/Mild Cognitive Impairment on the Alzheimer Disease Assessment Scale-Cognitive Subscale. Alzheimer Dis Assoc Disord 20, 16–22. [PubMed: 16493231]
- Zec RF, Landreth ES, Vicari SK, Feldman E, Belman J, Andrise A, Robbs R, Kumar V, Becker R (1992) Alzheimer Disease Assessment Scale: Useful for Both Early Detection and Staging of Dementia of the Alzheimer Type. Alzheimer Dis Assoc Disord 6, 89–102. [PubMed: 1389084]
- 16. Kaufman DM, Geyer HL, Milstein MJ, Rosengard J (2023) Kaufman's Clinical Neurology for Psychiatrists 9th Edition Elsevier.
- 17. Raghavan N, Samtani MN, Farnum M, Yang E, Novak G, Grundman M, Narayan V, DiBernardo A (2013) The ADAS-Cog revisited: Novel composite scales based on ADAS-Cog to improve efficiency in MCI and early AD trials. Alzheimers Dement 9, S21–S31. [PubMed: 23127469]
- Tzeng RC, Yang YW, Hsu KC, Chang HT, Chiu PY (2022) Sum of boxes of the clinical dementia rating scale highly predicts conversion or reversion in predementia stages. Front. Aging Neurosci, 14.
- Cedarbaum JM, Jaros M, Hernandez C, Coley N, Andrieu S, Grundman M, Vellas B (2013) Rationale for use of the Clinical Dementia Rating Sum of Boxes as a primary outcome measure for Alzheimer's disease clinical trials. Alzheimers Dement 9, S45–S55. [PubMed: 22658286]
- 20. Coley N, Andrieu S, Jaros M, Weiner M, Cedarbaum J, Vellas B (2011) Suitability of the Clinical Dementia Rating – Sum of Boxes as a single primary endpoint for Alzheimer's disease trials. Alzheimers Dement 7, 602–610.e2. [PubMed: 21745761]
- 21. Farlow MR, He Y, Tekin S, Xu J, Lane R, Charles HC (2004) Impact of APOE in mild cognitive impairment. Neurology 63, 1898–1902. [PubMed: 15557508]
- 22. Buckley RF, Mormino EC, Chhatwal J, Schultz AP, Rabin JS, Rentz DM, Acar D, Properzi MJ, Dumurgier J, Jacobs H, Gomez-Isla T, Johnson KA, Sperling RA, Hanseeuw BJ (2019) Associations between baseline amyloid, sex, and APOE on subsequent tau accumulation in cerebrospinal fluid. Neurobiol Aging 78, 178–185. [PubMed: 30947113]
- 23. Hansson O, Seibyl J, Stomrud E, Zetterberg H, Trojanowski JQ, Bittner T, Lifke V, Corradini V, Eichenlaub U, Batrla R, Buck K, Zink K, Rabe C, Blennow K, Shaw LM (2018) CSF biomarkers of Alzheimer's disease concord with amyloid-β PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts. Alzheimers Dement 14, 1470–1481. [PubMed: 29499171]
- 24. Suárez-Calvet M, Capell A, Araque Caballero MÁ, Morenas-Rodríguez E, Fellerer K, Franzmeier N, Kleinberger G, Eren E, Deming Y, Piccio L, Karch CM, Cruchaga C, Paumier K, Bateman RJ, Fagan AM, Morris JC, Levin J, Danek A, Jucker M, Masters CL, Rossor MN, Ringman JM, Shaw LM, Trojanowski JQ, Weiner M, Ewers M, Haass C (2018) CSF progranulin increases in the course of Alzheimer's disease and is associated with sTREM2, neurodegeneration and cognitive decline. EMBO Mol Med 10.
- 25. Bodner TE (2008) What Improves with Increased Missing Data Imputations? Struct Equ Model A Multidiscip J 15, 651–675.
- 26. Alonso A, Litière S, Laenen A (2010) A Note on the Indeterminacy of the Random-Effects Distribution in Hierarchical Models. The American Statistician 64, 318–324.
- Schielzeth H, Dingemanse NJ, Nakagawa S, Westneat DF, Allegue H, Teplitsky C, Réale D, Dochtermann NA, Garamszegi LZ, Araya-Ajoy YG (2020) Robustness of linear mixed-effects models to violations of distributional assumptions. Methods Ecol. Evol 11, 1141–1152.
- Bell A, Fairbrother M, Jones K (2019) Fixed and random effects models: making an informed choice. Quality & Quantity 53, 1051–1074.
- Knief U, Forstmeier W (2021) Violating the normality assumption may be the lesser of two evils. Behav. Res. Methods 53, 2576–2590. [PubMed: 33963496]
- 30. White IR, Carlin JB (2010) Bias and efficiency of multiple imputation compared with completecase analysis for missing covariate values. Stat Med 29, 2920–2931. [PubMed: 20842622]
- Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38, 963–974. [PubMed: 7168798]

- 32. Kennedy RE, Cutter GR, Schneider LS (2014) Effect of APOE genotype status on targeted clinical trials outcomes and efficiency in dementia and mild cognitive impairment resulting from Alzheimer's disease. Alzheimers Dement 10, 349–359. [PubMed: 23712001]
- 33. von Hippel PT (2017) Should a Normal Imputation Model Be Modified to Impute Skewed Variables? Sociol Methods Res 42, 105–138.
- Doraiswamy PM, Krishen A, Stallone F, Martin WL, Potts NL, Metz A, DeVeaugh-Geiss J (1995) Cognitive performance on the Alzheimer's Disease Assessment Scale: Effect of education. Neurology 45, 1980–1984. [PubMed: 7501145]
- 35. Schultz RR, Siviero MO, Bertolucci PH (2001) The cognitive subscale of the "Alzheimer's Disease Assessment Scale" in a Brazilian Sample. Braz J Med Biol Res 34, 1295–1302. [PubMed: 11593304]



Figure 1:

Mean (left) and standard deviation (right) of imputed values for different variables across iterations for the $A\beta_{42}$ model. Each line (represented by a different color) is one of 50 imputations. Each of the three rows is for a different variable being imputed. Convergence is achieved when the lines stabilize at some mean and standard deviation region, and so long as there is no observable trend between the 50 imputation lines.



Figure 2:

Distributions of random intercepts from the ACA for each of the five models: $A\beta_{42}$ (top left), pTau (top right), tTau (middle left), log-transformed pTau/A β_{42} (middle right), and log-transformed tTau/A β_{42} (bottom).



Figure 3:

(top left) Residual plot of linear model when observed tTau (not imputed) is the dependent variable and the independent variables are baseline features (Age, Sex, Education, APOE4, Family History of AD, baseline CDR-SB, and baseline ADAS-Cog 13). The x-axis (the fitted values) is estimated tTau values from the linear model. (top center) Histogram of standardized residual values from the same model. (top right) Quantile-quantile plot of standardized residual values from the same model. Theoretical quantiles are for the standard normal distribution. (bottom row) The same set of plots as in the top row, but for the linear model when observed ADAS-Cog 13 score at month 6 (not imputed) is the dependent variable and the independent variables are baseline features (Age, Sex, Education, APOE4, Family History of AD, baseline CDR-SB, and tTau) and ADAS-Cog 13 scores at all other times for which missingness was below 50% tolerance, with the exception of month 6 (since this is the month considered in the dependent variable).



Figure 4:

Comparison of ACA and MICE estimates with 95% confidence intervals for the effect of Age at baseline, APOE $\varepsilon 4$ allele presence, CSF level at baseline, and level of education. (top row) A β_{42} analysis model, (middle row) pTau analysis model, and (bottom row) tTau analysis model. Black dots represent the estimates for the quantities in the subplot titles. Confidence intervals which cross the dashed line at 0 suggest that the estimate is not significant.



Figure 5:

Comparison of ACA and MICE estimates with 95% confidence intervals for the effect of Age at baseline, APOE $\varepsilon 4$ allele presence, CSF level at baseline, and level of education. (top row) pTau/A β_{42} analysis model and (bottom row) tTau/A β_{42} analysis model. Black dots represent the estimates for the quantities in the subplot titles. Confidence intervals which cross the dashed line at 0 suggest that the estimate is not significant.



Figure 6:

Boxplots of observed and imputed CSF values for each of the five models for the first 15 imputations: $A\beta_{42}$ (top left), pTau (top right), tTau (middle left), log-transformed pTau/ $A\beta_{42}$ (middle right), and log-transformed tTau/ $A\beta_{42}$ (bottom). Observed data are shown in blue and imputed data sets are in red. Only the first 15 of the 50 imputed data sets are plotted.



Figure 7:

Marginal proportion of presence (TRUE) or absence (FALSE) of patient-reported family history of AD. Black bars are observed data while blue bars are each different imputed data sets from the tTau model. Only the first 15 of the 50 imputed data sets are plotted. Plots were generated by the propplot function obtained from GitHub.



Figure 8:

Boxplots of observed and imputed ADAS-Cog 13 values for the tTau models for the first 15 imputations at 12 months (left) and 36 months (right). (top) data imputed by Bayesian linear regression. (bottom) data imputed by predictive mean matching. Observed data are shown in blue and imputed data sets in red. Only the first 15 of the 50 imputed data sets are plotted.

Table 1:

Demographic analysis of baseline information for MCI patients.

		% missing (<i>n</i> missing)
Age at baseline ^a	72.9 (7.6)	0.4 (4)
Sex (male %)	58.9 %	0 (0)
APOE e4 (%)	49.7 %	4.6 (49)
Education (years) b	16 (14–18)	0 (0)
CDR-SB at baseline ^a	1.53 (0.92)	0 (0)
Family history AD (%)	59.9 %	48.4 (515)
Aβ 42 ^b	836.8 (609.9 - 1308.0)	
pTau ^b	24.0 (17.4 - 34.3)	
tTau ^b	256.5 (192.0 - 343.4)	42.0 (446)
$\log(pTau/A\beta_{42}) b$	-3.50 (-4.32.9)	
$\log(tTau/A\beta_{42}) b$	-1.1 (-1.90.6)	

^aMean (Standard deviation) n=1063

^bMedian (1st and 3rd quartiles)

Table 2:

Percentage of ADAS-Cog 13 scores missing for each visit time in the protocol within the first 72 months. For all other months, missingness prevalence is greater than 85%.

Visit time (months from baseline)	% ADAS-Cog 13 missing
0	0.56 %
6	24.37 %
12	16.09 %
18	69.24 %
24	32.36 %
36	44.40 %
48	62.65 %
60	77.42 %
72	83.54 %