Available online at www.sciencedirect.com

SCIENCE DIRECT®

Psychiatry Research 129 (2004) 209–215

PSYCHIATRY RESEARCH

www.elsevier.com/locate/psychres

Estimating treatment effects from longitudinal clinical trial data with missing values: comparative analyses using different methods

Patricia R. Houck[a,*], Sati Mazumdar[b], Tulay Koru-Sengul[c], Gong Tang[b], Benoit H. Mulsant[a,d], Bruce G. Pollock[a], Charles F. Reynolds III[a]

[a]*Department of Psychiatry, University of Pittsburgh, Western Psychiatric Institute and Clinic, UPMC Health System, Thomas Detre Hall, 3811 O'Hara Street, Pittsburgh, PA 15213-2593, USA*

[b]*Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15260, USA*

[c]*Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA*

[d]*Geriatric Research, Education, and Clinical Center, VA Pittsburgh Health Care System, Pittsburgh, PA 15213, USA*

Received 12 March 2004; received in revised form 30 June 2004; accepted 2 August 2004

Abstract

The selection of a method for estimating treatment effects in an intent-to-treat analysis from clinical trial data with missing values often depends on the field of practice. The last observation carried forward (LOCF) analysis assumes that the responses do not change after dropout. Such an assumption is often unrealistic. Analysis with completers only requires that missing values occur completely at random (MCAR). Ignorable maximum likelihood (IML) and multiple imputation (MI) methods require that data are missing at random (MAR). We applied these four methods to a randomized clinical trial comparing anti-depressant effects in an elderly depressed group of patients using a mixed model to describe the course of the treatment effects. Results from an explanatory approach showed a significant difference between the treatments using LOCF and IML methods. Statistical tests indicate violation of the MCAR assumption favoring the flexible IML and MI methods. IML and MI methods were repeated under the pragmatic approach, using data collected after termination of protocol treatment and compared with previously reported results using piecewise splines and rescue (treatment adjustment) pragmatic analysis. No significant treatment differences were found. We conclude that attention to the missing-data mechanism should be an integral part in analysis of clinical trial data.

Keywords: Intent-to-treat; Missing data; Mixed model; Multiple imputation; Maximum likelihood; Depression

1. Introduction

Incomplete data sets due to attrition or noncompliance present a major and persistent challenge to

* Corresponding author. Tel.: +1 412 246 6402; fax: +1 412 246 5300.

E-mail address: houckpr@upmc.edu (P.R. Houck).

0165-1781/$ - see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.psychres.2004.08.001

clinical trials research. Though the importance of examining and accounting for the different patterns and reasons for the missing data is recognized by researchers, reports from clinical research often fall short of addressing these issues. We aim here to illustrate this important point through a sensitivity analysis of a longitudinal data set from a psychiatric clinical trial comparing treatment effects.

According to Little and Rubin (2002), the missing-data mechanism is called "missing completely at random" (MCAR) when the missingness does not depend on either the observed values or the missing values (example, missing appointment for a snow day); "missing at random" (MAR) when the missingness may depend on the observed values but not the missing values (example, patient stays stable for a few weeks and decides to skip an appointment for assessment). Otherwise, the missingness that depends on the unobserved missing values is called "not missing at random" (NMAR; example, the patient feels bad and cannot come for assessment). Recognition of the underlying missing-data mechanism is important in selecting an appropriate statistical technique for analysis, since methods that disregard the missing-data process often lead to biased and inefficient estimates.

Mixed models for longitudinal data (Laird and Ware, 1982) from clinical trials were introduced into the psychiatric literature by Gibbons et al. (1993). The maximum likelihood method (ML; Schafer and Graham, 2002) has been shown to do well in estimating parameters in the presence of dropout bias in the context of different models (Verbeke and Molenberghs, 2000; Mallinckrodt et al., 2003; Hogan et al., 2004; Gueorguieva and Krystal, 2004). Multiple imputation (MI; Rubin, 1996; Lavori et al., 1995; Allison, 2001; Schafer and Graham, 2002) is another option for analysis of missing data where multiple data sets are created by imputing missing values utilizing information from the study. Curry et al. (2003) demonstrate the use of MI in a randomized clinical trial to remove bias that is due to differential attrition between the treatment groups.

In psychiatric clinical trials, traditional but ad hoc methods such as last observation carried forward (LOCF) and analysis with a completer sample are commonly reported without any validation of their underlying assumptions about the missing-data process. LOCF analysis uses all subjects and imputes the missing values with the last observed value, a method that assumes that the outcomes would not have changed from the last observed value. Analysis with a completer sample, usually referred to as a completer analysis or complete-case analysis, includes only those subjects who complete the trial and results in loss of information by the exclusion of subjects who do not complete the trial. The completer analysis uses a chosen set and the modeling does not depend on the missing data. Neither of these two methods is desirable in clinical trials that have missing data due to attrition (Liu and Gould, 2002). A completer analysis is often biased when the stringent MCAR assumption is not satisfied and less efficient than the maximum likelihood method even when this assumption holds. The assumption of constant profile after dropout for the LOCF analysis is neither scientifically nor statistically sound.

Recent methods such as ignorable maximum likelihood (IML) and MI that require less stringent assumption, such as MAR, are more robust and favored by the statistical community (Schafer and Graham, 2002; Collins et al., 2001). They are increasingly used in clinical research due to their availability in software packages (Horton and Lipsitz, 2001). When the data are MAR and the missing-data process does not involve parameters of the complete data model, then the missing-data process is ignorable such that the ML method without modeling the missing-data process yields consistent and asymptotically efficient estimates (Rubin, 1976). Under such circumstance, the ML method is usually referred to as the IML method.

The intent-to-treat (IT) analysis based on the realistic principle of using all data as randomized (Lavori, 1992) provides unbiased estimates of the clinical effectiveness of treatments (Lachin, 2000). It has been a central goal of statisticians in the field to reinforce the IT analysis as a standard methodology (Lavori, 1992; Sheiner and Rubin, 1995; Mazumdar et al., 1999) for clinical trials. In an IT analysis, the common practice is to analyze only the data that are obtained under the blinded condition (on-randomized treatment), which can be referred to as an "efficacy" or "explanatory" analysis. An alternative approach, which accounts for changes in treatment practice, is called an "effectiveness" or a "pragmatic" analysis (Heyting et al., 1992). The pragmatic analysis includes all observations, from both "on-" and "off-" randomized treatment periods, after change or dropout in initial treatment assignment. If dropouts are due to non-adherence to treatment protocols, side effects or other treatment-related reasons, ignoring the missing data violates the strict IT principles of measuring all patient outcomes regardless of protocol adherence or no matter what treatment patients received (Lavori, 1992). A pragmatic approach is desirable to decide the effect of the treatment in all subjects, including those subjects who are unable to tolerate the treatment or change treatment for other reasons.

In the present article, both IML and MI methods used in a mixed model are compared to the LOCF analysis and the completer sample analysis in an explanatory approach. The IML and MI are repeated under the pragmatic approach and compared with two other pragmatic approaches recently presented by our group, piecewise spline model (Mazumdar et al., 2002) and rescue regression analysis (Houck et al., 2003). This sensitivity analysis compares the final point estimates from several statistical approaches to arrive at a reliable conclusion regarding treatment effects.

2. Methods

Several methods to estimate treatment effects are applied to depression severity data measured by the 17-item Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960) from a double-blind randomized study comparing nortriptyline and paroxetine over 12 weeks (Mulsant et al., 2001). The sample comprised 116 geriatric inpatients and outpatients at Western Psychiatric Institute and Clinic who met DSM-IV criteria (First et al., 1997) for a major depressive episode, without psychotic features or a past history of bipolar or schizoaffective disorder. An entry HRSD score of 15 or above and a Mini-Mental State Examination (MMSE; Folstein et al., 1976) score of 15 or above were required for study entry. Subjects with contraindication to treatment with either medication were excluded.

Qualified patients who consented to participate were randomly assigned to either nortriptyline (n=54) or paroxetine (n=62) and dosing was titrated according to protocol. Depression severity was assessed weekly with the HRSD for 12 weeks. Patients who failed to tolerate the randomized treatment or did not show satisfactory progress were terminated from the randomized portion of the study and were treated openly with the other protocol drug or augmented or replaced by other drugs, determined by the treating physician. Monitoring depression severity was continued during the open treatment. For simplicity, we call the data from the portion of the trial with randomized treatments as "on-randomized treatment data" and the data after the dropout and change of treatment portion of the study as "off-randomized treatment data."

A total of 58 of the 116 subjects (50%) completed the 12-week randomized treatment. About 30% of the subjects were terminated from the randomized protocol by week 6 of the trial. Forty-five percent of the nortriptyline-treated subjects and 55% of the paroxetine-treated subjects completed the trial. Side effects led to dropout from the randomized treatment in 28 subjects; nine subjects withdrew consent; five were noncompliant and the rest withdrew for other reasons including new medical conditions and reasons related to subject burden (i.e., inability to arrange for transportation).

To test the MCAR assumption, we applied Little's test (Little, 1988), which divides the sample into groups based on the patterns of missingness for the study outcome. Test statistics are based on the pattern-specific means and the pooled estimates of the population mean and covariance. A step-by-step procedure for calculating the test statistics is available along with SAS coding (Fairclough, 2002). A significant result indicates that the missing data are not MCAR. There are statistical methods available to analyze data under the assumption of MAR or NMAR, but no statistical test is currently available to test those two missingness mechanisms. Generally, MAR is assumed when it is believed that subjects drop out because of the observed history of their response values. In addition to Little's test, we also performed a logistic model using HRSD score regressed on dropout for the subsequent week for weeks 3, 6, and 9. A significant result indicates that the missing data are not MCAR.

For the LOCF analysis, we used an analysis of covariance, entering the final HRSD score as the dependent variable and the baseline HRSD score as a covariate. The analysis with the completer sample uses all time points in a mixed model repeated measures approach with random intercept after excluding subjects who had intermittent missing data or dropped out early.

The IML method was performed using SAS PROC MIXED (SAS, 2000; Brown and Prescott, 1999). It provides unbiased and efficient estimates, robust to deviations from MAR, and accommodates intermittently missing values. The MI method, based upon Markov Chain Monte Carlo (MCMC), is a parametric approach for creating multiple imputed data sets that provides some solution to the complex problem of handling both dropout and intermittent missing values (Allison, 2001). It is robust to minor departures from multivariate normality (Schafer and Graham, 2002). SAS PROC MI was used to generate 10 complete data sets by the MCMC method of imputation, and PROC MIANALYZE was used to combine the results for the final inference (Yuan, 2000). We used a linear mixed model, which included a random intercept and slope, on the repeated measures of depression severity over time for both IML and MI methods. Since most of the movement was in the early part of the trial, week was transformed via natural logarithm before statistical analyses (Gibbons et al., 1993).

The IML and MI methods were repeated under the pragmatic approach, using both “on-” and “off-” randomized data, and compared with previously published results using a piecewise spline model (Mazumdar et al., 2002) and rescue regression analysis (Houck et al., 2003). The piecewise spline model fits two lines joined at the dropout point. The rescue regression analysis incorporates the week of dropout in the analysis as a covariate.

3. Results

Little’s test was calculated in two different ways: one for the entire sample and one for each treatment group. Based on the test statistics for testing the MCAR of the entire sample, it was found that the missing data mechanism is not MCAR (χ^2=1411, *df*=157, *P*<0.001). We also performed this test in each treatment group and arrived at the same conclusions (nortriptyline, χ^2=506, *df*=78, *P*<0.001; paroxetine, χ^2=1165, *df*=129, *P*<0.001). These results clearly indicate that data are not MCAR. The simpler logistic model using HRSD predicting subsequent week dropout was significant at both 6 and 9 weeks (χ^2=4.16, *df*=1, *P*<0.05 and χ^2=5.69, *df*=1, *P*<0.05, respectively), implying similar results.

The table and the figure report the estimated 12-week depression scores for each of the treatment arms derived from the different methods. The estimates using only the on-randomized treatment data (explan-

Table 1
Point estimates from different analytic methods (mean, S.E.) at 12 weeks by treatment group

(a) Explanatory approach: on-randomized treatment data

	LOCF using ANCOVA	Repeated-measures using completers sample	Mixed model (MM) with IML	Mixed model (MM) with MI
Nortriptyline	9.54 (0.78)	7.13 (0.75)	6.41 (0.64)	6.92 (0.64)
Paroxetine	11.72 (0.72)	7.37 (0.69)	8.81 (0.60)	8.36 (0.60)
Mean difference (S.E.)	2.18* (1.06)	0.24 (1.01)	2.40* (0.88)	1.43 (0.80)

(b) Pragmatic approach: on- and off-randomized treatment data

	Mixed model (MM) with IML	Mixed model (MM) with MI	Piecewise spline[a]	Rescue regression[b]
Nortriptyline	6.75 (0.65)	7.09 (0.63)	5.80 (2.25)	6.55 (0.68)
Paroxetine	8.39 (0.61)	8.04 (0.59)	5.68 (1.41)	6.03 (0.19)
Mean difference (S.E.)	1.65 (0.89)	0.95 (0.82)	0.12 (2.57)	0.52 (0.66)

[a] From Mazumdar et al. (2002).

[b] From Houck et al. (2003).

* Significant treatment difference at *P*<0.05.

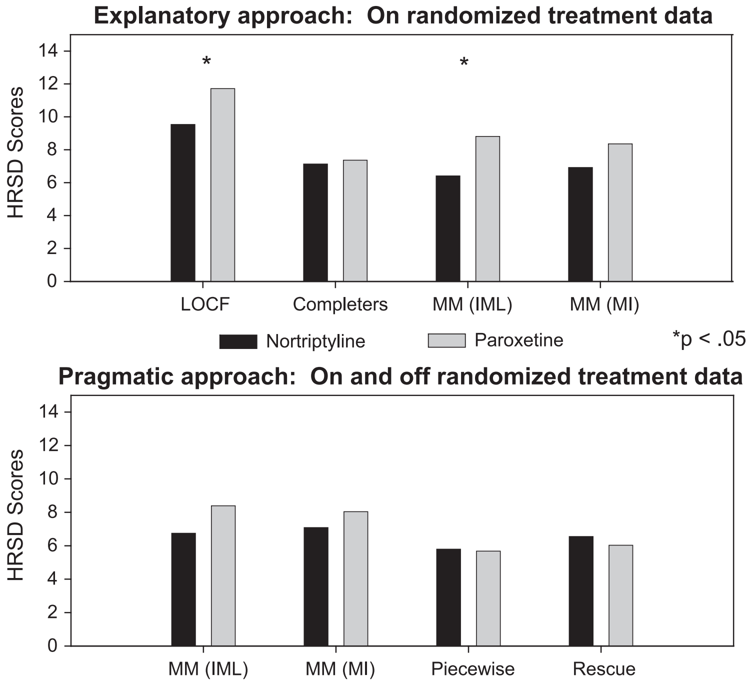

Fig. 1. Final 12-week point estimates using different analytic methods. Top panel illustrates the use of on-randomized treatment data while the lower panel illustrates the use of both on- and off-randomized treatment data. Significant differences at $P<0.05$ between the point estimates are also indicated.

atory approach) are reported in Table 1a and the top panel of Fig. 1. The estimates using both on- and off-randomized treatment data (pragmatic approach) are reported in Table 1b and the bottom panel of Fig. 1.

A significant treatment effect using LOCF and IML was found under the explanatory approach. The IML and MI methods did not show a significant treatment effect under the pragmatic approach and agree with the previously reported results. The largest estimates are seen with the LOCF analysis. This may be due to high last-observed values for the early dropouts that may have been obtained early in the treatment before much improvement.

4. Discussion

The significant MCAR test indicates that the assumption of the analysis with the completer sample is not met. However, we presented the results from this analyses and LOCF to show how biased results can be obtained and erroneous conclusions can be reached if analytic assumptions are violated. The code for Little’s test entails some advanced level programming; however, a simpler approach, such as a logistic model using HRSD score to predict outcome, can be done with standard software.

The LOCF analysis estimated the worse outcome and a significant difference between the treatments. The IML under the explanatory condition also found a significant treatment effect. None of the methods found a significant treatment effect under the pragmatic approach where both on- and off-randomized treatment data were used.

The IML and MI methods gave different results when only the on-randomized treatment data were used. These two methods are expected to agree for a large sample and the difference seen here could be due to the limited sample size. Sinharay et al. (2001) point out that adding important covariates in the multiple imputation step improves the chance that the MAR assumption is plausible, but this also introduces more parameters and thus more variability. These authors suggest that even if the data are truly NMAR, MI may give reasonable estimates with the use of appropriate covariates. The rescue regression analysis adds a covariate to the model and thus might have helped in conforming to the MAR mechanism pattern.

LOCF analysis requires a strong and unrealistic assumption, i.e., no change in the profile after the last observed value. The completer analysis requires the stringent MCAR assumption and is not efficient even when this assumption holds. IML and MI are superior to the completer analysis in that they are valid and efficient under the more flexible assumption of MAR. They are also more efficient than the completer analysis even when MCAR holds. Furthermore, the current availability of flexible software for analyzing longitudinal data under ignorable missingness allows researchers to implement the IML and MI methods comfortably in clinical trials.

The assumption of MAR is not testable; the observed data do not supply information about whether this MAR assumption on the missing-data mechanism is correct. The possibility of the data being NMAR cannot be ruled out, but instead of blindly shifting to NMAR analyses with the current computing difficulties, we support the notion of doing sensitivity analyses. When the missing data are indeed NMAR, methods for modeling the dropout mechanism should be incorporated (Little and Rubin, 2002; Hedeker and Gibbons, 1997). Sensitivity analysis addressing the impact of alternative assumptions or models on the NMAR mechanism should be conducted before drawing conclusion. Realizing that there is no generally “correct” model for an NMAR mechanism, Schafer and Graham (2002) regard IML and MI as the “practical state of the art” for handling missing data.

The use of traditional methods such as LOCF and completer analysis should be avoided in psychiatric clinical trials, especially when the missing data are not MCAR. We have illustrated the use of two readily available methods that are appropriate for data under MAR. We believe that this illustration demonstrates the need to examine the missing data pattern, to the extent possible, before selecting an analytic method.

Analyses of clinical trials should also include the examination of off-treatment data in a pragmatic approach. This will give an overall treatment effect resembling real-world clinical practice.

Acknowledgments

The research reported was supported in part by the Mental Health Intervention Research Center for the Study of Late Life Mood Disorders (MH52247) and the Mental Health Intervention Research Center for Mood and Anxiety Disorder (MH30915). Drs. Mulsant, Pollock, and Reynolds have received research support and honoraria from GlaxoSmithKline and other manufacturers of antidepressant medications.

References

Allison, P.D., 2001. Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Sage, Thousand Oaks, CA.

Brown, H., Prescott, R., 1999. Applied Mixed Models in Medicine. John Wiley and Sons, New York.

Collins, L.M., Schafer, J.L., Kam, C.M., 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods 6 (4), 330–351.

Curry, S.J., Ludman, E.J., Grothaus, L.C., Donovan, D., Kim, E., 2003. A randomized trial of a brief primary-care-based intervention for reducing at-risk drinking practices. Health Psychology 22 (2), 156–165.

Fairclough, D.L., 2002. Design and Analysis of Quality of Life Studies in Clinical Trial. CRC Press, Boca Raton, FL.

First, M.B., Gibbon, M., Spitzer, R.L., Williams, J.B.W., 1997. Structured Clinical Interview for DSM-IV Axis I Disorders—SCID I: Clinician Version, Administration Booklet. American Psychiatric Press, Washington, DC.

Folstein, M., Folstein, S., McHugh, P., 1976. Mini-mental state. Journal of Psychiatric Research 12, 189–198.

Gibbons, R.D., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H.C., Greenhouse, J.B., Shea, T., Imber, S.D., Sotsky, S.M., Watkins, J.T., 1993. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. Archives of General Psychiatry 50, 739–750.

Gueorguieva, R., Krystal, J.H., 2004. Move over ANOVA. Archives of General Psychiatry 61, 310–317.

Hamilton, M., 1960. A rating scale for depression. Journal of Neurology, Neurosurgery and Psychiatry 23, 56–62.

Hedeker, D., Gibbons, R.D., 1997. Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychological Methods 2, 64–78.

Heyting, A., Tolboom, J.T.B.M., Essers, J.G.A., 1992. Statistical handling of drop-outs in longitudinal clinical trials. Statistics in Medicine 11, 2043–2061.

Hogan, J.W., Roy, J., Korkontzelou, C., 2004. Tutorial in biostatistics: handling drop-out in longitudinal studies. Statistics in Medicine 23, 1455–1497.

Horton, N.J., Lipsitz, S.R., 2001. Statistical computing software reviews: multiple imputation in practice: comparison of software packages for regression models with missing variables. American Statistician 55 (3), 244–254.

Houck, P.R., Mazumdar, S., Mulsant, B.H., Pollock, B.G., Dew, M.A., Reynolds, C.F., 2003. An intent-to-treat method for enhancing analysis of clinical trials with rescue medication: a mixed-model approach. Psychopharmacology Bulletin 37 (1), 79–89.

Lachin, J.M., 2000. Statistical considerations in the intent-to-treat principle. Controlled Clinical Trials 21, 167–189.

Laird, N.M., Ware, J.R., 1982. Random-effects models for longitudinal data. Biometrics 38, 963–974.

Lavori, P.W., 1992. Clinical trials in psychiatry: should protocol deviation censor patient data? Neuropsychopharmacology 6 (1), 39–48.

Lavori, P.W., Dawson, R., Shera, D., 1995. A multiple imputation strategy for clinical trials with truncation of patient data. Statistics in Medicine 14, 1913–1925.

Little, R.J.A., 1988. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association 83, 1198–1202.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. Wiley-Interscience, Hoboken, NJ.

Liu, G., Gould, A.L., 2002. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. Journal of Biopharmaceutical Statistics 12 (2), 207–226.

Mallinckrodt, C.H., Sanger, T.M., Dube, S., DeBrota, D.J., Molenberghs, G., Carroll, R.J., Potter, W.Z., Tollefson, G.D., 2003. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. Biological Psychiatry 53, 754–760.

Mazumdar, S., Liu, K.S., Houck, P.R., Reynolds, C.F., 1999. Intent-to-treat analysis for longitudinal clinical trials: coping with the challenge of missing values. Journal of Psychiatric Research 33, 87–95.

Mazumdar, S., Houck, P.R., Liu, K.S., Mulsant, B.H., Pollock, B.G., Dew, M.A., Reynolds, C.F., 2002. Intent-to-treat analysis for clinical trials: use of data collected after termination of treatment protocol. Journal of Psychiatric Research 36, 153–164.

Mulsant, B.H., Pollock, B.G., Nebes, R., Miller, M., Sweet, R., Stack, J., Houck, P.R., Bensasi, S., Mazumdar, S., Reynolds, C.F., 2001. A twelve-week double-blind randomized comparison on nortriptyline and paroxetine in older depressed inpatients and outpatients. American Journal of Geriatric Psychiatry 9, 406–414.

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

Rubin, D.B., 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association 91 (434), 473–489.

SAS Institute, 2000. SAS/STAT Software, Version 8. Author, Cary, NC.

Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychological Methods 7 (2), 147–177.

Sheiner, L.B., Rubin, D.B., 1995. Intention-to-treat analysis and the goal of clinical trials. Clinical Pharmacology and Therapy 57, 6–15.

Sinharay, S., Stern, H.S., Russell, D., 2001. The use of multiple imputation for the analysis of missing data. Psychological Methods 6 (4), 317–329.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

Yuan, Y.C., 2000. Multiple imputation for missing data: concepts and new development. Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (Paper No. 267). SAS Institute, Cary, NC.