Title

Approach to addressing missing data for electronic medical records and pharmacy claims data research.

Journal

Pharmacotherapy, 35(4)

ISSN

0277-0008

Authors

Bounthavong, Mark
Watanabe, Jonathan H
Sullivan, Kevin M

Publication Date

2015-04-01

DOI

10.1002/phar.1569

Peer reviewed

ORIGINAL RESEARCH ARTICLE

Approach to Addressing Missing Data for Electronic Medical Records and Pharmacy Claims Data Research

Mark Bounthavong,[1,2,*] Jonathan H. Watanabe,[3] and Kevin M. Sullivan,[4]

[1]Pharmaceutical Outcomes Research and Policy Program, University of Washington, Seattle, Washington; [2]Veterans Affairs San Diego Healthcare System, San Diego, California; [3]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California; [4]Rollins School of Public Health, Emory University, Atlanta, Georgia

Objective The complete capture of all values for each variable of interest in pharmacy research studies remains aspirational. The absence of these possibly influential values is a common problem for pharmacist investigators. Failure to account for missing data may translate to biased study findings and conclusions. Our goal in this analysis was to apply validated statistical methods for missing data to a previously analyzed data set and compare results when missing data methods were implemented versus standard analytics that ignore missing data effects.

Design Using data from a retrospective cohort study, the statistical method of multiple imputation was used to provide regression-based estimates of the missing values to improve available data usable for study outcomes measurement. These findings were then contrasted with a complete-case analysis that restricted estimation to subjects in the cohort that had no missing values. Odds ratios were compared to assess differences in findings of the analyses. A nonadjusted regression analysis ("crude analysis") was also performed as a reference for potential bias.

Setting Veterans Integrated Systems Network that includes VA facilities in the Southern California and Nevada regions.

Patients New statin users between November 30, 2006, and December 2, 2007, with a diagnosis of dyslipidemia.

Main outcome measure We compared the odds ratios (ORs) and 95% confidence intervals (CIs) for the crude, complete-case, and multiple imputation analyses for the end points of a 25% or greater reduction in atherogenic lipids.

Results Data were missing for 21.5% of identified patients (1665 subjects of 7739). Regression model results were similar for the crude, complete-case, and multiple imputation analyses with overlap of 95% confidence limits at each end point. The crude, complete-case, and multiple imputation ORs (95% CIs) for a 25% or greater reduction in low-density lipoprotein cholesterol were 3.5 (95% CI 3.1–3.9), 4.3 (95% CI 3.8–4.9), and 4.1 (95% CI 3.7–4.6), respectively. The crude, complete-case, and multiple imputation ORs (95% CIs) for a 25% or greater reduction in non–high-density lipoprotein cholesterol were 3.5 (95% CI 3.1–3.9), 4.5 (95% CI 4.0–5.2), and 4.4 (95% CI 3.9–4.9), respectively. The crude, complete-case, and multiple imputation ORs (95% CIs) for 25% or greater reduction in TGs were 3.1 (95% CI 2.8–3.6), 4.0 (95% CI 3.5–4.6), and 4.1 (95% CI 3.6–4.6), respectively.

Conclusion The use of the multiple imputation method to account for missing data did not alter conclusions based on a complete-case analysis. Given the frequency of missing data in research using electronic health records and pharmacy claims data, multiple imputation may play an important role in the validation of study findings.

Key Words pharmacists, research, multiple imputation, missing data, adherence, dyslipidemia, statins, logistic regression, complete-case analysis, pharmacoepidemiology.

(Pharmacotherapy 2015;35(4):380–387) doi: 10.1002/phar.1569

Widespread use of electronic health records, medication administration records, and pharmacy claims data has allowed health services researchers, pharmacoepidemiologists, and other outcomes researchers to perform rapid analyses that can be used for clinical decision support, policy development, formulary management, and pharmacovigilance. However, missing data in pharmacy claims and other electronic records are common problems with important potential consequences.[1, 2] For example, in the QRISK study, where three-fourths of patients had missing data, there was no association between cholesterol and cardiovascular events based on multiple imputation. However, a complete-case analysis of the same data set yielded a significant association.[3] Researchers who perform statistical inference with missing data risk reporting conclusions that may be invalid.[4–9] Reliance on methods such as complete-case analysis, where the missing data are ignored and observations are dropped from analysis, may be appropriate if the missing data mechanism is missing completely at random (MCAR).[6, 10–12] An example of MCAR would be the absence of a laboratory value because a test tube randomly fell out of a rack while being transported. In this situation, there is no causal link between the absence of the data and the actual laboratory value. However, in situations where missing data cannot be confirmed to be MCAR, the absent data is deemed missing at random (MAR). In MAR, researchers should consider whether more sophisticated methods are needed to estimate the missing data.[6, 10, 11]

According to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE), proper reporting of missing data should include an explanation of methods used to handle missing data.[13] In practice, however, missing data are commonly unreported in observational research.[14, 15] A review of a subset of randomized trials between July and December 2001 published in the *British Medical Journal*, the *Journal of the American Medical Association*, *The Lancet*, and the *New England Journal of Medicine* reported that 89% of the 71 studies reviewed had partial or missing data.[15] The authors concluded that 92% of studies unjustifiably used complete-case analysis to handle missing data.

Multiple imputation is one form of missing data analysis appropriate to a missing data mechanism that is MAR.[6, 12, 16–20] The multiple imputation method inputs plausible values for missing data based on the observed values. Multiple imputation is a process of imputation that creates multiple data sets using regression methods and data simulations.[6, 16, 19] Each simulated data set is analyzed using standard methods (e.g., logistic regression), which are then combined to produce outcomes estimates and confidence intervals (CIs).

This article provides health services researchers, pharmacoepidemiologists, outcomes researchers, and decision makers with a demonstration of multiple imputation to support the published results of an observational study that relied on complete-case analysis.[21] After a brief description of the original study, we compare multiple imputation and complete-case analysis results. We conclude with a discussion of our research implications in the context of pharmacy research. Appendix 1 provides details of the multiple imputation method.

Objective

A previous study evaluated the association between medication adherence levels and improvements in lipid profile.[21] That analysis used a complete-case analysis that removed subjects who did not have a complete list of values for all variables in the multiple regression model. We sought to investigate whether multiple imputation would yield different conclusions from the complete-case analysis.

Methods

Case Study

We used data from a previous publication[21] as a case study to compare findings of a complete-case analysis with the findings of an analysis that applied multiple imputation for missing data.

This project was supported by grant T32HS013853 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality or the Department of Veterans Affairs.

A portion of this study was presented at the AcademyHealth 2014 Annual Research Meeting in San Diego, California, on June 8, 2014.

*Address for correspondence: Mark Bounthavong, Pharmaceutical Outcomes Research and Policy Program, University of Washington, 1959 NE Pacific Street, HSB H-375, Box 357630, Seattle, WA 98195-7630; e-mail: mbounth@uw.edu.

The retrospective cohort study evaluated the association between adherence to statin medication for dyslipidemia and a 25% or greater change in low-density lipoprotein cholesterol (LDL-C), non–high-density lipoprotein cholesterol (non–HDL-C), and triglycerides (TGs). Further details of the observational study are described elsewhere.[21]

Study Population

The study population[21] was drawn from the Veterans Integrated Systems Network 22 (Desert Pacific Healthcare Network) that includes VA facilities in the Southern California (Los Angeles, Long Beach, Loma Linda, and San Diego) and Nevada (Las Vegas) regions that service ~ 1.4 million veterans.[22] Patients were included if they were a new statin user between the periods of November 30, 2006, and December 2, 2007, with a diagnosis of dyslipidemia (or related disorders) based on the International Classification of Disease, Ninth Revision (Appendix 1), older than 18 years, and had been continuously enrolled in the VA health plan for at least 2 years.[21] Patients were considered new statin users as defined by a 6-month washout period before filling their first statin prescription. Patients were followed for a 1-year observation period after the index date on LDL-C, non–HDL-C, and TGs. Subjects were required to be eligible for VA medical and pharmacy services 6 months prior to the index date and throughout the study period. Patients were excluded if they switched statins during the 12-month follow-up period or had an admission for more than 30 consecutive days. The Veterans Affairs San Diego Healthcare System and University of California, San Diego, institutional review board granted exemption status for the study protocol.[21]

Summary Measure

Adherence was the main exposure variable of interest and categorized into adherent or nonadherent based on a medication possession ratio (MPR) threshold level of 0.80. Patients who were at or above the threshold were considered adherent; patients who were below the threshold were considered nonadherent.[21] MPR was calculated as the days supplied of prescription medication divided by the number of days the subject was designated to be on therapy during the study period.[23, 24]

The dependent variable was reduction in lipid panel levels for LDL-C, non–HDL-C, and TGs at 12 months. In terms of study end points, subjects were dichotomized into those that achieved a 25% or greater reduction from baseline or those that had less than a 25% reduction from baseline for each lipid type.[21] A 25% reduction in atherogenic lipids was described in prior studies as a clinically significant improvement.[25, 26]

Multiple Imputation Analysis

Using the data from the study,[21] we applied multiple imputation to address the missing data. This framework requires several data assumptions. First, the pattern of missing values is considered missing at random.[6, 19, 27, 28] Second, the variables in the multivariate model should have a normal distribution.[6] Finally, all subjects must have some observed values for imputation to proceed.[7] To address sampling variability, five data sets were created using multiple imputation, and the effect estimates were averaged.[6] Appendix 1 provides further details about multiple imputation.

Data Analysis

In this case study,[21] a variable was created for all subjects that indicated whether each individual had missing data or otherwise. Balance of demographic characteristics for patients with and without missing data was then evaluated to confirm assumptions of MAR data.

A logistic regression model was used to evaluation the association between adherence and lipid reduction. OR estimates and CIs from the regression models using multiple imputation were compared with complete-case analysis and crude estimates. Crude estimates were included to reflect the unadjusted model. Statistical significance was defined as $p<0.05$. Data analysis was performed using SAS v.9.3 (Cary, NC). Details about the logistic regression model are provided in Appendix 1.

Results

A total of 7739 patients were identified based on the inclusion and exclusion criteria. Overall, 78% of the patients (6074) had complete values and were included in the complete-case analysis. For the complete-case analysis, 2827 (47%) of the subjects were adherent (MPR of 0.80 or higher) and 3247 (53%) were nonadherent

(MPR lower than 0.80) (Table 1). A large proportion of patients (5786 [95%]) were male, prescribed simvastatin (5154 [85%]), and had hypertension (4366 [72%]). Patients in the adherent group were older (64 vs 62 yrs, p<0.001), had higher starting medication count (7.9 vs 6.8, p<0.001), and lower LDL-C (133.9 vs 141.0 mg/dl, p<0.001), non-HDL-C (167.3 vs 174.9 mg/dl, p<0.001), and TGs (209.3 vs 217.6 mg/dl, p<0.001). Rates of diabetes (p=0.005), hypertension (p<0.001), and vascular disease (p<0.001) were lower in adherent patients, but more adherent patients had congestive heart failure (p=0.006) relative to nonadherent patients.

For the baseline lipid panel variables of LDL-C, HDL-C, and TG values, 16.1%, 16.6%, and 17.9% of variable values were missing, respectively (Figure 1). In terms of the follow-up lipid panel variables, 19.2%, 19.6%, and 19.4% of values were missing for LDL-C, HDL-C, and TG values, respectively. Approximately 7% of baseline body mass index (BMI) values were missing.

Comparison of Subjects with Missing and Complete Data

Similar gender proportions were observed for groups with missing data for baseline values for LDL-C, HDL-C, and TGs compared with groups with complete data (95.67%, 95.65%, and 95.88% vs 95.85%, 95.32%, and 95.28%, respectively). A higher proportion of patients with complete data were categorized as adherent compared with those with missing data for baseline LDL-C, HDL-C, and TGs (38.70%, 38.96%, and 39.14% vs 46.26%, 46.26%, and 46.17%, respectively) (Table 2).

Regression Findings

The logistic regression model for the multivariate analysis controlled for age, BMI, gender, baseline lipid values, comorbid conditions (diabetes, hypertension, congestive heart failure, history of myocardial infarction, angina, and vascular disease), statin use, ethnicity, and starting medication count. The regression results for the crude, complete-case analysis, and multiple imputation method were similar with CI overlap for all lipid panel end points (Table 3). The ORs (95% CIs) in achieving a 25% or greater reduction in LDL-C for crude, complete-case, and multiple imputation analyses were 3.5 (95% CI 3.1–3.9), 4.3 (95% CI 3.8–4.9), and 4.1 (95% CI 3.7–4.6), respectively.

Table 1. Baseline Demographics Between Adherent and Nonadherent Subjects

Variables[a]	Adherent N=2827	Nonadherent N=3247	p value[b]
Age, yrs	64.07 (10.79)	62.28 (11.29)	<0.001
Male, n (%)	2701 (95.54)	3085 (95.01)	0.330
Starting medication count	7.94 (4.62)	6.79 (4.14)	<0.001
LDL-C	133.89 (40.32)	141.04 (39.22)	<0.001
Baseline non–HDL-C	167.30 (46.43)	174.85 (45.32)	<0.001
Baseline TGs	169.72 (150.19)	172.47 (144.95)	0.826
Baseline TC	209.25 (48.39)	217.57 (47.03)	<0.001
Ethnicity, n (%)			
White	1474 (52.14)	1474 (45.40)	<0.001
Black	345 (12.20)	548 (16.88)	
Hispanic	266 (9.41)	393 (12.10)	
Asian	91 (3.22)	121 (3.73)	
American Indian	37 (1.31)	46 (1.42)	
Unknown	614 (21.72)	665 (20.48)	
Statin use, n (%)			
Simvastatin	2378 (84.12)	2776 (85.49)	0.057
Atorvastatin	9 (0.32)	23 (0.71)	
Rosuvastatin	187 (6.61)	170 (5.24)	
Lovastatin	187 (6.61)	212 (6.53)	
Pravastatin	34 (1.20)	30 (0.92)	
Fluvastatin	32 (1.13)	36 (1.11)	
Copayment, n (%)	1847 (65.33)	2196 (67.63)	0.058
Diabetes, n (%)	1113 (39.37)	1165 (35.88)	0.005
Hypertension, n (%)	2142 (75.77)	2224 (68.49)	<0.001
Vascular disease, n (%)	969 (34.38)	984 (30.30)	<0.001
Congestive heart failure, n (%)	157 (5.55)	131 (4.03)	0.006
History of myocardial infarction, n (%)	84 (2.97)	99 (3.05)	0.860
Angina, n (%)	70 (2.48)	66 (2.03)	0.244

HDL-C = high-density lipoprotein cholesterol; LDL-C = low-density lipoprotein cholesterol; TC = total cholesterol; TGs = triglycerides.

[a]Unless otherwise stated, data are presented as mean (SD).

[b]Student *t* test and χ^2 test for continuous and discrete data, respectively.

The crude ORs (95% CIs) in achieving a 25% or greater reduction in non–HDL-C for crude, complete-case, and multiple imputation analyses were 3.5 (95% CI 3.1–3.9), 4.5 (95% CI 4.0–5.2), and 4.4 (95% CI 3.9–4.9), respectively. The ORs (95% CIs) in achieving a 25% or greater reduction in TGs for crude, compete-case, and multiple imputation analyses were 3.1 (95% CI 2.8–3.6), 4.0 (95% CI 3.5–4.6), and 4.1 (95% CI 3.6–4.6), respectively.

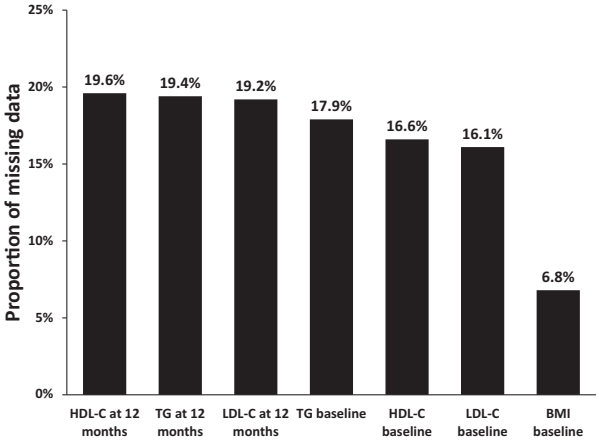

Figure 1. Proportion of missing data for several variables in the observational study. The proportions of missing data for the other baseline parameters were 0%. BMI = body mass index; HDL-C = high-density lipoprotein cholesterol; LDL-C = low-density lipoprotein cholesterol; TGs = triglycerides.

Discussion

In this case study, the use of multiple imputation did not alter the conclusions of the complete-case and crude analyses. It functions as a supporting sensitivity analysis that the complete-case conclusions were appropriate despite the absence of data for 22% of patients. We cannot conclude necessarily that multiple imputation has eliminated missing data bias because this would require having the missing data to determine. It does serve as a robust data-driven approach to reduce uncertainty in estimates by using all available data for estimation rather than discarding observations because of absent values.

Researchers with an interest in using large databases should consider performing missing data analysis to confirm the results of complete-case analysis. Ideally, this would be performed in conjunction with the primary complete-case analysis in accord with the STROBE recommendations.[9] If there is suspicion that missing data could influence results, systematic examination of the possible effects should be conducted. Given the growing importance of research based on electronic medical records and its application in clinical care, mitigating the risk of spurious conclusions due to missing data demands greater attention. Multiple imputation applications are now included in commonly used statistical software packages including those available at no charge. Hence production of robust results is accessible to researchers in virtually any setting.

Table 2. Comparison of Groups with Missing and Nonmissing Data

	Missing			Nonmissing		
Variable of interest	LDL-C at baseline	HDL-C at baseline	TGs at baseline	LDL-C at baseline	HDL-C at baseline	TGs at baseline
No. of patients	1248	1286	1239	6491	6453	6500
Age, yrs, mean (SD)	65.49 (11.82)	65.59 (11.85)	65.82 (11.81)	63.13 (11.15)	63.09 (11.14)	63.07 (11.14)
Males, n (%)	1194 (95.67)	1230 (95.65)	1188 (95.88)	6187 (95.32)	6151 (95.32)	6193 (95.28)
Adherent, n (%)	483 (38.70)	501 (38.96)	485 (39.14)	3003 (46.26)	2985 (46.26)	3001 (46.17)

HDL-C = high-density lipoprotein cholesterol; LDL-C = low-density lipoprotein cholesterol; TGs = triglycerides.

Table 3. Odds of Achieving 25% Reduction or Greater in Lipid Panel Levels for Adherent vs Nonadherent Patients on a Statin in the Veterans Integrated Systems Network 22

Outcome	Crude analysis N=6074 OR (95% CI)	Complete-case analysis[a] N=6074 OR (95% CI)	Multiple imputation[a] N=7739 OR (95% CI)
≥ 25% reduction in LDL-C	3.5 (3.1–3.9)	4.3 (3.8–4.9)	4.1 (3.7–4.6)
≥ 25% reduction in non–HDL-C	3.5 (3.1–3.9)	4.5 (4.0–5.2)	4.4 (3.9–4.9)
≥ 25% reduction in Triglycerides	3.1 (2.8–3.6)	4.0 (3.5–4.6)	4.1 (3.6–4.6)

CI = confidence interval; HDL-C = high-density lipoprotein cholesterol; LDL-C = low-density lipoprotein cholesterol; OR = odds ratio.
[a]Adjusted for age, gender, body mass index, baseline lipid values, comorbid conditions (diabetes, hypertension, congestive heart failure, history of myocardial infarction, angina, vascular disease), statin use, ethnicity, and starting medication count.

Inclusion of multiple imputation results can serve to strengthen the veracity of the conclusions.

Most statistical software (e.g., SAS, SPSS, and STATA) utilizes complete-case analysis for multiple regression models by default. Researchers unaware of this default setting may find a reduction in their study sample size and altered CIs when performing analyses. This article serves to illuminate the common and routinely ignored phenomenon of missing data. We accomplished this by examining a real-world data set to quantify the amount of missing data and attempted to address it.

A study limitation is the possibility of having a missing data pattern where the absence of data is reflective of the outcome. This scenario is termed not missing at random (NMAR). An example of NMAR data would be a survey conducted to determine the relationship between worker satisfaction and the number of hours worked per week in which information was missing for all employees working overtime who were too busy to respond to the survey. No statistical method is currently available to determine if the missing data pattern was either MAR or NMAR. Application of multiple imputation with NMAR data may amplify bias rather than eliminate bias.[29] To ensure that NMAR was not a sizable risk, we compared several baseline characteristics (age, gender, and adherence status) between the patient sample with missing data and the patient sample with complete data.

The importance of using electronic medical records and pharmacy claims data for assessing outcomes, performance measurement, and health care forecasting has grown exponentially as the analysis of large data sets has become easier for researchers.[30] Clinical decision making and health care policy rely increasingly on statistical analyses of medication records and pharmacy-derived claims. However, the absence of complete data is common in electronic medical records and pharmacy claims data that are designed for clinical management, not necessarily investigational studies. Prior research found that missing data can influence the results and conclusions generated from studies based on pharmacy claims records.[1] Few studies appropriately address missing data or describe the methods applied to contend with the absence of study information.[15] The convenience of assuming data is MCAR is offset by the knowledge that missing data bias may potentially reverse the study findings.

In this report, we describe a method to account systematically for missing data to validate findings of a complete-case analysis that measured the association between adherence and atherogenic lipid reduction in statin users. This was motivated by the potential loss of large portions of study data when complete-case analysis is performed for pharmacy investigations. Health services researchers, pharmacoepidemiologists, and other outcomes researchers using electronic medical records could benefit from additional analyses that account for the missing data to bolster the robustness of study findings.

Conclusions

The use of multiple imputation for addressing missing data did not alter conclusions that relied on a complete-case analysis, despite missing data for 22% of study subjects. Application of missing data methods served as statistical support that findings from the complete-case analysis were robust. Multiple imputation represents a valid and accessible means of accounting for missing data.

Acknowledgments

We acknowledge the assistance of the late Michael Juzba for validating and extracting the data. We are

also grateful for additional validation provided by Josephine N. Tran. We appreciate the helpful comments from the reviewers that strengthened this manuscript.

Conflict of Interest

The authors have nothing to disclose.

References

1. **Polinski JM, Schneeweiss S, Levin R, Shrank WH.** Completeness of retail pharmacy claims data: implications for pharmacoepidemiologic studies and pharmacy practice in elderly patients. Clin Ther 2009;31:2048–59.
2. **Fitzmaurice G.** Missing data: implications for analysis. Nutrition 2008;24:200–2.
3. **Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P.** Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. Heart 2008;94:34–9.
4. **Acock AC.** Working with missing values. J Marriage Fam 2005;67:1012–28.
5. **Gorelick MH.** Bias arising from missing data in predictive models. J Clin Epidemiol 2006;59:1115–23.
6. **Little RJA, Rubin DB.** Statistical Analysis with Missing Data, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2002.
7. **Raghunathan TE.** What do we do with missing data? Some options for analysis of incomplete data. Annu Rev Public Health 2004;25:99–117.
8. **Rotnitzky A, Wypij D.** A note on the bias of estimators with missing data. Biometrics 1994;50:1163–70.
9. **Gallo V, Egger M, McCormack V, et al.** STrengthening the Reporting of OBservational studies in Epidemiology-Molecular Epidemiology STROBE-ME: an extension of the STROBE statement. J Clin Epidemiol 2011;64:1350–63.
10. **Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM.** Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087–91.
11. **Rubin DB.** Inference and missing data. Biometrika 1976;63:581–92.
12. **Knol MJ, Janssen KJM, Donders ART, et al.** Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol 2010;63:728–36.
13. **Von Elm E, Altman DG, Egger M, et al.** The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Ann Intern Med 2007;147:573–7.
14. **Fielding S, Maclennan G, Cook JA, Ramsay CR.** A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. Trials 2008;9:51.
15. **Wood AM, White IR, Thompson SG.** Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials 2004;1:368–76.
16. **Rubin DB.** Multiple imputation after 18+ years. J Am Stat Assoc 1996;91:473–89.
17. **De Goeij MCM, van Diepen M, Jager KJ, Tripepi G, Zoccali C, Dekker FW.** Multiple imputation: dealing with missing data. Nephrol Dial Transplant 2013;28:2415–20.
18. **Horton NJ, Kleinman KP.** Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007;61:79–90.
19. **Schafer JL.** Multiple imputation: a primer. Stat Methods Med Res 1999;8:3–15.
20. **Janssen KJM, Donders ART, Harrell FE Jr, et al.** Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 2010;63:721–7.
21. **Watanabe JH, Bounthavong M, Chen T.** Revisiting the medication possession ratio threshold for adherence in lipid management. Curr Med Res Opin 2013;29:175–80.
22. **U.S. Department of Veterans Affairs.** VISN 22: Desert Pacific Healthcare Network. Available from http://www.va.gov/directory/guide/region.asp?ID=1022. Accessed July 23, 2013.
23. **Andrade SE, Kahler KH, Frech F, Chan KA.** Methods for evaluation of medication adherence and persistence using automated databases. Pharmacoepidemiol Drug Saf 2006;15:565–74; discussion 575–7.
24. **Hess LM, Raebel MA, Conner DA, Malone DC.** Measurement of adherence in pharmacy administrative databases: a proposal for standard definitions and preferred measures. Ann Pharmacother 2006;40:1280–8.
25. **Grundy SM, Cleeman JI, Merz CNB, et al.** Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III Guidelines. J Am Coll Cardiol 2004;44:720–32.
26. **Gotto AM Jr, Grundy SM.** Lowering LDL cholesterol: questions from recent meta-analyses and subset analyses of clinical trial Data Issues from the Interdisciplinary Council on Reducing the Risk for Coronary Heart Disease, ninth Council meeting. Circulation 1999;99:E1–7.
27. **Rubin DB.** Multiple Imputation for Nonresponse in Surveys. New York, NY: John Wiley & Sons, 1987.
28. **Schafer JL, Graham JW.** Missing data: our view of the state of the art. Psychol Methods 2002;7:147–77.
29. **Allison PD.** Multiple imputation for missing data a cautionary tale. Sociol Methods Res 2000;28:301–9.
30. **Raghupathi W, Raghupathi V.** Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3.1

APPENDIX

ICD-9-CM Diagnosis Codes for Dyslipidemia

ICD-9-CM diagnosis code	Description
272	Disorders of lipid metabolism
272.1	Pure hyperglyceridemia
272	Mixed hyperlipidemia
272.3	Hyperchylomicronemia
272.4	Hyperchylomicronemia

Logistic regression model

Logistic model expressed as:

$$\mathrm{p}(D = 1|X_1, X_2, \ldots, X_k) = \mathrm{p}(X) = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^{k} \beta_i X_i + \varepsilon)}},$$

where $i = 1, 2, 3, \ldots, k$, $D = 1$ denotes the outcome of interest (e.g., achieving 25% in lipid reduction), X_i denote the k number of independent variables in the regression model, α and β denote model parameters, $\mathrm{p}(X)$ denotes the probability of achieving the clinical goals ($X = 1$) given that the following independent

variables (X_k) are present, and ε denotes the error term. The logit form of the logistic model is expressed as:

$$\text{Logit } p(X) = \alpha + \sum_{i=1}^{k} \beta_i X_i + \varepsilon,$$

where $i = 1, 2, 3, \ldots, k$, X_k denotes the k number independent variables for in the regression model, and ε denotes the error term. The odds ratio (OR) is computed as the product of exponentials:

$$\text{Odds ratio (OR)} = \prod_{i=1}^{k} e^{\beta_i (X_{1i} - X_{0i})},$$

where X_1 and X_0 are two specifications of the collection of k independent variables X_1, X_2, $X_3, \ldots, X_k$.

Multiple imputation

Multiple imputation uses Bayesian methods to generate posterior probability for the parameter estimate using a specified prior distribution with the likelihood function.[6, 11, 16, 27] The target variable, which contains the unobserved value (Y_{mis}), dependents on the available observed value (Y_{obs}) on Y. Predictive distribution for Y_{mis} is generated using Markov chain Monte Carlo (MCMC) simulations [$p(Y_{mis} \mid Y_{obs})$], an iterative process that ends when the posterior distribution of Y_{mis} stabilizes and converges.[6, 19] This iterative process generates predictive Y_{mis} for each subject resulting in different estimates for the missing values. Each data set has n number of Y_{mis} that are imputed using MCMC processes that results in a single data set. As m number of data set is created, they can be combined for use in statistical analysis. The Bayesian process is determined as follows[6, 11, 16]:

$$p(Q|Y_{obs}) = \int p(Q|Y_{obs}, Y_{mis}) p(Y_{mis}|Y_{obs}) dY_{mis},$$

where $p(Q \mid Y_{obs})$ is the actual posterior distribution of Q and $p(Y_{mis} \mid Y_{obs})$ is the posterior predictive distribution of Y_{mis} given Y_{obs}.[6, 16] The likelihood function is denoted by $p(Q \mid Y_{obs}, Y_{mis})$.

Data sets are combined by averaging the parameter estimates (Q_i) over m number of data sets:

$$\overline{Q}_m = \frac{1}{m} \sum_{i=1}^{m} Q_i,$$

where Q_i is the point estimate generated from each of the i-th imputed data set.[6, 11, 16, 27]

The within-imputation variance or variability (U_m) is determined by the following:

$$\overline{U}_m = \frac{1}{m} \sum_{i=1}^{m} U_i,$$

where $\overline{U}_m$ is the average within-imputation variance for m imputations and U_i is the variance for each i-th imputed data set.[6, 16]

The between-imputation variance or variability (B_m) is determined by the following:

$$B_m = \frac{1}{m-1} \sum_{i=1}^{m} (Q_i - \overline{Q}_m)^2,$$

where $(Q_i - \overline{Q}_m)^2$ represents the difference between the predicted point estimate from each of the i-th imputed data set (Q_i) and the average predicted point estimate over m number of data sets ($\overline{Q}_m$).

The total variance or variability (T_m) is determined by combining the within-imputation variance (U_m) and the between-imputation variance (B_m):

$$T_m = \overline{U}_m + \left(1 + \frac{1}{m}\right) B_m,$$

where $\overline{U}_m$ is the average within-imputation variance for m imputations. Overall standard error (SE) is the square root of T_m.[6, 11, 16, 27]

Confidence intervals at the 95% level (95% CIs) are determined using the following:

$$\overline{Q}_m \pm t_v \left(\frac{\alpha}{2}\right) T_m^{1/2},$$

where $t_v(\alpha/2)$ represents that upper and lower confidence bounds as determined by $100(\alpha/2)$. For a 95% CI with lower and upper bounds of 2.5% and 97.5%, respectively, $t_v(\alpha/2) = 1.96$.[6, 16, 19]

Several assumptions were required for multiple imputation process to be valid. First, the pattern of missing values must be "ignorable," which is achieved when the missing data mechanism is MAR.[6, 19, 27, 28] The variables in the multivariate model must have a normal distribution; however, multiple imputation is robust to violation of this assumption.[6] The data set must also follow an item nonresponse pattern where all subjects contain some observed values (Y_{obs}) of Y.[16] If the data set has unit nonresponse patterns where the subjects or groups of subjects have none of the observed values (Y_{obs}) of Y, then the multiple imputation procedure will not be suitable. In our analysis, five imputed data sets ($m = 5$) were used to combine the results into the regression methods because this was considered to be efficient.