Missing Data: A Special Challenge in Aging Research

Susan E. Hardy, MD, PhD[*], **Heather Allore, PhD**[†], and **Stephanie A. Studenski, MD, MPH**[*,‡]

[*] University of Pittsburgh School of Medicine

[†] Yale University School of Medicine, Department of Internal Medicine, New Haven, CT

[‡] GRECC, Pittsburgh Veterans Affairs Health Care System

Abstract

Evidence about care of older adults informs practice but is influenced by special methodological challenges. Missing data, ranging from lack of individual items in questionnaires to complete loss to follow up, affect the quality of the evidence and are more likely to occur in studies of older adults because older adults have more health and functional problems that interfere with all aspects of data collection. The purpose of this article is to promote knowledge about the risks and consequences of missing data in clinical aging research, and to provide an organized approach to prevention and management. While it is almost never possible to achieve complete data capture, efforts to prevent missing data are more effective than analytic "cure". Strategies to prevent missing data include 1) selecting a primary outcome that is easy to determine and devise valid alternate definitions, 2) adapting data collection to the special needs of the target population, 3) pilot testing data collection plans, and 4) monitoring missing data rates during the study and adapt data collection procedures as needed. Key steps in the analysis of missing data include 1) assessing the extent and types of missing data prior to analysis, 2) exploring potential mechanisms that contributed to the missing data, and 3) using multiple analytic approaches to assess the effect of missing data on the results. Manuscripts should 1) disclose rates of missing data and losses to follow up, 2) compare drop outs to participants who completed the study, 3) describe how missing data was managed in the analysis phase, and 4) discuss the potential impact of missing data on the conclusions of the study.

Keywords

clinical research; study design; statistics; missing data; aging

INTRODUCTION

Missing data are a special challenge in clinical aging research because older adults are more likely than younger adults to experience health and functional problems that limit data collection. In longitudinal studies, death and loss to follow-up increase with age.[1] Cognitive or physical deficits can lead to inability to perform some assessments, leading to incomplete

data.2 Missing data from any of these causes can bias results, reduce generalizability and limit power. The ultimate consequence of missing data is distortion from the truth; reducing the internal and external validity of study results (Table 1). For example, in a hypothetical study of the course of dementia, persons who become unable to follow directions may not complete formal cognitive testing and will have missing test scores. Over time, as those who are unable to complete the tests do not contribute data, the group mean and range of cognitive test scores will appear better than they really are. In a clinical trial of an intervention to prevent disability, missing data might occur if persons with disability have difficulty coming to a central site for testing. If the intervention was effective, the control group might develop more disability than the treatment group, be less able to come in for testing and subsequently have more missing data. Using only the data obtained from persons who came in for testing, the difference between treatment arms in disability scores will be smaller than truly occurred.

Investigators within the field of aging research have developed successful strategies to minimize missing data during studies of complex older people. These strategies may be useful to all investigators who wish to extend participation to a greater range of age and health. While missing data in older adults are the focus of this manuscript, similar issues and solutions may apply to other populations with complex, multisystem chronic illnesses and unique social issues, such as persons with AIDS, renal failure on dialysis or multiple developmental disabilities.3

Our objective is to promote knowledge about the risks and consequences of missing data in clinical aging research, and to provide an organized approach to its prevention and management. Everyone who creates or uses data, including investigators, trainees, grant sponsors, providers, policy makers, older adults and their families, has a stake in the creation of reliable evidence to improve care for the rapidly growing aged population. Creating strong evidence requires special attention to the prevention and management of missing data.

OVERVIEW

Missing data can range from loss of single items, for example when a participant refuses or is unable to answer a question, to loss of all follow-up data, as when a participant withdraws from a study. For any kind of missing data, prevention is more effective than analytic "cure" and should be part of every phase of research. Planning for missing data begins with the development of the research question and design of the study, and then continues throughout planning, piloting, implementation, monitoring, and data management and analysis (Table 2). In each phase, the challenges of an aging population are anticipated and strategies to reduce the risk of missing data are implemented. In general, the most effective strategies include 1) use easily obtainable primary outcomes, 2) prioritize data collection, 3) prespecify alternative data collection strategies and 4) anticipate the resources needed to maintain participants with health and functional problems in the study. Analytic techniques for management are a last resort and can rarely fully account for the effects of missing data.

THE RESEARCH QUESTION AND THE DESIGN

In an ideal (but unachievable) study, the participants reflect the true referent population and are all retained with complete data. In reality, any study is a trade-off between internal validity and generalizability. Scientific issues, such as need for a homogenous population or risks of study interventions, usually guide the choice of inclusion and exclusion criteria, but these decisions also have an impact on missing data rates. Participants excluded from research due to comorbidity or frailty are also more likely to generate missing data. A

compromise that increases generalizability is to minimize exclusions while simultaneously adapting study procedures to maximize data completion. Recruitment and retention strategies for older adults are discussed in detail in another Research Methods article.4

In order to minimize missing data, studies of older adults are likely to require an investment of substantial resources for this purpose. The investigator who is designing the study must weigh the best use of limited resources; there may be competition between the need to maximize sample size versus the need to prevent missing data. For example, a clinical trial with fixed resources might invest in a sample size of 200 and achieve 95% outcome data collection or for the same cost, obtain a sample size of 400 but only achieve 70% data capture. While the final sample size will be larger in the latter, the results may be more distorted and less valid.

MEASURES

The impact of missing data varies depending upon the type of variable: primary outcome, secondary outcome, primary predictor or covariate. Some strategies to reduce missing data are specific to the type of measure, while others apply to all. In general, consider the impact of health and functional limitations on data collection and minimize the time and effort of the participant associated with data collection.

Missing data due to inability to perform a test is a special concern in aging research. Since this is a predictable problem, it can be anticipated. Reasons for missing data, such as physical inability, cognitive state, or equipment failure, can be predefined, coded and used later in analysis. Some performance measures incorporate a code for "can't do". For example, the Short Physical Performance Battery assigns a score of 0 to inability to perform a task.5 Tests that count the number of completed items (such as the Digit Symbol Substitution Test6) or record a distance moved within a specified time frame (such as the six minute walk7) accommodate failure to perform with a score of zero. Sometimes the number of missing items, such as the number of missed tones in hearing tests,8 is itself the outcome.

Proxy respondents are a commonly used alternative data collection source for observable phenomena such as dependence in functional activities.9, 10 Some data can be obtained from proxies when participants are unable to answer for themselves due to cognitive decline, intercurrent illness, or death. Both proxy characteristics and the type of data requested have an impact on reliability. Respondents who live with the participant, as opposed to those who see the participant less often, provide responses that have the highest reliability as proxies for the absent participant.9, 10 High caregiver burden can lead to a negative bias in proxy reports of health and function.10 Agreement between proxies and participants is highest for observable phenomena such as functional domains and diagnosed conditions, and lower for factors that are more subjective, such as emotional state and symptoms.9, 11 In general, proxy respondents tend to overestimate the presence of health problems and disability.9−11 To enhance the reliability of proxy measures, it is important to identify a proxy with adequate knowledge of the participant and to use proxies only for measures for which proxy reports have been validated.

Outcome or Dependent Variables

Outcome or dependent variables measure the observed consequences of an exposure or intervention studied. For any study that is not cross-sectional, participants must be monitored over time. Changes in health or intercurrent events may precipitate losses to follow-up and incomplete data. If persons with and without outcome data are different, results will be biased. Missing outcomes also decrease power. For these reasons, the first priority in data collection is to minimize loss of outcome data. Strategies to promote

acquisition of outcome data include use of passively available information, alternative data acquisition for essential data, protocol modifications for follow up data collection, and use of combined outcomes. Passively available data, such as mortality data from the National Death Index, functional status acquired in nursing homes from mandated data sources such as the Minimum Data Set, or health care utilization from Medicare claims data, can be acquired without the direct involvement of the participant. However, many outcomes important to aging research, such as symptoms, depend on participant involvement. For such measures, alternatives include offering alternate sites and methods for data collection and standardized decision tools for determining outcomes. For example, home visits or telephone calls might capture important data on participants who are no longer able to come to a central testing site. While such additional efforts can increase the cost of data collection, their value in reducing bias often outweighs other considerations. For missing data on physical performance or cognitive tests, logical decision processes can allow for unbiased determination of some outcomes. For example, in a recent multi-site trial,12 the main outcome is observed inability to walk 400 meters. This outcome can be defined using a decision logic that states it has occurred in a participant who cannot perform the 400 meter test because he is bedridden or unable to walk 10 feet.

Missing data can occur if the main outcome is measured at a specific time point, such as after 12 weeks or one year in a clinical trial, and the participant could not be assessed at that time. Alternative forms of the outcome variable or analysis strategy can reduce this problem. One alternative form for the outcome variable is "time in state." Examples include use of diaries or activity monitors to define the outcome as proportion of time spent in activity or proportion of restricted activity days.13 Such high-frequency, relatively low burden measures of health, function or symptoms can yield an outcome that is a proportion of observed time in the condition and are less dependent on a specific follow up time. There are still problems with such measures since they are dependent on participant compliance with data recording. If the data are not collected systematically, the final outcome could be an underestimate of the proportion of time in the condition or state. For anticipated events or conditions, a novel approach is "triggered sampling". In this approach, participants are monitored using frequent low-burden assessments, such as telephone calls. If the participant had a change in status, an in-person interview is scheduled to capture more detailed information before the subject may no longer be able to participate.14 Alternate analytic methods, such as repeated measures or survival analysis with time-to-event as the outcome, maximize the use of available data from all participants, even those with incomplete follow-up

Competing events, such as death before a primary outcome event like stroke, pneumonia or disability, can lead to bias because the primary outcome will not have occurred before the participant is out of the study.15 Strategies to address this problem include predefined combined outcomes such as "death or primary outcome" 16 or analyzing data in a manner that accommodates competing risks (discussed more in the analysis section below).15

Independent Variables

Independent variables include both the primary intervention or risk factor and covariates representing potential confounders, mediators, or moderators. Missing data for the primary independent variable are more difficult to manage, and thus, the overall priority sequence for data collection is: outcome variable, primary independent variable, then other variables.

Since many geriatric problems are multifactorial,17 studies may include multiple independent (predictor) measures. When many factors must be assessed, participants with worse health and function will have more difficulty completing all assessments due to fatigue and the increased duration of data collection when many responses or tasks take

longer.[18] Independent measures can be prioritized so that the most critical are captured first.[19] Measures not expected to change, such as gender or education, should be assessed only once in a longitudinal study.[20] To reduce fatigue, data collection can be paced with time for breaks, distributed across several encounters, and divided among telephone, in home and on-site encounters.[19]

INTERVENTIONS

While missing data most often refers to data that were included in the study protocol, but not actually collected, missing data can also include data that were never included in the protocol, but are necessary for interpreting the study. Frequently overlooked types of data include reasons for missing data, details of study participation, and aspects of blinding. Codes for reasons for missing measures or study withdrawal help evaluate the potential for bias. Study results can be more interpretable if there are measures of adherence to the intervention, the success of blinding in study participants and personnel, and assessment of expectations in participants and controls in unblinded intervention studies. These types of data can also help with data imputation, as discussed further in the analysis section below.

THE PILOT PHASE

Pilot studies provide insights into the characteristics of older participants, estimates of missing data rates for proposed measures, and assessments of the duration of encounters and the prioritization of measures. This is the time to identify measures that participants dislike or are unable to perform, which should be modifed or eliminated. The pilot phase is a good time to test the reliability of proxy reports for key measures. Community Advisory Board members can serve as pilot participants to provide feedback about multiple aspects of the data collection process.[21]

IMPLEMENTATION

For all aspects of a study, a key to reducing missing data is to be as flexible as possible within the constraints of scientific rigor. Convenient and flexible follow-up can increase data collection rates.[22] Indicators of impending withdrawal such as difficulty making a study appointment, reports of declining health, or reluctance to complete interviews can be used to identify potential for withdrawal or missing data. When such persons are identified, preventive protocols can increase personal attention and adapt scheduling the participant's needs.[23] It is wise to have pre-established protocols for data collection alternatives, such as home visits, telephone follow-ups, and proxy interviews. Consider further modifying protocols if missing data problems develop during the intervention phase.

DATA MANAGEMENT

Throughout the conduct of the study, it is important to track follow-up assessments and monitor data collection. Data management systems can track participants as they move through the study and generate reports of missing data and late follow-up evaluations.[24, 25] Timely data entry can help detect missing or inconsistent data, which can be used to find problems with measures or protocols. These issues can be addressed promptly by exploring possible causes and alternatives. Remedies might include staff retraining, revised protocols for data collection or revisions of coding systems for missing data.

DATA ANALYSIS

Assess the Magnitude and Impact of Missing Data

How Much Missing Data Is There?—When the number of cases with missing data is small (ex., <5% in larger samples), some statisticians suggest that the observations with missing data can be deleted with no or small biases in the effect estimates.[26] However, if participants with missing data are very different than those with complete data, or if data are missing for key variables, then substantial bias can still result from even a small amount of missing data.[27]

Characterization of Missing Data—Once the missing data are quantified, it is important to identify any systematic patterns. Compare the frequencies of missing data by participant characteristics, such as age, gender, or health status and conditions, to determine whether "missingness" (the presence of missing data) is related to other known factors. Types of missing data are defined in Table 3. Data can be considered "missing completely at random (MCAR)" only if there are no measured or unmeasured differences in the characteristics of those with missing data and those without. Most analytic methods to account for missing data assume that data are either MCAR or missing at random (MAR).[28] If the characteristics or outcomes of participants with missing data differ from those without missing data after adjusting for other measured factors, then data are "missing not at random (MNAR)". Since we have substantial evidence that participants in longitudinal studies who are lost to follow-up have worse outcomes, even after adjustment for baseline characteristics,[29–31] most missing data in clinical studies will be MNAR. It is thus unlikely that data analysis can ever completely adjust for the effects of missing data.

Why is the Data Missing?—There are three types of non-response (Table 4): unit, item, and wave.[26] In unit non-response no data are collected on an individual participant, and there is no way to include the participant in the analysis. Item non-response refers to missing data for individual items due to participant fatigue or inability, or to a participant's reluctance to respond to the item due to privacy issues or other factors. In wave non-response, all data for a given assessment point in a longitudinal study are missing. Codes for reasons for missing items and waves can be developed and recorded. The reasons for item or wave non-response can sometimes be explained using other available data from the study. For example, since proxies can only provide certain types of data,[9–11] performance tests or information that must be self-reported will be missing for known reasons from proxy interviews.

In addition to methods described previously, researchers can anticipate and plan for some conditions that result in item or wave non-response. For example, if at the time of follow up, participants might be in skilled nursing facilities, it might be wise to recruit likely institutions as study sites. In some studies, the majority of medical data are collected at the discretion of the participant's physician. Although study data might be obtained from these routine clinical evaluations, high priority clinical data cannot be assured unless it is collected as a part of the study itself.

One of the primary reasons for missing data in geriatric research is the death of the participant. Because many outcomes of interest in aging, such as disability, often precede death, alternate methods must be used to account for the bias that results when decedents are excluded from analysis. In addition to using death as an outcome or using triggered sampling to collect data prior to death, proxy interviews are often used to collect data about outcomes that occurred between the last study evaluation and death.[32] Unfortunately, even when measurements proximal to death are included in the analyses, failure to incorporate death in the analysis can still bias the results.[33] For example, when an estimate of the

probability of death is not incorporated into analytic models of health status change over time, the results will assume that the trajectory in decedents resembles that of survivors. In general, when health status and death are associated, it is difficult to discriminate between changes due to time versus those related to death.34 Sensitivity analyses can test assumptions about adverse health events prior to death in order to provide an estimate of the potential severity of bias. Graphical methods of sensitivity analysis can provide a more nuanced evaluation.35

Analytic Problems Associated with Missing Data

All missing data decreases the statistical power to detect significant effects. If data are missing in detectable patterns associated with participant or intervention characteristics, the results are less generalizable and may be biased. The calculated point estimate of the effect, its variance (and thus p-values and confidence intervals) or both may be distorted. Because missing data can lead to incorrect interpretation of study results, authors should include a discussion of the amount and reasons for missing data as well as the methods used to handle missing data in the presentation of study results.

Analytic Methods That Include Participants with Partial Outcome Data—Some analytic methods for longitudinal studies can use available data for participants with incomplete follow-up. One common method is survival (time-to-event) analysis, which uses all participants with complete predictors up to the time they either experience the outcome or are censored (lost to follow-up due to death, drop-out, or other factors). Unfortunately, if the censoring is informative (i.e. the censored participants are either more or less likely than those not censored to experience the outcome) then the results may be severely biased. There are no ways to test for informative censoring. For example, if participants in a study of nursing home-acquired pneumonia were censored when they transferred to another care unit, and most transfers were due to increased functional dependence (a risk factor for pneumonia), then censoring would be informative.

For longitudinal studies with multiple outcome assessments per participant, mixed models or generalized estimating equations can include participants as long as they have data on predictors and at least one outcome assessment. Both models, however, have significant limitations when missing data results from death. Mixed models assume the trajectory for the longitudinal response after death is similar to the trajectory among participants who do not die. Generalized estimating equations (GEE)36 can make inferences only on the overall population trajectory for the longitudinal response, but not for individual trajectories. When there is missing data due to death, this population approach makes it difficult if not impossible to sort out the associations among population trajectories, individual trajectories, and the risk of death.

Several advanced statistical methods can be used to account more specifically for data missing due to death. Shared latent variable models use two linked models, one for the change over time and one for measurement cessation, and assume that measurement cessation and longitudinal change are independent after adjustment for other covariates.37–39 Although this conditional independence assumption between change and cessation may not be always satisfied, shared latent variable models are more appropriate than other options (pattern mixture40 and selection models41) when missing data are caused by death. A particularly useful example of the shared latent variable technique is Gao and colleagues' analyses of longitudinal dementia.39, 42

Can We Ignore The Missing Data Mechanism?—Standard analytic approaches to missing data assume that the missing data mechanism is ignorable, i.e. that the data are

MCAR or MAR.[28] Modeling data that are non-ignorable (MNAR) requires very good prior knowledge about the mechanism that caused the missing data, so that the missing data process can be modeled as a component of the of the overall estimation process.[28] Because knowledge of the mechanism is rarely available and there is no general method or statistical software to model missing data mechanisms,[28] the best method to handle non-ignorable data is to prevent it. Formal statistical tests of non-ignorability have recently been developed.[43, 44]

Analytic Approaches to Missing Data

Listwise Deletion—The default response of most statistical software packages to missing data is to delete all observations with any missing items. This listwise deletion results in reduced power, a skewed referent population, and, if the data are not MCAR, incorrect variances and biased effect estimates.[26–28] As a rule of thumb, if any variable has more than 5% missing values, listwise deletion should not be used.[26]

Dummy Variable Adjustment—Including a dummy variable for missingness is an intuitively appealing method for handling missing data on predictors, but it has been shown to always result in biased estimates even when data are missing completely at random. [45, 46] It should not be used.

Imputation—Imputation methods assign plausible values to missing data.[28] Single imputation methods substitute a single value for a missing value and include replacement with mean, regression imputation, hot-deck, maximum likelihood estimation, propensity scoring and approximate Bayesian bootstrap.[26, 28] Most of these methods incorporate multiple assumptions and can lead to biased estimates when these assumptions are not met. Last observation carried forward, a technique used commonly in longitudinal clinical trials, leads to biased estimates of both effects and variances, even when the data are missing at random, and cannot be recommended.[47] The most commonly used method, maximum likelihood estimation,[28] assumes missing values are MAR, but often results in artificially reduced variances and can lead to over-correction or modeling of noise.

Multiple imputation addresses the underestimation of variance that occurs with single imputation by representing missing data uncertainty.[26–28] Most methods assume that variables are normally distributed and can be represented by a linear function of all the other variables, and only produce unbiased results when the data are MAR or MCAR. The basic method involves replacing each missing value with a set of plausible values (based on correlated variables), resulting in multiple different complete data sets. Each set is then analyzed using standard procedures and the results are combined, yielding correct variance and parameter estimates.

Use of Missingness Screens—Missingness screens are new statistical techniques that help address the impact of missing data and provide guidance in regression modeling and model selection. A two-step approach to model selection in the presence of missing data is recommended. First, a complete case analysis is performed to eliminate variables that have weak associations with the outcome or strong correlations among themselves, and thus to yield a manageable group of candidate variables. Given appropriate results from the missingness screens,[43] multiple imputation can be used. Next, a second step of model selection should be undertaken on each imputed dataset.[48]

Reporting on missing data in publications

In order to understand the magnitude and impact of missing data on evidence, authors and readers of manuscripts should attend to key elements as described in Table 5. In general, the

magnitude of missing data should be reported; participants with missing data, especially primary outcomes, should be compared to participants who had the data; the analytic approach to missing data should be explicitly disclosed; and the potential impact of missing data on the interpretation of study findings should be considered in the discussion section. These elements will allow everyone who is interested in the evidence to weigh the potential for bias in the findings.

SUMMARY

Missing data present a serious challenge to researchers in the field of aging. The best way to handle missing data is to prevent it by careful attention to study design and implementation. The most effective preventive strategies include 1) develop plans to minimize missing data throughout every phase of research; 2) be prepared to adapt to participant needs; 3) monitor missing data during the study; and 4) plan for additional resources to support efforts that reduce missing data. While there are limits to the role of statistics in correcting potential biases due to missing data, it is possible to assess the magnitude and patterns of missing data and to consider their effects on the interpretation of the results.

Acknowledgments

References

1. Di Bari M, Williamson J, Pahor M. Missing-data in epidemiological studies of age-associated cognitive decline. J Am Geriatr Soc. 1999; 47:1380–1381. [PubMed: 10573452]

2. Atkinson HH, Rosano C, Simonsick EM, et al. Cognitive Function, Gait Speed Decline, and Comorbidities: The Health, Aging and Body Composition Study. J Gerontol A Biol Sci Med Sci. 2007; 62:844–850. [PubMed: 17702875]

3. Rochon PA, Berger PB, Gordon M. The evolution of clinical trials: Inclusion and representation. CMAJ. 1998; 159:1373–1374. [PubMed: 9861206]

4. Mody L, Miller DK, McGloin J, et al. Recruitment and retention of older adults in aging research. J Am Geriatr Soc. (In Press).

5. Guralnik JM, Simonsick EM, Ferrucci L, et al. A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol. 1994; 49:M85–94. [PubMed: 8126356]

6. Wechsler, D. Wechsler Memory Scale-Revised Manual. New York: Psychological Corporation; 1987.

7. Troosters T, Gosselink R, Decramer M. Six-minute walk test: A valuable test, when properly standardized. Phys Ther. 2002; 82:826–827. [PubMed: 12147012]

8. Lichtenstein MJ, Bess FH, Logan SA. Validation of screening tools for identifying hearing-impaired elderly in primary care. JAMA. 1988; 259:2875–2878. [PubMed: 3285039]

9. Magaziner J, Zimmerman SI, Gruber-Baldini AL, et al. Proxy reporting in five areas of functional status. Comparison with self-reports and observations of performance. Am J Epidemiol. 1997; 146:418–428. [PubMed: 9290502]

10. Neumann PJ, Araki SS, Gutterman EM. The use of proxy respondents in studies of older adults: Lessons, challenges, and opportunities. J Am Geriatr Soc. 2000; 48:1646–1654. [PubMed: 11129756]

11. Magaziner J, Bassett SS, Hebel JR, et al. Use of proxies to measure health and functional status in epidemiologic studies of community-dwelling women aged 65 years and older. Am J Epidemiol. 1996; 143:283–292. [PubMed: 8561163]

12. Rejeski WJ, Fielding RA, Blair SN, et al. The lifestyle interventions and independence for elders (LIFE) pilot study: Design and methods. Contemp Clin Trials. 2005; 26:141–154. [PubMed: 15837437]

13. Gill TM, Desai MM, Gahbauer EA, et al. Restricted activity among community-living older persons: Incidence, precipitants and health care utilization. Ann Intern Med. 2001; 135:313–321. [PubMed: 11529694]

14. Fried TR, Byers AL, Gallo WT, et al. Prospective study of health status preferences and changes in preferences over time in older adults. Arch Intern Med. 2006; 166:890–895. [PubMed: 16636215]

15. Satagopan JM, Ben-Porat L, Berwick M, et al. A note on competing risks in survival data analysis. Br J Cancer. 2004; 91:1229–1235. [PubMed: 15305188]

16. Ferrucci L, Guralnik JM, Studenski S, et al. Designing randomized, controlled trials aimed at preventing or delaying functional decline and disability in frail, older persons: a consensus report. J Am Geriatr Soc. 2004; 52:625–634. [PubMed: 15066083]

17. Inouye SK, Studenski S, Tinetti ME, et al. Geriatric syndromes: clinical, research, and policy implications of a core geriatric concept. J Am Geriatr Soc. 2007; 55:780–791. [PubMed: 17493201]

18. Chatfield MD, Brayne CE, Matthews FE. A systematic literature review of attrition between waves in longitudinal studies in the elderly shows a consistent pattern of dropout between differing studies. J Clin Epidemiol. 2005; 58:13–19. [PubMed: 15649666]

19. Berkman CS, Leipzig RM, Greenberg SA, et al. Methodologic issues in conducting research on hospitalized older people. J Am Geriatr Soc. 2001; 49:172–178. [PubMed: 11207872]

20. Cornoni-Huntley, J.; Brock, DB.; Ostfeld, AM., et al., editors. Established Populations for Epidemiologic Studies of the Elderly: Resource Data Book. Bethesda, MD: National Institute on Aging; 1986.

21. Ross F, Donovan S, Brearley S, et al. Involving older people in research: methodological issues. Health Soc Care Community. 2005; 13:268–275. [PubMed: 15819748]

22. Good M, Schuler L. Subject retention in a controlled clinical trial. J Adv Nurs. 1997; 26:351–355. [PubMed: 9292370]

23. Cassidy EL, Baird E, Sheikh JI. Recruitment and retention of elderly patients in clinical trials: Issues and strategies. Am J Geriatr Psychiatry. 2001; 9:136–140. [PubMed: 11316617]

24. Parks PL. Development and use of a database management system for longitudinal research. Comput Nurs. 1989; 7:110–111. [PubMed: 2731065]

25. Fulcher SF, Burris TE. A computerized recall system for clinical trials. Ann Ophthalmol. 1988; 20:10–16. [PubMed: 3345053]

26. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999; 8:3–15. [PubMed: 10347857]

27. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2. New York: John Wiley & Sons; 2002.

28. Allison, PD. Missing Data. In: Lewis-Beck, MS., editor. Quantitative Applications in the Social Sciences. Thousand Oaks: Sage Publications; 2002.

29. Hille ET, Elbertse L, Gravenhorst JB, et al. Nonresponse bias in a follow-up study of 19-year-old adolescents born as preterm infants. Pediatrics. 2005; 116:e662–666. [PubMed: 16263980]

30. Pennefather PM, Tin W, Clarke MP, et al. Bias due to incomplete follow up in a cohort study. Br J Ophthalmol. 1999; 83:643–645. [PubMed: 10340968]

31. Vestbo J, Rasmussen FV. Baseline characteristics are not sufficient indicators of non-response bias follow up studies. J Epidemiol Community Health. 1992; 46:617–619. [PubMed: 1494079]

32. Koster A, Patel KV, Visser M, et al. Joint effects of adiposity and physical activity on incident mobility limitation in older adults. J Am Geriatr Soc. 2008; 56:636–643. [PubMed: 18284534]

33. Diehr P, Patrick DL. Trajectories of health for older adults over time: accounting fully for death. Ann Intern Med. 2003; 139:416–420. [PubMed: 12965968]

34. Diehr P, Williamson J, Burke GL, et al. The aging and dying processes and the health of older adults. J Clin Epidemiol. 2002; 55:269–278. [PubMed: 11864798]

35. Hollis S. A graphical sensitivity analysis for clinical trials with non-ignorable missing binary outcome. Stat Med. 2002; 21:3823–3834. [PubMed: 12483769]

36. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.

37. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997; 53:330–339. [PubMed: 9147598]

38. Lin HQ, McCulloch CE, Mayne ST. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. Stat Med. 2002; 21:2369–2382. [PubMed: 12210621]

39. Gao SJ. A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. Stat Med. 2004; 24:211–219. [PubMed: 14716723]

40. Pauler DK, McCoy S, Moinpour C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. Stat Med. 2003; 22:795–809. [PubMed: 12587106]

41. Touloumi G, Pocock SJ, Babiker AG, et al. Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. Stat Med. 1999; 18:1215–1233. [PubMed: 10363341]

42. Shen C, Gao S. A mixed-effects model for cognitive decline with non-monotone non-response from a two-phase longitudinal study of dementia. Stat Med. 2007; 26:409–425. [PubMed: 16345034]

43. Troxel AB, Ma G, Heitjan DF. An index of sensitivity to nonignorability. Stat Sin. 2004; 14:1221–1237.

44. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc. 1988; 83:1198–1202.

45. Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. J Am Stat Assoc. 1996; 91:222–230.

46. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol. 1991; 134:895–907. [PubMed: 1670320]

47. Liu G, Gould AL. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. J Biopharm Stat. 2002; 12:207–226. [PubMed: 12413241]

48. Van Ness P, Murphy T, Araujo K, et al. The use of missingness screens in clinical epidemiologic research has implications for regression modeling. J Clin Epidemiol. 2007; 60:1239–1245. [PubMed: 17998078]

Table 1

Why is missing data a special problem in clinical aging research?

Contributors	Consequences
High rates of intercurrent events, including deaths	Biased results and conclusions
Disability or illness that interferes with data collection	Reduced generalizability
	Reduced power
	Reduced range of effects

Table 2

Overview of Strategies to Prevent and Manage Missing Data

Phase of research	Main strategies
Research Question and Study Design	When planning the study, consider how the research question, the target population and key variables can be defined to promote both high rates of complete data and a representative population
	Adapt the frequency of study visits, sites, and duration of participation to the capabilities of the target population
All measures	Anticipate data collection needs of participants with varying health and function.
	Anticipate the need for proxy informants. Identify potential proxies at enrollment and use key measures that have been validated for proxy use when possible
	Code reasons for missing data, especially inability to perform a test
Outcome Measures	Prespecify alternate data collection strategies to use when the primary strategy fails
	Prespecify alternate definitions and logical sequences for adjudication of major outcomes
	Anticipate need for combined outcomes
	Consider alternatives to a single fixed time point for outcome assessment
Predictor Measures	Prioritize data collection sequence
Intervention	Measure adherence and fidelity to treatment protocol
	Measure success of blinding in participants and study personnel
	Measure expectations in controls, especially if trial participants are not blinded
Pilot studies	Assess problems with data collection
	Revise study plans to reduce problems with data collection
Implementation	Plan for flexibility in schedules, sites and protocols.
	Have protocols for identifying participants at risk of missing data
	Be prepared to modify protocol if missing data problems develop
Data management	Develop and implement real time tracking and reporting system for missing data
Missing data assessment	Quantify amount of missing data (problems minor when < 5%)
	Characterize missing data rates by items, waves, and participants
	Examine potential reasons and mechanisms for missing data
	Compare participants with and without types of missing data to assess potential biases
Analysis	Weigh analytic options in the context of the limitations of each
	Determine whether imputation can be used for some missing data
	Perform sensitivity analyses to examine potential biases due to missing data

Table 3

Mechanisms of Missing Data in Clinical Research

Mechanism	Definition	Example	Prevalence
Missing Completely at Random (MCAR)	If the likelihood of being missing is not related to either the value of the missing variable or to the values of any other variables in the data set.	A set of samples are "lost" due to lab error or an instrument is wrongly calibrated for a day on which random sample of subjects were measured.	Almost never occurs.
Missing at Random (MAR)	If the likelihood of missing data can be completely explained by other variables in the analysis.	The probability of missing data on ADL can be explained by cognition, comorbidity, and living arrangements.	Other data can sometimes provide a good prediction, but missingness is rarely completely explained.
Missing Not at Random (MNAR)	If missing values are not randomly distributed across participants, and the probability of being missing cannot be predicted from the other variables.	The probability of missing data on CESD is related to cognitive status, which was never measured.	Most missing data.

Table 4

Types of Non-response

Type	Definition	Example
Unit non-response	No information is collected	Person did not return survey
Item non-response	Partial data available	Either part of a scale, such as CESD is missing or the person stopped the interview or testing before the end.
Wave non-response	Missing for some waves in a longitudinal study	Person was missing for one or more whole follow-up visits in a longitudinal study.

Table 5

Reporting on Missing Data in Research Manuscripts

1	Report participant sample size at key study time points
2	Provide sample sizes for measures when they are not the same as the participant sample size
3	Compare characteristics of participants who completed the study versus those who did not
4	Describe and justify analytic strategies used to either delete or impute observations and data
5	Consider sensitivity analyses to examine potential biases of missing data and chosen analytic strategies
6	Discuss how missing data might have influenced study findings.