ELSEVIER

Journal of Substance Abuse Treatment 37 (2009) 54 – 63

Journal of Substance Abuse Treatment

Regular article

The impact of loss to follow-up on hypothesis tests of the treatment effect for several statistical methods in substance abuse clinical trials

Sarra L. Hedden, (Ph.D.)[a,b,c,*], Robert F. Woolson, (Ph.D.)[a], Rickey E. Carter, (Ph.D.)[a], Yuko Palesch, (Ph.D.)[a], Himanshu P. Upadhyaya, (M.B.B.S., M.S.)[b], Robert J. Malcolm, (M.D.)[b]

[a]*Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC 29425, USA*
[b]*Department of Psychiatry and Behavioral Research, Medical University of South Carolina, Charleston, SC 29425, USA*
[c]*Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA*

Received 20 March 2008; received in revised form 23 September 2008; accepted 24 September 2008

Abstract

"Loss to follow-up" can be substantial in substance abuse clinical trials. When extensive losses to follow-up occur, one must cautiously analyze and interpret the findings of a research study. Aims of this project were to introduce the types of missing data mechanisms and describe several methods for analyzing data with loss to follow-up. Furthermore, a simulation study compared Type I error and power of several methods when missing data amount and mechanism varies. Methods compared were the following: Last observation carried forward (LOCF), multiple imputation (MI), modified stratified summary statistics (SSS), and mixed effects models. Results demonstrated nominal Type I error for all methods; power was high for all methods except LOCF. Mixed effect model, modified SSS, and MI are generally recommended for use; however, many methods require that the data are missing at random or missing completely at random (i.e., "ignorable"). If the missing data are presumed to be nonignorable, a sensitivity analysis is recommended.

Keywords: Substance abuse clinical trial; Loss to follow-up; Treatment effect; Hypothesis tests; Longitudinal data

1. Introduction

In substance abuse clinical trials, much of the missing data are due to losses to follow-up, that is, individuals who drop out of the study after randomization to treatment and whose data are lost thereafter. This type of dropout is differentiated from individuals who may discontinue treatment (clinical dropout) but who are followed throughout the duration of the study, a scenario common with intention-to-treat analyses, or individuals who are lost but later located for future or follow-up assessment (intermittent missing data). In this article, we discuss the impact of losses to follow-up on hypothesis tests of the treatment effect where missing data are prevalent in outpatient substance abuse clinical trials. This is particularly warranted given that many substance abuse treatment clinical trials demonstrate substantial losses to follow-up after the first dose of treatment with missing data percentages ranging from 10% to 50% (Dutra et al., 2008; Edwards & Rollnick, 1997; Higgins & Budney, 1997; Howard, Cox, & Saunders, 1990; Mattson et al., 1998; McRae, Hedden, Carter, Malcolm, & Brady, 2006; Nich & Carroll, 2002).

This high percentage of loss to follow-up in substance abuse research may interfere with the evaluation of treatment programs and could call into question the validity of study findings. Furthermore, extant research suggests that individuals who are lost to follow-up tend to have poorer functioning compared to individuals who complete treatment, thereby biasing treatment outcome (Sobell, Sobell, & Maisto, 1984). Traditional methods of longitudinal data

* Corresponding author. Johns Hopkins University Bloomberg School of Public Health, Department of Mental Health, Baltimore, MD 21205, USA. Tel.: +1 410 502 9515; fax: +1 410 955 0237.
E-mail address: shedden@jhsph.edu (S.L. Hedden).

0740-5472/08/$ – see front matter
doi:10.1016/j.jsat.2008.09.011

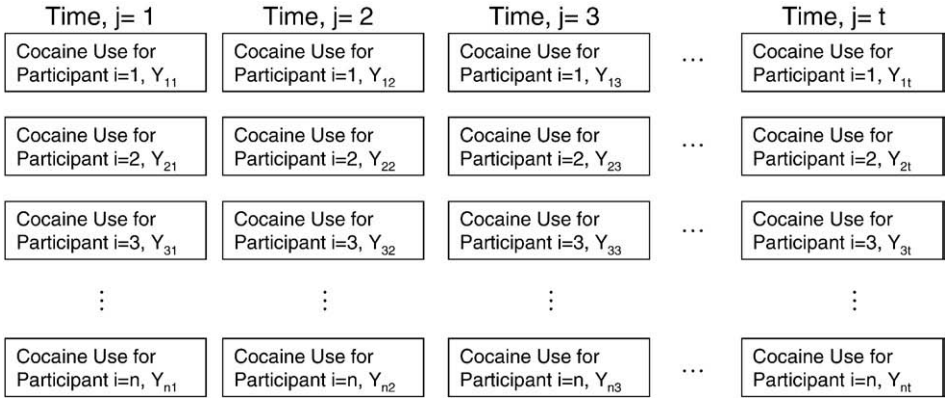

Fig. 1. Representation of a complete longitudinal dataset for, *n*, individuals and, *t*, time points.

analysis such as data deletion (complete case analysis) or single imputation may be biased or otherwise invalidated in the presence of substantial missing data (Figueredo, McKnight, McKnight, & Sidani, 2000). Furthermore, the performance of various statistical methods for missing data needs to be evaluated in the presence of different missing data scenarios and, in particular, in scenarios involving substantial losses to follow-up, as seen in substance abuse clinical trials.

Therefore, the aim of this study is to describe the types of missing data mechanisms that may occur in substance abuse clinical trials, as well as various methods used to test hypothesis of the treatment effect when missing data are present. Methods extensively described in this article include, but are not limited to, stratified summary statistics (SSS), imputation, and mixed effect models. Further, a Monte Carlo simulation study is performed to compare several of the methods described in terms of their Type I error and power. We end with recommendations for the analysis of longitudinal data in the presence of loss to follow-up.

The primary purpose of this article is to inform the substance abuse researcher of the various types of missing data that may occur in substance abuse clinical trials as well as describe a variety of methods that may be used for missing data scenarios. We do this to aid researchers in applying modern and appropriate longitudinal analysis when individuals are lost to follow-up.

1.1. Missing data mechanisms

Subjects who miss a study visit in substance abuse research are often lost thereafter. This complete loss to follow-up gives rise to the probability that the missing data mechanism is not random and may be dependent upon observed and unobserved values of the outcome. Several missing data mechanisms defined by their dependence on observed and unobserved data points, including the outcome, have been classified by Rubin (1987). The specific case of missing data mechanisms for data that are lost to follow-up in the longitudinal setting has been further described by Schafer and Graham (2002) and is similarly expressed below.

To describe missing data mechanisms, some notation and illustrations are helpful. Assume that the outcome variable (e.g., amount of cocaine used) can be measured on each participant repeatedly over the course of the study. In mathematical notation, let Y_{ij} represent the jth $j = 1,\ldots,t$, measurement of the outcome variable for the ith participant, $i = 1,\ldots,n$. Fig. 1 demonstrates this notation for a longitudinal study for n individuals over t points in time. This illustration serves as the framework from which all missing data mechanisms will be discussed.

It should be noted that Fig. 1 demonstrates the case where all individuals' outcomes are measured at all points in time, that is, a complete dataset. However, should an individual miss a visit, this can also be defined. A missed visit, m, can be defined as the time at which a subject drops out of a study and does not return, $j = m$.

Missing data can further be classified by its dependence on observed and/or unobserved outcomes, that is, a missing data mechanism may be defined. The most ideal missing data scenario occurs when data are missing completely at random (MCAR). This occurs when the missing data are independent of any outcome variables and any covariates of interest (Rubin, 1987; Schafer & Graham, 2002). For example, suppose a participant is followed continuously through the first $(m - 1)$ visits but fails to return to the study starting at visit m. In this setting, the values for the outcome variables are unobserved from week m to the end of the study, t (i.e., $Y_{im},\ldots, Y_{it}$ are all unobserved). For this missing data to be MCAR, the missing outcome data must not be associated with any observed or unobserved data point. A practical example of such a scenario is when a participant changes location during the course of the study for reasons completely unassociated with the study and or disease (e.g., work transfer, military deployment, etc.).

The second most common missing data mechanism is missing at random (MAR). MAR occurs when the missing outcome is dependent on any of the observed outcomes, $y_{i1},\ldots,y_{i(m-1)}$, or observed covariates until the time of the