Development of Longitudinal Data Analysis in Mental Health Research For Military Service Members

Xian Liu*†; Daniel P. Evatt*†; Bradley E. Belsher*†

ABSTRACT
Introduction: Since the 1991 Gulf War, mental health conditions of military Service members have received increasing public attention and are a major focus for the U.S. government. A substantial proportion of military health research has been devoted to examining the pattern of change over time in mental health symptoms and diagnostic status among Service members. Unfortunately, many researchers continue to use somewhat obsolete methods to analyze trends and transitions in mental health, despite advances in statistical methodology that permit attention to the unique features inherent in longitudinal data. Materials and Methods: This article defines and describes data features and structures, and basic specifications of longitudinal data analysis to military health researchers. In particular, we highlight the respective impacts of missing data and intra-individual correlation on longitudinal data analysis. Based on the descriptions of the basic features in longitudinal data, we introduce several popular techniques to analyze a variety of longitudinal data types. Results: We demonstrate that traditional analytic techniques do not properly account for missing data and intra-individual correlation inherent in longitudinal data. Failure to use correct, appropriate models and methods can result in major bias in analytic results and mental health predictions. Conclusions: Failure to use correct, appropriate models and methods in longitudinal data analysis can have unfortunate repercussions on a military health system that needs accurate findings to support valid policy decisions. By applying adequate models and methods, military health researchers will be able to better understand the complex interactions of biological, psychological, and social factors on mental health trends and transitions among military Service members.

SIGNIFICANCE OF LONGITUDINAL DATA ANALYSIS
The prolonged wars in Iraq and Afghanistan have produced elevated rates of mental health conditions among active duty Service members. Rates of post-traumatic stress disorder (PTSD) and depression among Service members who served in Iraq and Afghanistan are between 5% and 20%, and one-third return with some type of self-reported mental health or cognitive problem. The developmental course of mental health conditions likely reflects a complex interaction among predispositional, neurobiological, and environmental factors. A better understanding of the onset and course of mental health conditions is critical to prevention and treatment efforts. However, while considerable research has been devoted to cross-sectional descriptions of how mental health conditions are distributed among Service members, there has been a dearth of research properly controlling for factors that affect the onset, course, and management of mental health conditions over time.

In the developmental course of mental health conditions, disease onset and pattern of change over time can be influenced and determined by various risk and interrelated factors, such as socio-demographics, physical illness, and exposure to stressful life events. Consequently, mental health trajectories of Service members can differ significantly, determined by the variables that govern the timing and the rate of change in the developmental process of mental health conditions. This is a particularly important consideration for military Service members who are exposed to unique challenges and stressors over the course of their career. These may include combat-related exposures that may accompany deployment, and other military-specific factors and responsibilities that can generate considerable effects on longitudinal mental health outcomes. Given the complexities associated with the onset and progression of mental health conditions, military health researchers have become increasingly interested in examining the variables that influence the development of these conditions. The military mental health research that contains theoretically relevant predictors can inform policy-makers and planners who seek evidence-based approaches to minimize long-term health consequences to uniformed personnel.

The traditional approach to the analysis of military health research typically involves the cross-sectional examination of data; i.e., analysis of data collected at a single point of time. This approach, by definition, however, does not allow a display of change in mental health conditions and its pattern over time. Cross-sectional data only designate a snapshot of a long-term developmental course and do not possess...
capacity to reflect change, growth, or development of mental health conditions among military Service members. Aware of the limitations in cross-sectional studies, some military health agencies have advanced the data management perspective by collecting data with repeated measurements of the same Service members.10–11 By measuring the same variable of interest repeatedly, a pattern of change in mental health over time can be revealed, and constructive findings derived about the dynamic nature of mental health conditions. If repeated measures are organized appropriately into a longitudinal dataset, analytic approaches are available to examine trends and transitions in mental health outcomes while taking into account missing data, covariates, and correlation of multiple assessments for the same person.12–15 These approaches have been specifically developed to address limitations and weaknesses in the application of traditional methods in longitudinal data analysis.

In the following sections, we describe various longitudinal data structures, present appropriate use of each, and discuss the benefits and drawbacks of various approaches to analyzing these data to characterize mental health conditions among military personnel.

LONGITUDINAL DATA STRUCTURES
Methodologically, longitudinal data can be defined as the data of repeated measurements at a limited number of time points with predetermined designs on time scale, time interval, and other-related conditions.14–15 In the military health system, there is a long history of administrative health data recorded, archived, processed, and generally categorized by type of record and year.16 Additionally, information about deployments, sociodemographic characteristics, and other non-health statistics are recorded for each Service member enrolled in the military health administrative system. Such large-scale administrative health data can be readily reorganized and restructured into longitudinal datasets for scientific use. More recently, some DoD related agencies started collecting systematic, population-based longitudinal data for military Service members by following rigorous procedures of probability sampling, questionnaire design, and measurement specifications for representative surveys.17–18 The availability of the administrative and the observational longitudinal datasets can potentially facilitate military health researchers to perform longitudinal data analysis to assess trends and transitions in mental health conditions among Service members.

Univariate and Multivariate Data
Longitudinal data can be structured by two formats – multivariate and univariate. Classical repeated measure models and modern latent growth modeling follow the multivariate data structure. In this data structure, each individual is assigned a single row of data, with repeated measurements being recorded horizontally; i.e., a column is assigned to the measurement at each time point in the data matrix. Table I illustrates this arrangement using an example of repeated measures data on a mental status score, named MSS, at four-time occasions with two covariates (COV1 and COV2). In the multivariate data structure, we can see in Table I that the repeated measurements of MSS for each individual are specified as four variables lined in the same row, with time points indicated as suffixes attached to the variable name (e.g., MSS1, MSS2, MSS3, and MSS4). Given all four observations for this variable being lined horizontally, the multivariate data structure of repeated measurements contains additional columns, therefore also referred to as the wide table format. By using the multivariate data structure, the most distinctive advantage is that each subject’s empirical growth record can be visually examined and assessed.19

The multivariate data structure, however, has distinct disadvantages for performing longitudinal data analysis. In the multivariate format, the time factor is designated by the suffix attached to each time occasion. An analysis of the time effect is inconvenient because time is not explicitly specified as an independent factor. Sometimes, intervals between two successive waves are designed to be unequally spaced or vary across individuals, and the multivariate data structure cannot reflect such variations in time interval spacing. Additionally, in longitudinal data analysis, values of some covariates may vary over time which, in turn, can generate unique influences on the changing pattern of mental health conditions. Failure to address such time-varying effects in predictor variables can result in biased parameter estimates and erroneous model-based predictions on mental health outcomes. There are some complex, cumbersome ways to specify time-varying covariates within the multivariate data structure; these approaches, however, are not user-friendly.14

Given the aforementioned disadvantages in the multivariate data structure, a majority of modern longitudinal analyses rely on data of a univariate structure illustrated in Table II. In the univariate format, each subject is assigned a block that contains multiple rows in the data matrix, and time is explicitly specified as a primary predictor on the trajectory of individuals. Each subject has multiple lines of data on the outcome variable, with each record corresponding to a specific time point. Consider the aforementioned example again. In Table II, the repeated measurements of the MSS score now are set vertically under the same variable name MSS, with suffixes removed. A new covariate, TIME, is added to the data matrix to indicate a specific time point, and a combination of values for the MSS and the time variables reflects

| TABLE I. Multivariate Data of Repeated Measurements |
|----------------|---|---|---|---|---|---|
| ID   | MSS1 | MSS2 | MSS3 | MSS4 | COV1 | COV2 |
| 1    | 66   | 31   | 58   | 39   | 27   | 0    |
| 2    | 48   | 56   | 43   | 44   | 44   | 1    |
| 3    | 37   | 50   | 53   | 47   | 38   | 0    |
repeated measurements at a number of time points. At the same time, values of covariates (COV1 and COV2) can be specified differently across various time points so that time-varying effects of those variables can be captured in longitudinal data analysis. Because subject-specific observations are set vertically, the researcher needs to specify fewer columns but more rows than for the multivariate data structure. Accordingly, the univariate longitudinal data structure is also referred to as the long table format. In this long-table design, the observation at a given time for a given individual, not the individual per se, is the unit of analysis for longitudinal data analysis.

**PRIMARY FEATURES OF LONGITUDINAL DATA**

**Missing Data**

Given the complex structure of longitudinal data, performing longitudinal data analysis is not an easy, straightforward undertaking. The most troublesome feature of longitudinal data is the presence of missing data in repeated measurements. For the military health data administrative system in particular, the loss of observations on the variables of interest occurs frequently. In a clinical trial on the effectiveness of a new medical treatment for a mental health condition, Service members may be lost to follow-up due to deployment, retirement, or health problems. In a longitudinal observational survey of active duty military personnel, some baseline respondents may be unavailable due to unforeseen events such as trainings or emerging occupational requirements. With high fluidity among military Service members, missing data problems can be especially serious, which can cause concern for military health researchers in conducting longitudinal data analysis.

A rich body of literature regarding longitudinal data analysis is devoted to the classification of missing data into different types. Some missing data are not related to either the outcome or the explanatory variables and thus represent a random sample of all cases; this type of missing data is referred to as missing completely at random (MCAR). More often, missing observations are not random but the information inherent in missingness is related to certain observed variables such as deployment, physical health functioning, or sociodemography. This type of missing data is referred to as missing at random (MAR). In the above two situations, because missing data do not pose serious threats to the quality of a longitudinal data analysis, they can be ignored. In some special circumstances however, missing data are related to missing values of the outcome variable. Ignoring such systematic missing data can cause substantial bias in parameter estimates, and in model-based predictions on longitudinal trajectories of mental health conditions. This non-ignorable type of missing data is referred to as missing not at random (MNAR). Military health researchers might want to understand various missing data patterns and mechanisms before proceeding with a formal longitudinal data analysis.

**Intra-individual Correlation**

Another primary feature in longitudinal data is that the repeated measurements for the same individual tend to be intimately correlated, referred to as intra-individual correlation. Given various biological, genetic, and environmental predispositions, the values of repeated measurements for the same individual tend to be more similar than those obtained from several randomly selected individuals. When multiple assessments of mental health conditions for the same Service member are strongly related, the individual-specific data are clustered within individuals, thereby violating the conditional independence hypothesis. The conditional independence hypothesis, routinely applied in the application of multivariate regression modeling, assumes model residuals to be mutually independent of each other conditional on model parameters. Statistical analyses that violate this requirement can result in seriously biased parameter estimates, including both the point and the standard error estimates, which, in turn, lead to erroneous predictions on longitudinal outcomes.

Scientists have developed two perspectives to handle intra-individual correlation in longitudinal data, each linked to a specific source of variability. In longitudinal processes, variability can be summarized into three components: between-subjects variability, within-subject variations, and the remaining random errors. The first two components address the non-random, systematic variations of longitudinal data. Because between-subjects and within-subject variations are intimately related, usually only one source of systematic variability needs to be considered to make observations conditionally independent. The third component is the random term for uncertainty as routinely specified in general linear models and generalized linear models.

**METHODS OF LONGITUDINAL DATA ANALYSIS**

**Mixed-Effects Modeling**

Over the years, quantitative methodologists have developed advanced techniques to address the weaknesses and limitations in conventional regression models for the analysis of
longitudinal data in which model parameters are fixed and contain non-random quantities. One popular approach is mixed-effects modeling in which some of the effects in regression are specified as random to account for intra-individual correlation. Mixed-effects modeling allows some of the model parameters to vary randomly over individuals, rather than to be fixed. The random parameters so specified are referred to as the random effects, and can include the random effects for the intercept, and the random effects for the time factor. Here, the model parameters are specified as random variables. For the parameter with the random effects, the regression analysis estimates a unique value for each individual to account for the unobserved or unrecognized heterogeneity across individuals. If the individual-specific random effects are correctly specified, intra-individual correlation tends to be explained by those added effect parameters; consequently, residuals in a regression model become conditionally independent thereby yielding valid parameter estimates and model-based predictions on longitudinal processes.

**Specifying Covariance**

Another perspective frequently applied in longitudinal analysis is to specify the covariance structure of within-subject random errors in regression models while leaving the between-subjects random effects unspecified. The use of such a design in longitudinal data analysis becomes necessary when the application of a mixed-effects model does not yield reliable analytic results or when within-subject variability is sizable in comparison with between-subjects variance. In this perspective, residuals are assumed to follow a multivariate normal distribution with zero expectation and a specific covariance matrix across the specified time occasions. With the inclusion of a residual covariance matrix in the regression model, intra-individual correlation is addressed directly thereby validating the conditional independence hypothesis for random errors. To pattern the residual covariance structure across discrete time occasions, the time factor must be specified as a classification factor to reflect the repeated effects. In the literature of longitudinal data analysis, there are a variety of covariance pattern models for use for empirical analyses. However, for nonnormal longitudinal outcome data such as binary, multinomial, or counts, the specification of a residual covariance pattern is not applicable due to the difficulty in defining residuals for those data types.

Numerous statistically complex models and methods have been developed to address intra-individual correlation in analysis of longitudinal data. Popular techniques include linear mixed models to analyze continuous longitudinal data, mixed-effect logistic regression models for the analysis of binary longitudinal outcomes, and mixed-effects multinomial logit models to deal with a longitudinal outcome variable taking more than two qualitative values. The methods to analyze nonnormal longitudinal data are particularly complex; they typically apply Bayesian techniques to approximate the random effects and functional transformation/retransformation steps in model-based predictions. For more details of these approaches, the interested reader is referred to West et al for linear mixed models, to Fitzmaurice et al for mixed-effects logistic models, and to Liu for mixed-effects multinomial logit models.

Considerable caution must be exercised in using those models and methods in longitudinal data analysis. Military health researchers might be more cautious to the application of simple analytic techniques in longitudinal data analysis that do not have sufficient capacity to handle missing data and intra-individual correlation. We contend that few, if any, traditional statistical approaches have sufficient capability to account properly for missing data. For example, the use of simple descriptive analysis to present frequencies of raw data over time can be problematic. These frequency analyses of longitudinal data tend to violate all of the aforementioned concerns and may produce and communicate biased rates of military health problems. In cases of continuous longitudinal data, whereas traditional repeated measures ANOVA may account somewhat for intra-individual correlation and confounding effects, only contemporary models are designed to account for all of them. Table III summarizes the strengths and weaknesses of the traditional approaches and the contemporary linear mixed models in the analysis of continuous longitudinal data.

Compared to the analysis of continuous longitudinal data, in analyzing binary and multinomial longitudinal data failure to properly model the population by accounting for the missing data mechanism and intra-individual correlation can result in more serious bias. To illustrate the possible degree of such bias, Figure 1 plots two sets of longitudinal trajectories of death rates and corresponding 95% confidence limits of older Americans using data from a previously published article. The two sets of longitudinal trajectories are derived from, respectively, the mixed-effects model and the conventional regression model including only the fixed effects. Each of the trajectory sets is compared to a solid growth

<table>
<thead>
<tr>
<th>Strengths and Weaknesses of Statistical Approaches in Analyzing Continuous Longitudinal Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TABLE III</strong>. Missing Data Intra-individual Correlation Confounding Effects Trajectory Over Time</td>
</tr>
<tr>
<td>Descriptive analyses (Frequencies)</td>
</tr>
<tr>
<td>Repeated measures ANOVA</td>
</tr>
<tr>
<td>Conventional multiple linear regression</td>
</tr>
<tr>
<td>Mixed models</td>
</tr>
</tbody>
</table>

*Limited by assumptions regarding normality; −Weakness; + Strength.
curve and a shaded area representing an unbiased 95% confidence region generated from the originally simulated data. In Panel A, which compares predictions of death rate between the mixed-effects model and the simulated, unbiased data, the two sets of trajectories almost coincide, thereby demonstrating the statistical efficiency and coverage of the mixed model. Panel B displays the predicted trajectories generated from the conventional multinomial logit model without specifying the random effects. Here, substantial and systematic bias can easily be identified. In later time points, the predicted probabilities are strongly biased: even the lower confidence limit of the probability appears much greater than the solid line.

**DISCUSSION**

Since the 1991 Gulf War, mental health conditions among military Service members have received increasing public attention and is a major focus for the U.S. government. In response to the Department of Defense (DoD) and the Institute of Medicine recommendations for a systematic, longitudinal, population-based assessment of Service members’ health, considerable research has been devoted to mental health conditions among Service members and their changing patterns over time. However, it is not uncommon among military health researchers to use inferior statistical methods to analyze longitudinal health data without paying sufficient attention to the unique features inherent in
longitudinal data. Traditional dated methods may be appealing because they produce and present data in simple ways that can be easily digested by clinicians and policy-makers who may not have advanced statistical training. However, military health researchers might want to increase their awareness of the fact that failure to use correct, valid methods can often result in considerable bias in analytic results and model-based predictions of longitudinal health trajectories. Such bias can have unfortunate repercussions on a military health system that needs accurate findings to support valid policy decisions.

REFERENCES