•Biostatistics in psychiatry (28)•

Introduction to longitudinal data analysis in psychiatric research

Xian LIU[1,2]

Summary: The onset, course, and management of mental health problems typically occur over relatively long periods of time, so a substantial proportion of psychiatric research – particularly the research that can provide clear answers about the complex interaction of biological, psychological, and social factors – requires multiple assessments of individuals and the environments in which they live over time. However, many psychiatric researchers use incorrect statistical methods to analyze this type of longitudinal data, a problem that can result in unrecognized bias in analytic results and, thus, incorrect conclusions. This paper provides an introduction to the topic of longitudinal data analysis. It discusses the different dataset structures used in the analysis of longitudinal data, the classification and management of missing data, and methods of adjusting for intra-individual correlation when developing multivariate regression models using longitudinal data.

Key words: Intra-individual correlation; longitudinal data; missing data; multivariate and univariate data structures; repeated measurements

[*Shanghai Arch Psychiatry.* 2015; **27**(4): 256-259. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.215089]

1. Significance of longitudinal data analysis in psychiatric research

We live in a world full of change. After birth, a person grows, ages, and dies. During such a dynamic process, we may contract various psychiatric disorders, develop functional disability, and lose mental ability. Accompanying the developmental course of psychiatric conditions, physical health may be affected. While poor physical health generally elevates the risk of developing psychiatric problems, the presence of a psychiatric disorder can also lead to an increased prevalence and severity of other diseases such as cardiovascular disease, cancer, and diabetes.[1] Social functioning can also be altered by psychiatric diseases; many individuals with psychiatric disorders experience marital disruption, unemployment, and occupational impediments. In order to understand the complex interaction between these physical, mental, and social processes, it is essential for psychiatric researchers to assess the gradual onset and course of psychiatric conditions over the lifetime of individuals who experience these disorders.

In the developmental course of psychiatric conditions, the pattern of change over time can be influenced and determined by various risk factors, such as genetic predisposition, physical illness, traumatic events, environment, or the like. Therefore, the illness trajectory can differ significantly among individuals due to the presence or absence of factors that can govern the direction, timing, and rate of change. Consequently, much of modern psychiatric research focuses on making comparisons among subgroups of specific populations with the goal of identifying the variables that influence the onset and course of psychiatric diseases. As psychiatric researchers and other medical scientists have placed increasing attention on the inherent mechanisms that are associated with the development of various psychiatric and medical conditions, there has been a corresponding development in the statistical methods and techniques used to describe and analyze underlying features of longitudinal processes. [2,3,4]

Data available at a single point of time cannot be used to analyze change in psychiatric conditions over time. Cross-sectional data, traditionally so popular and so widely used in many applied sciences, only provides a snapshot at one point in time of an ongoing trajectory and, thus, cannot be used to reflect change, growth, or development. Aware of the limitations of cross-sectional studies, many psychiatric researchers have

[1] DoD Deployment Health Clinical Center, Defense Center of Excellence for Psychological Health and Traumatic Brain Injury, Walter Reed National Military Medical Center, Bethesda, Maryland, United States

[2] Department of Psychiatry, F. Edward Hebert School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, Maryland, United States

correspondence: xian.liu@usuhs.edu

The views expressed in this article are those of the author and do not necessarily represent the official position of the government of the United States of America

A full-text Chinese translation of this article will be available at http://dx.doi.org/10.11919/j.issn.1002-0829.215089 on October 26, 2015.

advanced the analytic perspective by examining data with repeated measurements. By measuring the same variable of interest multiple times, the change in mental health is displayed, its pattern over time revealed, and, thus, it may be possible to identify factors that are associated with changes in psychiatric status. Such data with a limited number of repeated measurements are referred to as *longitudinal data*.[5] In many longitudinal data designs in psychiatric research, subjects receiving different interventions or those exposed to different potential risk factors are repeatedly evaluated at a number of time points separated by specified intervals.

2. Longitudinal data structures

Methodologically, longitudinal data can be regarded as a special case of classical repeated measures data. There are some conceptual differences between the two data types. Classical repeated measures data are a broadly defined type of data that can include a large number of time points and changing experimental or observational conditions over the course of the follow-up.[5] In contrast, longitudinal data are more specific. They are generally composed of multiple observations for the same group of individuals at a limited number of time points with equally or unequally spaced intervals. Therefore, 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.

Longitudinal data can be structured as either multivariate or univariate data. Traditionally, the data structure for repeated measurements follows a multivariate format. In this data structure, each individual only has a single row of data, with repeated measurements being recorded horizontally. That is, a column is assigned to the measurement at each time point in the data matrix. Consider an example of the repeated measures data on a posttraumatic stress disorder (PTSD) score. In the multivariate data structure, the repeated measurements of PTSD for each individual are specified as four variables placed in the same row of the data matrix, with time points indicated as suffixes attached to the variable name (e.g., PTSD1, PTSD2, PTSD3, and PTSD4). With all observations for this variable recorded in one row of the data matrix, the multivariate data structure of repeated measurements contains additional columns for each time point, referred to as the *wide table* format. The most distinctive advantage of using the multivariate data structure is that each subject's empirical growth record can be visually examined.[6]

There are, however, distinctive disadvantages of the multivariate data structure in performing longitudinal data analysis. First, in the multivariate format the time factor is indirectly reflected by the suffix attached to the variable name for each time the same assessment is repeated, so time is not explicitly specified as an independent factor, making it difficult to include the effect of time in the analysis. In some cases, assessment intervals between two successive waves are unequally spaced or vary across individuals, variations that cannot be captured using a multivariate data structure. Second, in longitudinal data analysis, values of some covariates may vary over time (e.g., age, marital status, economic status, employment, etc.); failure to address the time-varying nature of predictor variables can result in biased effects and erroneous predictions of longitudinal processes. There are some cumbersome ways to specify time-varying covariates within the multivariate data framework, but these approaches are not user-friendly and are inconvenient to apply.[4]

Given the aforementioned disadvantages in the multivariate data structure, the majority of modern longitudinal analyses are based on data with a univariate structure. In the univariate data format, each subject has multiple rows of data (one row for each time the outcome variable is assessed) and time is explicitly specified as a primary predictor of the trajectory of individuals. In this scenario, the repeated measures of PTSD in the example described earlier would be represented as a single variable that appears in a column within the data matrix, not as separate variables with different suffixes in a single row of the data matrix. A new covariate, TIME, is added to the data matrix to indicate a specific time point, and a combination of values for the PTSD and the time variables designate repeated measurements at a number of time points. As subject-specific observations are set vertically, fewer columns but more rows are specified than in the multivariate data structure. Correspondingly, the univariate longitudinal data structure is also referred to as the *long table* format.

3. Primary features of longitudinal data

Analyzing longitudinal data in psychiatric research poses considerable challenges to biostatisticians and other quantitative methodologists due to several unique features inherent in such data. The most troublesome feature of longitudinal data is the presence of missing data in repeated measurements. In a clinical trial on the effectiveness of a new medical treatment for a psychiatric disease, patients may be lost to a follow-up due to migration or health problems. In a longitudinal observational survey, some baseline respondents may lose interest in participating at subsequent times. There are different types of missing data, some of which do not threaten the quality of the longitudinal analysis and others that do. Missing data that represent a random sample of all cases or non-random missing data that can be accounted for by adjustments using observed variables (such as age, gender, illness severity, etc.) do not pose serious threats to the quality of a longitudinal data analysis. However, in some special circumstances missing data are related to missing values of the outcome variable, and ignoring such systematic missing data can be detrimental to the estimation and prediction of the pattern of change over time in the response variable. Thus, it is important for psychiatric researchers to understand the various types of missing data and the steps that should be taken in conducting

formal longitudinal data analysis when different types of missing data are present in the data set.[7]

Another primary feature in longitudinal data is the correlation in the repeated measurements of the same individual, referred to as *intra-individual correlation*.[4] Such correlation is a violation of the conditional independence hypothesis regularly applied in multivariate regression modeling, so biostatisticians and other quantitative methodologists have developed two primary ways to deal with this issue when performing longitudinal data analysis, each linked to a specific source of variability. Statistically, variability in longitudinal processes can be summarized into three components: between-subjects variability, within-subject variability, and the remaining variability due to random errors. Intra-individual correlation can be modeled by means of the first two components, that is, as either the between-subjects or the within-subject component. These two components are interrelated, so it is usually only necessary to consider one of the two sources of systematic variability to make longitudinal data conditionally independent and, thus, appropriate for use in multivariate regression modeling analysis.[4]

4. Longitudinal analysis

The importance of addressing intra-individual correlation and missing data has triggered the development of many advanced models and methods for longitudinal data analysis. One popular approach is 'mixed-effects modeling'. In this approach unobservable differences between individuals are accounted for by specifying specific effects that can vary over subjects. When conducting the analysis the researcher would specify that some of the parameters in the regression model can vary between subjects (i.e., 'random' parameters) while other parameters do not change between subjects (i.e., 'fixed' effects). The researcher-specified random parameters are referred to as *the random effects*; these can include random effects for the intercept, random effects for the time factor, and so forth. For example, when analyzing the trajectory of the PTSD score in patients receiving different types of treatment, the researcher could specify that each individual has a unique baseline value and a unique pattern of change over time (these added parameters would be the random effects in the model) before developing the model. After specification of the subject-specific random effects in the model, differences between observed and predicted PTSD results in the final regression model are considered conditionally independent (i.e., the basic requirement for reliable multivariate regression modeling) and, thus, the regression coefficients generated in the model of the longitudinal process are usually of high quality.

Another popular approach to longitudinal analysis is to include the pattern of correlation across repeated measurements in the regression model while leaving the between-subjects random effects unspecified. The use of such a design in modeling longitudinal processes becomes necessary when the application of the random-effects approach (above) does not yield reliable analytic results or when within-subject variability is sizable in comparison with between-subjects variance. Applying this approach to our PTSD example, the correlation of the repeated measurements of the PTSD score for the same subject is assumed to follow a known structure, referred to as a 'covariance matrix', that the researcher would specify before conducting the analysis. There is a variety of model covariance patterns (designed over the years by statisticians) that the researcher can select from when conducting the analysis.

The two approaches described above for the continuous response variables can be readily extended to longitudinal modeling of non-normal outcome variables, such as rates and proportions, multinomial outcomes, and count data. There are many statistically complex techniques and methods for modeling these data types.[8,9,10,11,12] Accompanying the rapid developments of statistical models and methods are the equally important advancements in computer science, particularly the powerful statistical software packages. The convenience of using computer software packages to create and utilize complex statistical models has made it possible for medical scientists, psychiatric researchers being no exception, to analyze longitudinal data by applying complex, efficient statistical methods and techniques. Among the variety of computing software packages, the Statistical Analysis System (SAS), a powerful software system for data analysis, consists of a group of computer programs that can be applied for longitudinal analyses on different data types.[13]

5. Conclusion

Over the years, biostatisticians and other scientists have developed a variety of statistical models and methods to analyze longitudinal data. Most of these advanced techniques are built for use in biomedical and behavioral settings, so the methodologically advanced techniques may be relatively unfamiliar to psychiatric researchers. Many psychiatrists still use incorrect statistical methods to analyze longitudinal mental health data without paying sufficient attention to the unique features inherent in such data. Failure to use correct analytic methods can result in tremendous bias in analytic results and outcome predictions. Psychiatric researchers need to familiarize themselves with the advanced models and methods developed specifically for longitudinal data analysis.

Conflict of interest

The author reports no conflict of interest related to this manuscript.

Funding

The preparation of this article was partially supported by the National Institute on Aging (NIH/NIA Grant No.: R03AG20140-01).

Shanghai Archives of Psychiatry, 2015, Vol. 27, No. 4

精神病学研究中纵向数据分析的介绍

Liu X

概述：精神卫生问题的发生、发展及对其管理都需要相对较长的时间，所以相当一部分精神病学研究—— 特别是能够明确回答有关生物、心理和社会因素复杂的相互作用的研究——要求对患者及其生活环境进行跨时长久的多种评估。然而，许多精神病学的研究人员使用不正确的统计方法来分析这一类型的纵向数据，这一问题会导致分析结果中出现无法识别的偏倚而由此得出不正确的结论。本文就纵向数据分析的话题做了介绍。文章探讨了纵向数据分析中使用的不同数据集结构、缺失数据的分类和处理以及使用纵向数据建立多元回归模型时对个体内相关性校正的方法。

关键词：个体内相关性；纵向数据；缺失数据；多元与一元数据结构；重复测量

本文全文中文版从 2015 年 10 月 26 日起在
http://dx.doi.org/10.11919/j.issn.1002-0829.215089 可供免费阅览下载

References

1. Goldberg D. The detection and treatment of depression in the physically ill. *World Psychiatry.* 2010; **9**(1): 16-20
2. Diggle PJ, Heagerty PJ, Liang K, Zeger SL. *Analysis of Longitudinal Data* (2nd ed.). Oxford: Clarendon Press; 2002
3. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Hoboken, New Jersey: Wiley; 2004
4. Liu X. *Methods and Applications of Longitudinal Data Analysis*. New York, NY: Academic Press; 2015
5. West BT, Welch KB, Gałecki AT (with contributions from Gillespie BW). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton, FL: Chapman & Hall/CRC; 2007
6. Singer JD, Willett JB. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press; 2003
7. Little RJA, Rubin DB. *Statistical Analysis with Missing Data (2nd ed.)*. New York, NY: Wiley; 2002
8. Breslow NR, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc.* 1993; **88**(421): 9-25. doi: http://dx.doi.org/10.2307/2290687
9. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986; **73**(1): 13-22. doi: http://dx.doi.org/10.2307/2336267
10. Liu X, Engel CC. Predicting longitudinal trajectories of health probabilities with random-effects multinomial logit regression. *Stat Med.* 2012; **31**(29): 4087-4101. doi: http://dx.doi.org/10.1002/sim.5514
11. McCulloch CE, Searle SR, Neuhaus JM. *Generalized, Linear, and Mixed Models*. Hoboken, NJ: Wiley; 2008
12. Molenberghs G, Verbeke G. *Models for Discrete Longitudinal Data*. New York, NY: Springer; 2010
13. SAS.SAS/STAT *12.1: User's Guide* (2nded.). Cary, NC: SAS Institute Inc; 2012

(received, 2015-07-22; accepted, 2015-07-25)

Dr. Xian Liu is Professor of Research at the Department of Psychiatry and Senior Scientist at the Center for the Study of Traumatic Stress, F. Edward Hebert School of Medicine at the Uniformed Services University of the Health Sciences in Bethesda, Maryland, USA. He also serves as Research Scientist/Senior Statistician in the Deployment Health Clinical Center, Defense Centers of Excellence at Walter Reed National Military Medical Center. His areas of expertise include longitudinal analysis in health research, survival analysis, aging and health, and development of advanced statistical models in behavioral and medical studies. Dr. Liu received his PhD in Sociology with specialization in Demography from the Population Studies Center, the Institute for Social Research at University of Michigan in 1991.