EXTRA
Dementia
and Geriatric
Cognitive Disorders

Dement Geriatr Cogn Disord Extra 2011;1:330–357
DOI: 10.1159/000330228
Published online: October 26, 2011

Review Article

An Overview of Longitudinal Data Analysis Methods for Neurological Research

Joseph J. Locascio[a] Alireza Atri[a, b]

[a]Massachusetts Alzheimer's Disease Research Center, Dept. of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, Mass., and [b]Geriatric Research, Education and Clinical Center (GRECC), ENRM VA Medical Center, Bedford, Mass., USA

Key Words
Analysis · Longitudinal studies · Methods · Neurology · Statistics

Abstract
The purpose of this article is to provide a concise, broad and readily accessible overview of longitudinal data analysis methods, aimed to be a practical guide for clinical investigators in neurology. In general, we advise that older, traditional methods, including (1) simple regression of the dependent variable on a time measure, (2) analyzing a single summary subject level number that indexes changes for each subject and (3) a general linear model approach with a fixed-subject effect, should be reserved for quick, simple or preliminary analyses. We advocate the general use of mixed-random and fixed-effect regression models for analyses of most longitudinal clinical studies. Under restrictive situations or to provide validation, we recommend: (1) repeated-measure analysis of covariance (ANCOVA), (2) ANCOVA for two time points, (3) generalized estimating equations and (4) latent growth curve/structural equation models.

Introduction

Purpose and Background

This paper provides a broad didactic survey of methods for statistical analysis of longitudinal, clinical, observational and experimental data, illustrated by applied examples,

Joseph J. Locascio, PhD
Neurology Department, Memory and Movement Disorders Units,
Massachusetts Alzheimer's Disease Research Center
Massachusetts General Hospital, Warren Building
Office 806, 55 Fruit Street, Boston, MA 02114 (USA)
Tel. +1 617 724 7192, E-Mail JLocascio@partners.org

KARGER

aimed to be of practical utility for clinical researchers with little background in statistical modeling. Substantively, the special focus is on neurological conditions, especially dementia, but the methods are more broadly relevant. We feel there is too often a general lack of understanding and confusion concerning appropriate longitudinal data analysis methods that has bred insecurity towards or prejudice against the use of newer, advanced and more powerful methods among some clinical researchers and journal reviewers of neurological literature. This lack of understanding can lead to inappropriate or inefficient analysis, inaccurate results, and simplistic or wrong interpretations, conclusions, and judgments. While we emphasize that sophisticated and advanced analytic models cannot, and should not, compensate for poor study design and execution, we also maintain that solely using simplistic analytic methods can scuttle detection of important signals and effects, even in well-designed and -conducted studies. In this review, we provide, using an informal and straightforward style, an organized overview of the types of methods available and suggest approaches for situations under which they may be appropriate. While we assume familiarity with basic methods of descriptive and inferential statistics for the biological, medical, and/or behavioral sciences (e.g., analysis of variance, ANOVA, and regression/correlation), our approach does not require specialized or advanced knowledge of statistics or modeling. To be accessible to a wide audience, our format leans toward verbal, intuitive, and graphical presentation with examples, software suggestions, and programming code/scripts in SAS [1] and MPlus [2] software. For readers wanting user-friendly drop-down menus, SPSS [3] and JMP [4] software provide analysis options with some of the advanced modeling techniques reviewed here (e.g., repeated-measure analysis of covariance, ANCOVA, and mixed-effect models).

In our discussions, we focus more on longitudinal observational research (prospective and retrospective), and to a lesser extent on randomized interventions or randomized clinical trials. By 'longitudinal research', we generally restrict ourselves here to data sets where typically each of a moderate/large number of subjects (10 to hundreds) has a relatively small number of repeat readings on a single, continuous, interval-level, numeric measure across time, usually 2–30 observations per subject over time. Depending on the method, the number of observations within subjects can vary across subjects, and the time intervals between observations can vary within as well as between subjects (in examples from our longitudinal studies, observations are often months to >1 year apart).

We also primarily focus examples on characterizing and modeling progression, i.e., assessing different forms of progression, not just computing 'rates' of change that presuppose only linear trajectories over time. Moreover, we focus mostly on research designs in which the dependent variable (outcome) is an essentially continuous numeric variable, the most common case, and where only one is studied at a time (univariate, not multivariate analysis with respect to the dependent variable). Extensions to categorical, binary, count, and ordinal dependent variables can often be dealt with through generalized linear model variants of the longitudinal methods we present, or essentially embedding logistic regression, Poisson models, or log-linear techniques, for example, within the methods we discuss, employing intrinsically nonlinear models, and/or other methods beyond the scope of our paper.

Emphasis of This Review and Further Reading

Although we aim for coverage of all major relevant methods of analysis, no review of longitudinal methods can hope to exhaustively cover every specialized, custom-built, ad hoc or improvised analysis method developed for this kind of research. Neither does this paper attempt to discuss all the methodological and design issues relevant to longitudinal studies, but focuses primarily on data analysis and those methodological issues closely tied to that.

Further, measurement issues not intimately connected to our analysis methods cannot be pursued here because of space limitations, though we feel they are important and too often overlooked by clinical researchers.

We do not emphasize research designs that involve a very large number of measurements (e.g., >50) across time on just one or a handful of entities – e.g., groups where values at each time point are averages across people, and where an intervention commences at a point part-way through time, and whose effects are of interest. Such data are suitable for Box-Jenkins-Type Time Series Analysis approaches. Similarly, Event History Methods, which analyze time *until some clinically meaningful event*, often death, e.g., 'survival analysis', are not stressed here. While time series and event history analyses are powerful methods, they are also relatively more developed and established techniques with a long history of abundant literature on their use and interpretation, and are also not usually applicable to the focus of examining predictive effects on a numeric variable assessed on relatively few observations across time for each of many subjects. We touch upon these and other methods for the sake of completeness, contrast, and clarification of which niche each method does and does not fill across a broad constellation of methods, and include them within a general flow chart of longitudinal methods.

We give priority to breadth of coverage over depth. Many important details, including mathematical derivations and formulas, can be pursued in references and elsewhere in the literature. We provide a systematic perspective on old and newly emerging techniques in the rapidly developing area of longitudinal research. While our primary purpose is to present a review of existing methods and not to introduce new techniques, we describe, mainly for illustrative purposes, some of our own variations and extrapolations in the application of these methods.

Another reason for writing this article is that we feel there is an, as yet, unmet need for a review in the clinical neuroscience literature that covers a broad overview of longitudinal analysis methods in a deliberate manner that is accessible to researchers without an advanced background in statistics or modeling. There are many examples of applications of longitudinal analysis, and methodological papers on longitudinal statistical techniques that are intended for statistically advanced audiences [5, 6] are more narrow in breadth in terms of the methods discussed [7–10] or are more focused on the specific concerns of a more restricted substantive area of neurological research [11]. An excellent article by Petkova and Teresi [12] provides a sophisticated discussion of random-effect models, but is more technical and less broad in coverage. Gibbons et al. [13] provide a more accessible treatment of the same techniques as Petkova and Teresi [12] within the context of psychiatric research. There are some excellent new or recently updated textbooks on longitudinal data analysis [14, 15], which we highly recommend for reference and further reading.

Pre-Data Analysis: Data Quality Assurance and Pre-Processing

The first step in analysis is data quality assurance (QA). Countless hours and days of 're-analysis' will be saved by ensuring your data are proofed, clean, complete (e.g., merging of data sets and creation of subject and visit level variables needed in the analysis), and in the format required for the software to be used before you start data analysis. While data QA and pre-processing are laborious and unexciting, they are essential first steps to ensure proper and efficient data modeling – if one cannot spare the time to redo everything, then one must give sufficient attention to QA up-front. Data cleaning includes thorough examination of missing data, searching for duplicate records, statistical and graphical screens, and setting up programming checks to alert you to improper data values due to input or transcription errors or outliers to be considered.

Importance of Iterative Graphical Data Analysis before, during and after Modeling Steps

Graphs should not just be limited to figures in manuscripts or slides in a presentation to illustrate a point. Graphical data analysis is a necessary component of good research methodology [16, 17]. Exploratory and confirmatory graphical analysis of raw and transformed data should be done preliminary to, concurrently with, and after numerical analysis. Preliminary graphs can serve in the QA process to screen raw data by highlighting obvious data errors that may otherwise be missed, like needles in a haystack of tabular data. Graphical displays go hand in hand with different steps in the analysis process. For example, graphs of residuals from a regression model plotted against predicted values are informative of model fit and whether assumptions of significance tests are met (e.g., normally distributed residuals with homogeneous variance across the predicted surface). Ideally, during analysis, an iterative cycle of graphical and numerical computations should be conducted. Post-analysis graphs can render numerical results that are difficult to interpret more understandable. For example, graphs of model-predicted values overlaid on raw data can make very intuitive what an 'adjustment' in ANCOVA means. Often, complex multiple regression models involving curvilinear terms, interactions, and combinations thereof are difficult to visualize even for mathematically sophisticated researchers, and frequently only a graph of predicted values can elucidate the nature of the model. Using the estimated model function to compute and plot predicted values at illustrative strata of covariate levels or scores held constant can be helpful.

Information richness is a key concept in good graphical analysis. Examples of information-rich graphs are scatterplots, side-by-side group dot plots, stem-leaf graphs, comparative frequency histograms, box-whisker plots, and 3D scatter- or surface plots. A simple two-dimensional scatterplot of raw data, for example, provides a wealth of information on: the univariate and bivariate distributions of the variables [Are they normal? Is there skewing? (important for consideration of assumptions of statistical tests or need for transformations, e.g., log or square root for positive skewing/powers for negative skewing)], whether there is any relation between the variables, and if so, whether the relation is linear or nonlinear, and what kind of nonlinear, the distribution of residuals from any relation, and whether they appear to meet assumptions of tests, a rough idea of the degree of correlation, the means/medians of the variables, modality, variability, and relative variability, whether there are ceilings or floors for the two variables, whether there are outliers which may be having a strong influence on statistics, out of range or nonsensical values indicating data errors, whether there are clusters that may have substantive meaning, or unexpected phenomena, for example. Fitted regression lines, polynomial curves, nonparametric smoothing curves (e.g., SAS Proc Loess), horizontal/vertical reference lines or a diagonal line where vertical and horizontal scores are equal can be overlaid on the scatterplot where relevant as visual aids. Incorporating group information into the scatterplot provides a quantum jump in information and can illustrate well ANCOVA or multivariate numerical results. Different groups can be indicated in a single graph with different symbols and/or colors for their respective points or a separate panel displayed for each group with uniform cross-panel horizontal and vertical axis ranges for easy group comparisons. An important concern in scatterplots (and dot plots) is whether multiple observations at the same spatial location are manifest or hidden. The latter can be dangerously misleading although many graphical software packages produce graphs with hidden observations without even warning the viewer that it is happening. Using different letters to indicate multiplicity of observations at a point is one way to avoid the problem ('a' = 1 observation, 'b' = 2, etc., as is done by default in the SAS plot procedure), or it might be advisable to add a slight random perturbation to values for purposes of the graph ('jittering' the data) so that multiple points at the same location are offset a little and at least partly distinguishable to convey a sense of the multiplicity in that region.

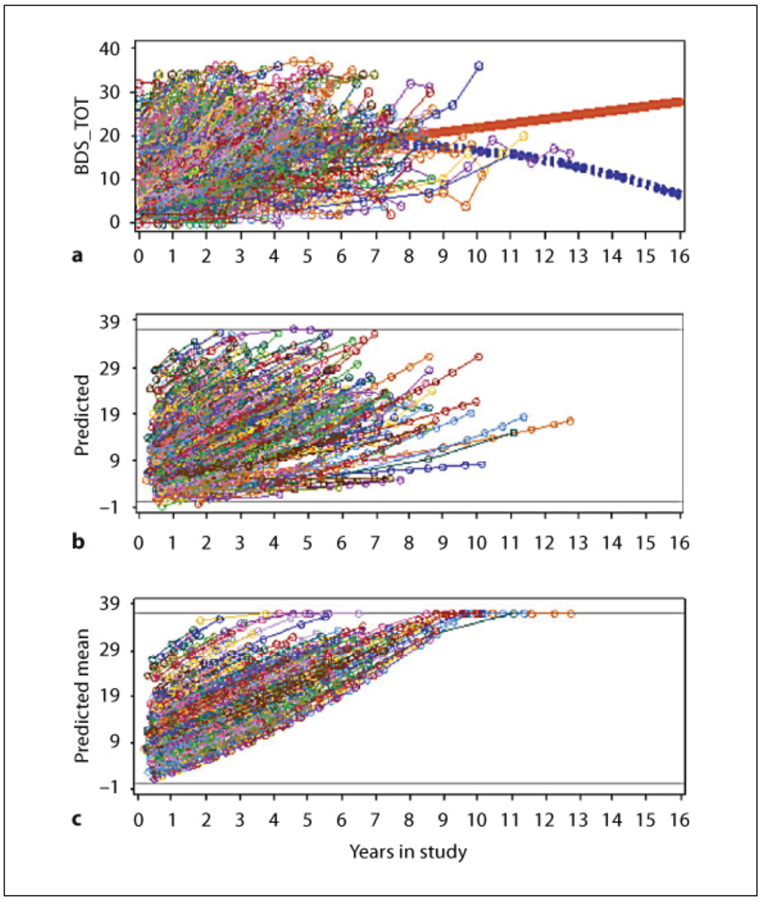

Fig. 1. a A 'spaghetti plot' of raw longitudinal data (example from Dodd et al. [28]). Raw BDS vs. years in study for 493 AD patients, each having 3–14 observations over time (years in study). The BDS score is the number of errors made on a measure of cognition (higher score means the patient is performing worse). Thin lines connect scores for an individual person. The thick straight solid line is the OLS regression line, and the thick dashed line is the OLS quadratic curve (this graph was produced with SAS Graph software, Proc Gplot). **b** The same data after removal of the pure time or visit level random error via a random-effect model, leaving subject level random quadratic and linear time terms and fixed effects. **c** The same data after additionally removing the subject level random quadratic and linear effects, leaving only fixed effects which included an interaction between baseline level of the BDS and a quadratic effect of time, shown in the figure as a predicted accelerating increase for subjects with low baseline levels but a decelerating increase for those with high baseline levels.

An example of an information-poor graph is a bar chart of group means. Even with error bars, they hide more than they reveal, though they may be helpful when there are many categories or variables. Box-whisker plots are usually an improvement.

With regard to longitudinal research, the value of graphical analysis becomes even more paramount. Our research group often examines 'spaghetti plots' of raw longitudinal data preliminary to data analysis, employing the Gplot procedure of SAS Graph software or the JMP interactive version of SAS (fig. 1a; Appendix). These graphs are essentially scatterplots of dependent variable scores versus the time variable with a separate line for each person connecting his/her scores over time. Spaghetti plots suggest likely models, especially whether effects are linear or not, whether there are ceiling or floor asymptotes, and in addition to

DOI: 10.1159/000330228
Published online: October 26, 2011

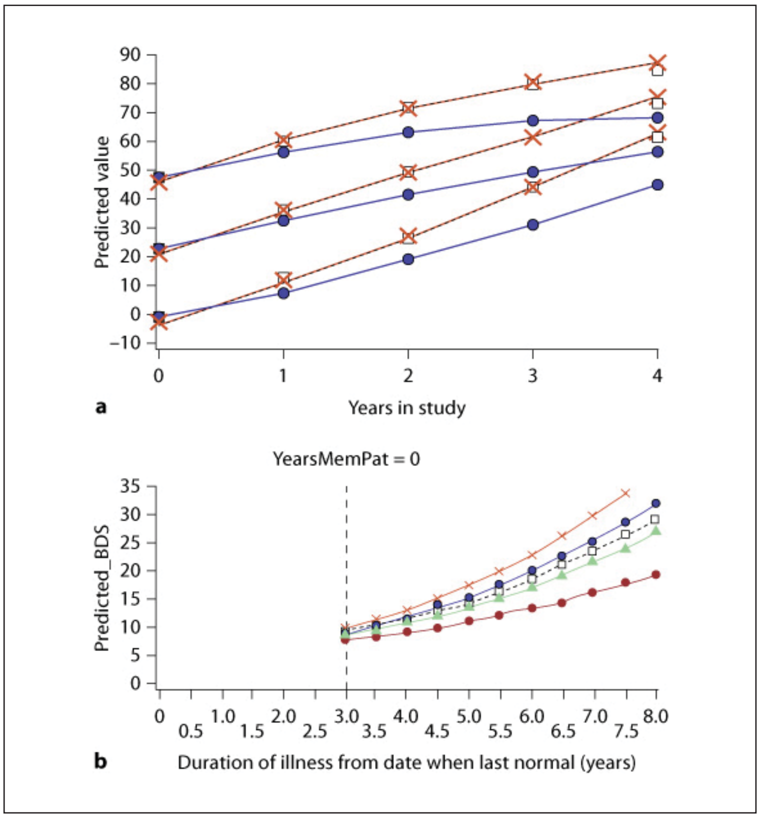

Fig. 2. a Illustrative mean ADL values vs. years in study, predicted by best-fitting longitudinal mixed-effect model for 382 AD patients treated with various medication regimens and starting at different initial mean ADL values (0, 25, 50). Score = Dependency (%) on other people; square = no medication; × = cholinesterase inhibitors only; dot = combination of cholinesterase inhibitors and memantine. Baseline ADL values and their linear/nonlinear interaction with time were included as fixed predictors. Note the differing trajectories depending on the baseline level, and superimposed on that is a medication group effect whereby the combination therapy apparently dampens clinical progression as measured by the ADL (from Atri et al. [29]). **b** Illustrative mean BDS scores across time predicted by the fitted mixed model in the longitudinal analysis for log plasma CRP for 122 AD patients, for selected levels of baseline log CRP and example time span. Illustrative levels of log CRP were chosen to correspond to the 1st, 25th, 50th (median), 75th and 99th percentiles of its distribution (from Locascio et al. [30]).

all the information provided by scatterplots noted above, they provide information on within-subject versus between-subject effects, and subjects who are outliers in terms of their pattern of progression, even if not in terms of the levels of the values themselves. As in the case of cross-sectional analyses, graphs of predicted means from a fitted longitudinal model are important and necessary when complex terms are significant which are difficult or impossible to understand or visualize otherwise (fig. 2). A picture is worth a thousand words as well as a thousand summary statistics.

Assessing and Modeling Linear versus Nonlinear Changes

A general consideration applicable to all the methods discussed below concerns the issue of whether change over time is linear or nonlinear. '*Rate of change* over the entire range of a dependent variable' assumes a linear model or one that is nearly so. Whether a change is nonlinear should always be raised as an initial question during the design of the study and/

or analysis. Rich, exploratory graphical analyses of raw data go hand in hand with addressing this concern. Often a nonlinear model is suggested by preliminary graphical analysis, or else it may simply be desirable to do tests to rule it out. Nonlinearity may be related to ceiling or floor effects. Sometimes a transformation of the data, e.g., logarithmic, is sensible and reparameterizes the model to a linear one that simplifies analysis, though complications may then occur during interpretation, unless results are transformed back to the original scale of the variable after analysis or the transformed data actually have more substantive meaning than the original. (Transformations can also be conducted for purely theoretical, substantive, and mechanistic reasons, and transformations are occasionally necessary to meet distributional assumptions of significance tests.)

Nonlinear models can be broadly classified as to whether they are *intrinsically* nonlinear or not. The latter or *'curvilinear'* type may be changed into a linear model by a simple transformation of one or more predictor variables in the model, followed by analysis with straightforward linear methods. The variables are then transformed back into the original variables during interpretation. For example, a curvilinear quadratic polynomial model which by definition has no more than a single bend can depict accelerating or decelerating change, but is not intrinsically nonlinear. A single predictor variable may be so modeled by simply squaring the predictor and entering the original variable (e.g., x) and its square (e.g., x^2) into a simultaneous multiple regression analysis just as if they were two different predictors. These predictor variables (e.g., x and x^2) will of course often be highly correlated, but the analysis will adjust for that. The chosen estimation method used to find parameter values that optimize model fit can be ordinary least squares [OLS; e.g., using the SAS regression procedure, Proc Reg, or the Proc GLM (general linear model procedure)] or can consist of maximum likelihood methods. In this example, the software is indifferent to the fact that one predictor is the square of the other and treats them like it would any other two (possibly correlated) predictors. Higher-order polynomials are handled similarly, e.g., a cubic (two-bend) model requires the addition of the cubed predictor (x^3) along with the squared and linear corresponding variables (x and x^2). An intrinsically nonlinear model is one that cannot be transformed, at least not in any straightforward manner, into a linear one, e.g., exponential (accelerating, decelerating, or asymptoting), or logistic or probit ('S'- or sigmoid-shaped curve) models.

Moving from common between-subject or 'cross-sectional' analyses to longitudinal methods, which include between- and within-subject and time effects (and their interactions), can involve a quantum jump in complexity. In the case of intrinsically nonlinear models, fairly specialized fitting methods and software (e.g., SAS Proc Nlin or Proc NLMixed) may be required, and in these cases getting iterative algorithms to converge on a solution can sometimes be difficult. Under specific situations and trade-offs, there may be an advantage to modeling nonlinearity with more simple polynomial transformations. Polynomial functions are more mathematically tractable, and can be fit using simpler procedures (e.g., SAS Proc Mixed). Thus, if the specific and detailed nature of the model function is not of paramount substantive interest (e.g. the model is not mechanistically explanatory) and having less complexity is more important than perfect fit, then the trade-off may favor a curvilinear over an intrinsically nonlinear approach even though the latter may be closer to a given true latent relation. For example, in such a scenario, data fitting using a quadratic function/model with adequate fit may be chosen over an exponential accelerating or decelerating model that fits the data only marginally better. Another example would be to utilize a cubic polynomial for an apparent sigmoid relation with floor and ceiling asymptotes, even when a logistic function may fit slightly better. Figure 3 displays simulated data generated to follow a logistic sigmoid curve with error variability incorporated and a floor and ceiling effect. However, as depicted, predicted values from a cubic function also fit the

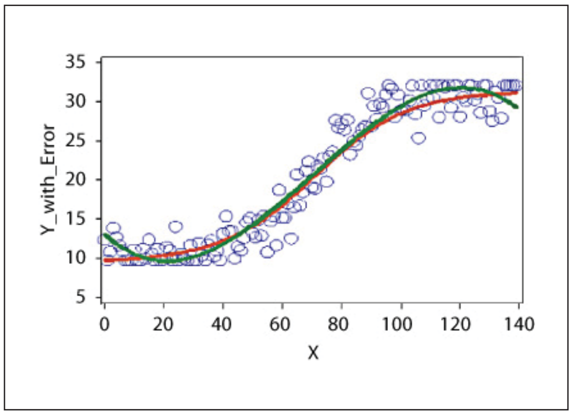

Fig. 3. Fitting sigmoid data with a cubic model. Data were created to approximate a sigmoid shape with a floor, ceiling, and some normally distributed random error (error std. dev. = 2). Best-fitting cubic and logistic curves are shown. Note the cubic curve bends slightly at tails in contrast to the logistic curve, a difference which may be trivial or unacceptable depending on the situation. The cubic function accounted for 95.2% of the variance in the dependent variable, and the logistic model accounted for 95.4%.

data well within the range of interest. A caveat to be heeded is that a polynomial model (e.g., quadratic or cubic function) may produce a slight non-realistic 'bend' where a monotonic asymptote ought to be, at the extremes of the range of the independent variable (due to the nature of the polynomial). These unrealistic predicted values should be recognized as a localized method artifact, and accounted for as such in interpretation and extrapolation of results.

If, for whatever reason, it is desired to fit an intrinsically nonlinear random-effect longitudinal model, and problems with convergence occur in using SAS Proc NLMixed, for example, it may be possible even for this intrinsic nonlinear situation to perform analyses in a series of less complex stages involving only linear modeling that produce essentially the same final result. We have had some success, e.g., in modeling sigmoid curves, by first transforming the dependent variable into a quasi-logit function based on the proportion each value is of its range of scores, which are then analyzed fairly easily with linear or polynomial random-effect models, using SAS Proc Mixed, for example. The final random and fixed predicted values from this latter model are then post-transformed with an appropriate exponential function back into sigmoid functions which fit the raw data well.

In all modeling, including modeling nonlinear relations, it is important to be explicitly mindful that even when a model fits well, the evidence is only supportive that the model is 'sufficient' to describe the results – it does not prove that model is 'necessary' – and it has been shown to fit only within a specified range of predictor values in the data. For example, a quadratic function may fit an asymptoting, decelerating relation well, but extrapolating beyond the limits of the observed data may incorporate a bend in the predicted curve that was not part of the fitted data model. Such a scenario may result in both poor as well as possibly nonsensical predictions based on the quadratic model for distant extrapolations. Most non-threshold-like relations can be estimated to be linear within a small enough data range (i.e., can be fit with a straight line for some small enough range of independent variables). Also, nonlinear models should be applied with caution when they are used to fit data that have a clear, sharp floor or ceiling, and not merely asymptotes. More sophisticated methods of dealing with floor and ceiling phenomena have been proposed [18, 19]. How to deal with these situations is best decided on a case-by-case basis depending on the questions being asked, study objectives, the nature and range of the observed data, and the overall goals of interpretation and prediction.

Allowing for nonlinear models in longitudinal studies may provide supportive evidence of or serve for exploration of theoretically driven or mechanistic considerations. For example, decelerating trajectories of symptom improvement in a medication treatment/intervention patient group, in contrast to the lack of the same in an untreated patient group, may further suggest an actual disease modification effect in the treatment group, whereas nonsignificant nonlinearity and only a significant change in intercept of a linear progression might be more suggestive of treatment effects on symptom reduction only. Tests of interactions of groups with differential linear and curvilinear terms could be sensitive to subtle and complex effects, missed by a linear analysis alone.

Analysis Methods

Overview

Figure 4 provides a general flow chart to assist the researcher in deciding what kind of analysis is appropriate to the specifics of his/her longitudinal study. Figure 4 is self-explanatory or will become so as one reads the remainder of the paper.

In the following, we discuss the different kinds of longitudinal analysis methods, not necessarily in the order they appear in the flow chart. Older, more traditional methods will be discussed first, followed by methods we consider most useful (random-coefficient, generalized estimating equation, GEE, and latent growth curve models, LGCM), and lastly we give briefer treatment to techniques that are more specialized or slightly out of the scope of research situations we are trying to cover. The methods below are not exhaustive of the full array of techniques for analysis of longitudinal data or methods closely related to them, but they are the most well-known and widely used ones.

Older, Traditional Methods

Simple Regression of the Dependent Variable on Time

Occasionally, researchers simply pool all the multiple records from multiple subjects and then just regress the dependent variable on the pertinent time measure using conventional OLS regression methods. Each observation point for each subject is simply treated as a separate record with the analysis being blind as to which scores are from the same or different persons. This method is not recommended, except perhaps to obtain a quick exploratory sense of anything striking. Although coefficient estimates may be unbiased with respect to parameters in the analogous referent population, conventionally computed standard error estimates can be very biased (up or down), and, as a result, so are tests of statistical significance based on them, because autocorrelation of scores within each subject is completely ignored and the assumption of the significance tests that observations are independent is grossly violated. In addition, effects of interest may be obscured with this method because relations within and between subjects, which could be very different, are just indiscriminately pooled. (Random-Coefficient Methods can avoid this problem as well as the others.)

Repeated-Measure Analysis of (Co)Variance

AN(C)OVA is used to analyze data with a continuous numeric dependent variable and one or more categorical/discrete predictor variables with the optional inclusion of some continuous numeric 'covariates' whose linear or nonlinear relations to the dependent variable are statistically separated from their otherwise confounded admixture with the other predictor variable effects. Classical repeated-measure ANCOVA, and variations thereof [20, 21], are suitable for longitudinal research designs that generally are well balanced, have the same, relatively few, and usually evenly spaced time points for each subject, with no missing values (various kinds of contrasts can address any uneven spacing within subjects). (The SAS GLM procedure with the repeated-statement option can perform these kinds of analyses.) For ex-

DOI: 10.1159/000330228
Published online: October 26, 2011

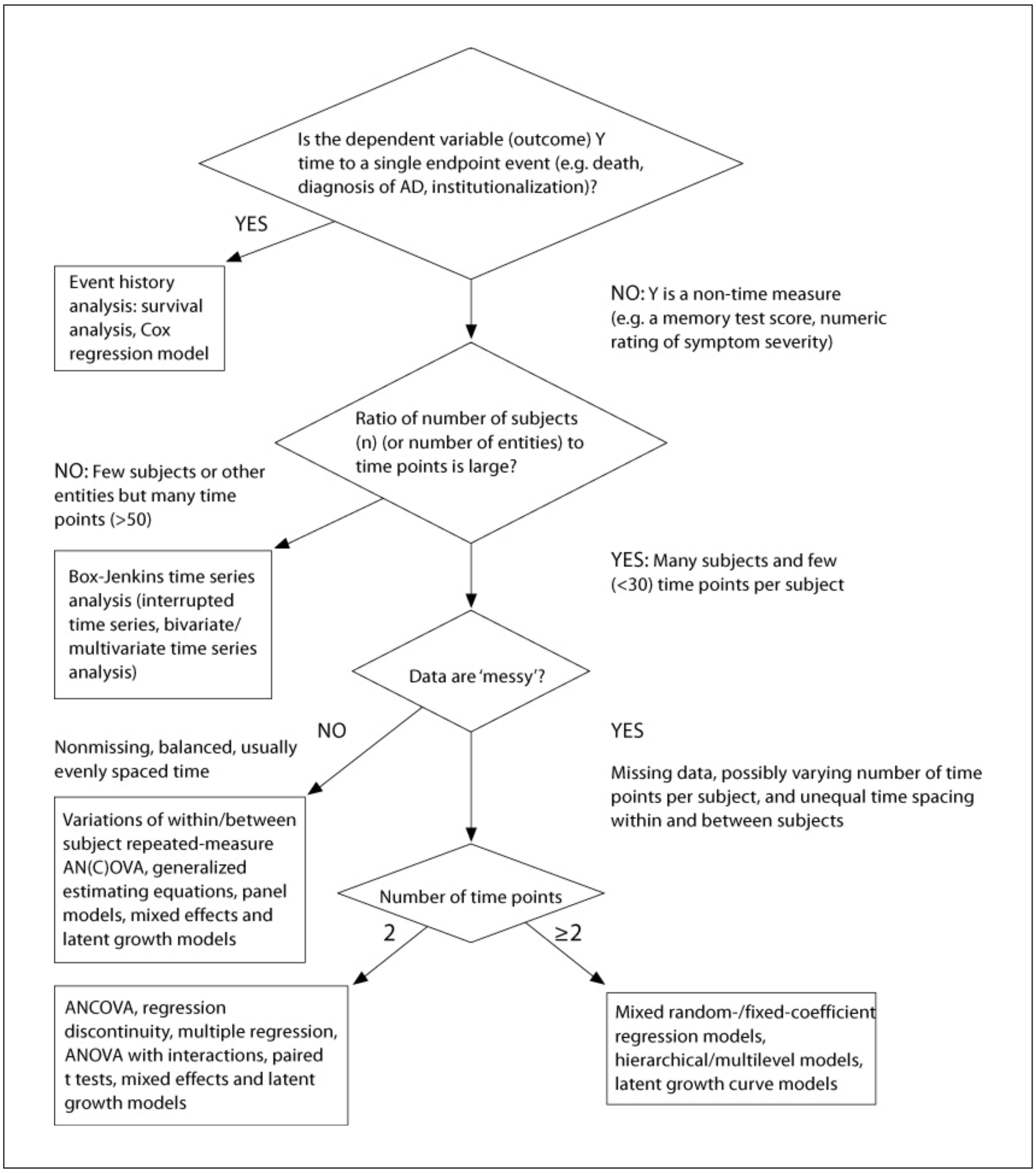
Is the dependent variable (outcome) Y time to a single endpoint event (e.g. death, diagnosis of AD, institutionalization)?
YES
Event history analysis: survival analysis, Cox regression model
NO: Y is a non-time measure (e.g. a memory test score, numeric rating of symptom severity)
Ratio of number of subjects (n) (or number of entities) to time points is large?
NO: Few subjects or other entities but many time points (>50)
Box-Jenkins time series analysis (interrupted time series, bivariate/ multivariate time series analysis)
YES: Many subjects and few (<30) time points per subject
Data are 'messy'?
NO
Nonmissing, balanced, usually evenly spaced time
Variations of within/between subject repeated-measure AN(C)OVA, generalized estimating equations, panel models, mixed effects and latent growth models
YES
Missing data, possibly varying number of time points per subject, and unequal time spacing within and between subjects
Number of time points
2
≥2
ANCOVA, regression discontinuity, multiple regression, ANOVA with interactions, paired t tests, mixed effects and latent growth models
Mixed random-/fixed-coefficient regression models, hierarchical/multilevel models, latent growth curve models

Fig. 4. Flow chart for deciding which method to use to analyze longitudinal data (with continuous numeric outcome) in neurological research. This flow chart should be considered only a rough guide; not all possible situations, exceptions, and combinations or variations of methods could be included.

ample, a 'one-way' repeated-measure ANOVA might involve testing each of 30 subjects on each of 6 consecutive days on the same test of memory. The same dependent variable on the same scale is assessed at each 'level' (days here) of the repeated-measure factor. Nonlinear change over time, e.g., polynomial contrasts, can be included as part of the analysis or assessed post hoc to it. There may be one repeated-measure factor or two or more crossed ones (factorial repeated-measure ANOVA), e.g., period of assessing some performance measure (a baseline and 4 follow-ups at 6-month intervals) may be crossed with a laterality (left vs. right) brain measure for each subject. There may also be various combinations of within-subject (repeated-measure) and between-subject crossed factors. For example, between-subject factors of medication group (those who receive a treatment drug versus others who receive only a placebo) and of gender might be crossed with each other and each also crossed with the within-subject factor of period of assessment. All combinations of multiway interactions among within-, among between-, and across within- and between-subject factors can be tested, in addition to all the main effects of the respective within- and between-subject factors. Subject level numeric covariates that are constant across time, e.g., years of education, might be incorporated into these designs (with interactions with other factors) or even covariates that vary across time, e.g., blood measures. Cross-over and parallel-group designs, as employed in many clinical trials, can be thought of as variations in repeated-measure ANOVA. An advantage of repeated-measure ANOVA over between-subject ('cross-sectional') ANOVA is that having each subject 'act as his/her own control' usually increases precision and permits more power for assessment of effects with less subjects.

A fundamental and common problem with repeated-measure ANOVA is that the repeated observations across a single subject can generate a correlation structure that violates an important assumption of ordinary between-subject ANOVA. This assumption is that the observations should not be correlated at all, in this case within subjects, or the correlations should be homogeneous across all pairs of levels of the within-subject factor(s). (If the within-subject factor has only two levels, as in a paired t test, or for other single degree of freedom effects, the assumption does not apply.) This assumption goes by various names, the most common being 'sphericity', 'compound symmetry' or non-correlated error. The issue is usually not a concern if the within-subject factor has levels that are not tied to time, there is little likelihood of any carryover effects from one level to another, the ordering of the levels for a subject does not matter, and/or the levels are randomized or counterbalanced across subjects, e.g., equivalent forms of a memory test known to show no practice effects are administered in 1-hour intervals in a high-facilitating, a low-facilitating and no-facilitation condition, for each subject in a random order with level of facilitation, not time, being the within-subject factor of interest. However, there is generally a problem in longitudinal research where the within-subject factor is by definition time or a variable tied to it, e.g., age or duration of illness. Usually, within-subject factor levels close in time have higher positive correlations than those more separated in time. Some software provides tests of whether the sphericity assumption is violated and if so, various methods are employed to eliminate the problem or adjust for it. Sometimes a model for the correlation is fit (e.g., autoregressive) and removed, but the more commonly employed methods are: (1) to re-parameterize the model so the within-subject levels are modeled as multiple variables in a multivariate (multiple dependent variable) analysis where any pattern of correlation would be permissible, or (2) an adjustment is made to the degrees of freedom (d.f.) of the within-subject factor(s) based on an estimate of the correlation structure of the within-subject factor(s), which lowers the d.f., essentially to allow for the fact that because there is time-to-time correlation, there are less actual d.f. than the nominal value indicates. We find the multivariate approach tends to be a little less powerful than the d.f. adjustment technique and is conceptually more confusing. Two variations in the d.f. adjustment method are the Greenhouse-Geisser [22] method and the slightly more liberal Huynh-Feldt [23] method, both of which in our experience almost always give the same result in terms of whether the effect is significant or not (see Girden [21] for when one is preferred over the other). SAS and SPSS software provide both these adjustment methods (as well as the multivariate technique).

The big disadvantage of repeated-measure ANOVA for longitudinal research is that it is usually restricted to only certain types of situations, e.g., few time points (usually <10), and well-balanced data – the same number of very similarly spaced time points across subjects. Most software algorithms do not allow missing values, i.e., if a subject is missing a value for even one of a number of time points, *all* that subject's data are removed from the analysis ('listwise deletion').

Analysis of a Single Number per Subject That Indexes Change

In this approach, the problem of analyzing longitudinal data is solved by summarizing the relevant aspect of longitudinal change for each subject with a single numeric value that can then be further analyzed with any of the traditional between-subject methods. For example, for each subject, the simple difference of a follow-up score minus a baseline score can be computed. Then a t test on these differences can be performed comparing a group of medication-treated patients with another group of patients receiving only placebo to see if these groups differ in their mean baseline to follow-up differences. Such a difference in two time points is generally a poor method of longitudinal analysis unless there are only two points of information available for each subject, and the time difference between them is fairly uniform across subjects or not thought to matter, or is adjusted for with a time interval covariate, for example. If there are more than two time points of data available for subjects, using only the two boundary points wastes hard-earned information and is blind to anything going on in between them, including any nonlinearity that might be of interest.

As single summary measures go, the slope of the OLS regression line of the dependent variable on the time variable fit separately for each subject is usually a better index of change than just subtracting let us say each person's first from last score (e.g., the Reg procedure in SAS will compute these subject-specific regression slopes with subjects as a 'by variable'). The intercept for each subject might also be of interest. Nonlinear change can be indexed by a summary measure if the coefficient of the quadratic term in a regression or the coefficient for a log-transformed variable is used as the summary measure of change, for example. There is still a major drawback with these regression methods in that the coefficient for a subject with only a few time points is given the same weight in the subsequent analyses and assumed just as reliable as the coefficient for a subject with many more time points. Adjusting for this somehow with some sort of weighting algorithm seems like more trouble than it is worth given that Random-Coefficient Models (discussed below) automatically avoid this problem.

Unusual circumstances notwithstanding, all of these summary measure methods seem sufficiently flawed to preclude recommending any of them, except possibly as a quick exploratory technique or to provide some reassuring confirmation of a more sophisticated, but less intuitively accessible approach.

GLM with Fixed-Subject Effects

This method is conceptually related to the repeated-measure ANOVA approach above though it is more flexible with regard to data imbalances, and unequal number of time points and spacings across subjects. A GLM (essentially a flexible ANCOVA) can be conducted with subjects as a fixed categorical variable (with d.f. = number of subjects – 1), using the SAS GLM procedure with the subject identifier in a Class or Absorb statement, for example. Level effects corresponding to subjects, and interactions of subjects with time measures if specified, are removed, and any correlation of scores due to their being from the same subject are taken into account on that basis. (A further estimation and adjustment for correlated error within subjects may be needed.) The problem with this method is that by treating subjects as a 'fixed', as opposed to a 'random', factor, strictly speaking, results are restricted to only those particular people in that study, and even though there may be many of them in the study, results are not considered formally reflective of a larger universe from which these people were (usually) randomly drawn. Since the point of most research is to generalize to a very large referent universe of primary interest, of which the sample is a convenient, hopefully representative, extract, there is a conceptual problem here. (There may also be computational and software issues because one is essentially analyzing a categorical variable, i.e., subjects, with a number of levels usually far in excess of common fixed-effect analyses.)

More Recent, Advanced Methods

Random-Coefficient Models

One of the variations of this broadly defined method is used for longitudinal data analysis. The method is often referred to as 'random-effect modeling' although for longitudinal analysis, the models are probably better labeled as 'mixed-fixed and random-coefficient regression models' because in longitudinal designs, fixed coefficients are almost always included in the model in addition to the random, and in fact the fixed are usually the coefficients of primary interest [14, 15, 24]. These models, in some of their variations, are also often termed 'multi-level' or 'hierarchical regression models' [25–27] when some data are naturally nested within other levels of the data. In the case of longitudinal designs, observations of a subject across time can be considered to be nested within the subject level of the data hierarchy, and subjects are then sometimes further nested within various groups, e.g., diagnostic groups or treatment/control groups, as a still higher level of the hierarchy. Applied to longitudinal analysis, this approach essentially deals with the mass of between- and within-subject data by specifying a model in which each subject is assumed to have his/her own unique functional relation between the dependent variable and time-related predictor(s). A straight regression line (if linearity is assumed) or curve that optimally fits the data for each given person is estimated. The fit is generally not perfect partly because there is assumed to be random, normally distributed error variation in the dependent variable at each time point for each person. The coefficients describing these lines or curves, e.g., intercept and slope (rate) in the case of straight lines or perhaps polynomial coefficients (e.g., quadratic or cubic) for curves, are assumed to vary randomly in the population of subjects according to a multivariate normal distribution (and may or may not be intercorrelated). This is in contrast to the fixed nature of the coefficients in the method of GLM with Fixed-Subject Effects (with its associated problems in inference) discussed above. The random-effect algorithm computes 'empirical Bayesian estimates' of each person's coefficients which are based both on that individual person's own data as well as the average corresponding coefficients for the entire sample of subjects or subgroup/covariate strata within which the subject is nested. If the subject has a relatively large number of observations across time relative to what other subjects have, a comparatively high weight is given to that person's own data in estimating his/her coefficients, whereas if the subject has relatively few observations, relatively greater weight is given to that subject's group/stratum average in computing the coefficient estimates for the subject. This method of estimation is superior to basing the coefficients solely on the subject's own data blind to how numerous or few the observations are for each respective subject (such a weakness was mentioned above as applying to the Analysis of a Single Number per Subject That Indexes Change). Usually corresponding to each random effect, there is a 'fixed' effect that is often of primary interest, e.g., in linear models, each subject has his/her own random regression coefficient that estimates rate of change for that person, but there is a single group (or subgroup) 'fixed' coefficient computed that indexes the average rate for the whole group (or subgroup/covariate strata). Figure 5 provides a graphic illustration of a random-effect model. The same kind of terms that could be introduced in an ANOVA situation can be included in the design of a random-effect model, e.g., between- and within-subject crossed factors, subject level (i.e., constant for a subject) fixed covariates (e.g., demographics or baseline clinical variables), or within-subject time-varying fixed or random covariates (e.g., time-varying physiological readings), interactions, polynomial terms, and other effects. One of a number of variations in maximum-likelihood methods using an it-

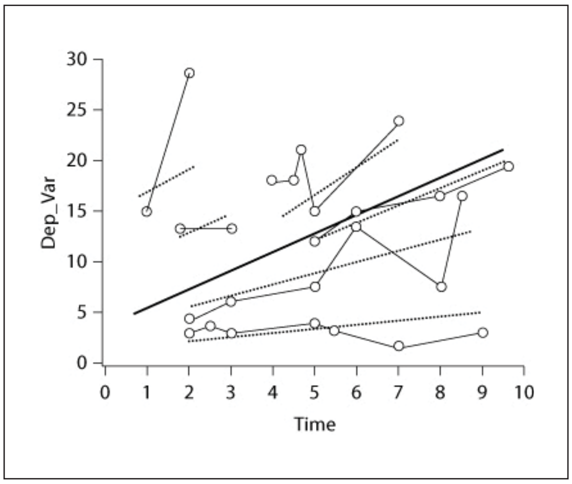

Fig. 5. Illustration of a random-effect model. Simple, simulated longitudinal data illustrate what a mixed-fixed and random-coefficient model does in the case of a simple linear model. Values on the dependent variable (Dep_Var) are indicated by circles with a thin solid line connecting scores for the same subject. The thick solid line in the middle is the estimated overall group regression line (a 'fixed' term). The dotted straight lines are the regression lines with random slopes and intercepts fit to the subjects' data, respectively. Note that when a subject has only a few observations, like the subject at the upper left, the slope and intercept of his regression line is weighted to be similar to the overall group average, whereas when a subject has relatively many observations, like the subject at the bottom of the graph, the regression line is more weighted in accordance to that subject's own values.

erative algorithm is usually employed to make the estimates. A widely used statistical routine for performing random-effect analyses is the SAS Mixed procedure (Proc Mixed) and there is an NLMixed procedure (Proc NLMixed) in SAS that handles random effects for intrinsically nonlinear models.

Besides the advantages noted above for this longitudinal analysis method, the method is very flexible in allowing for imbalanced data, missing values, differing number of time points from subject to subject, and unequal spacing of time point intervals within a given subject as well as across subjects. As in most other analysis methods, however, if missing values are not missing at random, biased results may be obtained (see the Missing Values section). It is also possible to include subjects in the analysis who have only one observation across time as this provides partial information. However, common sense suggests that when effects *within* subjects are of interest, most subjects should have two or more observations for reliable assessment of linear effects, three or more for quadratic relations, four or more for cubic, and so forth (in addition to any baseline score on the dependent variable that might be used as a subject level covariate). A disproportionately high number of these minimum frequencies is also undesirable, e.g., fitting a linear model to only 2 points per subject provides little advantage over a simple change score analysis or ANCOVA of 'posttest' covarying 'pretest'.

Note that this method deals with the correlation of scores within a person by basically separately modeling each person's data, fitting the overall elevation of a given subject's data (person average) with a random intercept, the slope or rate of change over time for that subject with a random slope, the degree of curvature of progression for that person with random polynomial coefficients or parameters of an intrinsically nonlinear model, and so on. Generally, the random terms included in the model are a random intercept, a random time-related variable, e.g., time in the study, duration of illness, age, and possibly corresponding random polynomial terms, e.g., a quadratic coefficient. Even more than one distinct variable can be made random in a given model, e.g., age as well as duration of illness, but too many random terms can cause failure of the estimation iterations to converge because of the complexity of the model, number of parameters to estimate, and possible difficulty in pulling apart random variables that are highly correlated especially within subjects.

Within the mixed-effect framework, methods can be used to separate out between- and within-subject relations, in case they are different – often it is the latter that are of greatest interest in a longitudinal study. Effects of time in the study, age and duration, for example, can be separated, if desired, by choosing one as the random term, e.g., time in the study, and then using the others as fixed-subject level baseline constants, e.g., age at baseline and/or duration at baseline. Care should be taken in deciding before the analysis how within- and between-subject effects are going to be disentangled and even greater care exercised when interpreting various relevant subject and time level coefficients after analysis. We find, for example, that separating a random variable indexing time in the study from fixed-subject level baseline age and baseline duration of illness predictors often works well. During interpretation of results, one must remember time in the study increments in tandem with age and duration of illness within a subject and consider whether something happens within a subject during the study additive to or in conflict with effects of aging and duration of illness as evidenced by between-subject relations. It is quite possible for between- and within-subject effects to differ or even be opposite in direction. For example, a biological variable may tend to decline with age, as suggested by its negative between-subject relation to baseline age, whereas a treatment applied during the study may cause increases within each subject in that same biological variable in opposition to the effect of increasing age within the same subject. Furthermore, linear/nonlinear baseline *dependent*-variable level adjustments can also be incorporated into a model as well as interactions of the same with the key time predictor, which may be a good way to model floor, ceiling and asymptoting trends tied to extreme baseline levels.

When the only random term in a model is the intercept, the correlations between dependent-variable scores at all pairs of time points are assumed to be the same ('compound symmetry'), but when the linear term for a time variable and possibly additional polynomial terms are introduced as random, heterogeneous correlations across time are modeled. SAS Proc Mixed can also optionally fit and allow estimation of correlations among the random coefficients themselves, thus permitting further complexity to be introduced into the model. Proc Mixed provides significance testing of these correlations (or identically of covariances) as well as tests as to whether the variances of the random terms are significantly different from zero as an aid in deciding which terms need to be considered random. (Strictly speaking, any sample variability rejects the zero variance null hypothesis, but the test can serve as a principal indicator of whether variance is substantial enough that an improvement in model fit will be worth the added complexity of introducing the pertinent random terms.) In our experience with complex models, our group usually finds it preferable to start with a fairly saturated fixed and random model and then use backward elimination of nonsignificant ($p > 0.05$) terms, one at a time, for both. However, depending on the situation, forward, stepwise, and other selection methods, with suitable cutoffs, may be advisable.

A brief description of some of our recent applications of longitudinal mixed-effect models may help to clarify what these models can do. The data displayed in figure 1 are from a study [28] in which we tried to determine, among other things, whether baseline levels of a clinical cognitive performance measure (the Blessed Dementia Scale Total; BDS) predicted differential trajectory of change in that same cognitive measure for 493 Alzheimer's disease (AD) patients, each having 3–14 observations over time. Figure 1a displays a 'spaghetti plot' of the raw longitudinal data for the BDS (a higher number indicates worse cognitive performance), where each thin line connects the scores for each respective subject over time (years in study). The thick solid line and dashed curve are the linear and quadratic OLS regression fits. It is apparent that the OLS straight line tends to underestimate the incline of *within*-subject trajectories and the quadratic is also misleading as an indicator of within-subject trajectories because both are blind to within- and between-subject distinctions, whereas a random coefficient analysis takes these distinctions into account. Figure 1b shows the same data after the time point level error variance in the BDS is statistically removed, leaving a fixed quadratic effect of time, dependent on the baseline level of BDS, i.e., an interaction of baseline BDS and the quadratic component of time. We tried to use this interaction to more realistically model the tendency for people starting at high baseline BDS to decelerate to a ceiling asymptote, whereas those starting at relatively low baseline values tended to accelerate (curve upward), perhaps partly due simply to their greater opportunity to increase in score. In addition, a subject level random component of this quadratic trajectory (a quadratic coefficient that varied randomly from subject to subject) is also meshed into figure 1b. Figure 1c removes this latter subject random effect leaving only the fixed baseline by quadratic time interaction which was of more substantive interest.

As another example, figure 2a shows illustrative mean progression curves predicted by a best-fitting longitudinal mixed-coefficient model for data for a measure of daily functioning (Activities of Daily Living Scale, ADL; higher scores indicate worse functioning) for 382 AD patients, each with 3–13 observations over time (years in study) and each in one of three mutually exclusive medication regimen treatment groups, indicated by the ×s, the dots, and the squares in the graph, respectively [29]. Our model here again included a fixed effect for the interaction of baseline level of ADL with the quadratic component of time as well as interactions of medication group with the quadratic component of time. Subject level random intercept, linear, and quadratic effects of time were also modeled. In this case, we felt the clearest visual explication of the important, estimated fixed-effect portion of the model would be obtained by taking the mathematical model which we estimated based on the actual data, but plugging into it a few simulated representative predictor values selected to produce predicted values of ADL which would be most illustrative of the nature of the model within the range of the actual predictor data and graphing those. (By contrast, the predicted values in figure 1 are the dependent-variable scores predicted on the basis of the actual, not simulated, predictor data, though in both cases the model used to make the predictions was estimated from real data.) One can see in figure 2a an interaction of baseline levels of ADL with the quadratic curvature of the trajectories over time, as well as a significant medication group effect superimposed on that, which was of special interest in the study (and was statistically significant).

The last example is from a study [30] in which, among other things, we estimated the predictive relation of baseline levels of a continuous numerically measured biomarker (log of plasma C-reactive protein; CRP) to trajectory of change in BDS over time for 122 AD patients, each with at least 2 to up to 25 observations across as much as a decade. Fixed and random quadratic effects of time were again estimated here. In figure 2b, predicted values across time (here shown in terms of duration of AD illness) were produced in a manner analogous to that of figure 2a. Illustrative levels of the log CRP predictor were chosen for the graph to correspond to the 1st, 25th, 50th (median), 75th and 99th percentiles of its distribution. (Thus, the spacings between the lines reflect the shape of the log CRP distribution, positive skewing in this case – lower lines correspond to higher values of log.) A significant fixed interaction effect of baseline CRP to the quadratic component of change in BDS over time is evident in the graph.

The Appendix gives an illustrative example of a SAS program which runs a random-effect model.

Generalized Estimating Equations

In a sense, GEE models [31] approach the problem of longitudinal data analysis from the 'top-down' in contrast to random-coefficient models that might be viewed as a 'bottom-up' method. Random-effect analysis basically focuses on individual subjects, modeling what is happening to them, and in the process is then able to assess average effects of interest, all of this done in essentially one big equation. GEE, on the other hand, is considered a 'marginal' longitudinal method. It directly tries to get an overview of the mean relations of interest on the one hand, i.e., how the mean dependent variable changes over time, while separately dealing with the nuisance covariances among the observations within subjects in order to remove the latter to get a better estimate and valid significance tests of the former. GEE estimates two different equations, one for the mean relations and one for the covariance structure. The SAS Genmod procedure with the repeated-statement option can be used for GEE analyses. The user can choose a variety of within-subject correlation models (e.g., autoregressive) to test, estimate, and remove, and most of these basically specify that positive correlations between temporally adjacent observations within subjects taper off as the observations get farther and farther away from each other in time. SAS users can now download an SAS 'macro' module that assesses which covariance structure seems to provide the best fit to the data (the quasi-likelihood independence criterion or QIC macro [32]).

An advantage of GEE over random-effect models is that it does not require the dependent variable to be normally distributed. However, a disadvantage of GEE is that it is less flexible and versatile – commonly employed algorithms for it require a small-to-moderate number of time points evenly (or approximately so) spaced, and similarly spaced across subjects. Nevertheless, it is a little more flexible than repeated-measure ANOVA because it permits some missing values and has an easy way to test for and model away the specific form of autocorrelation within subjects. Evenly spaced intervals are required because the user must specify a covariance (correlation) structure for the time point by time point covariance matrix which presupposes the same time points across subjects, and evenly spaced. Our tests seem to indicate that if the time points at least clump so they are mostly evenly spaced within and across subjects, the GEE is not severely effected, but it is unclear how to handle severe and numerous unevenness. In employing GEE to analyze data from studies with varying number of observations per subject, the estimated working covariance matrix has dimensions equal to the maximum number of observations any subject has.

We compared performance of GEE models for the study by Atri et al. [29] corresponding to figure 2a and the study by Locascio et al. [30] corresponding to figure 2b to that of the random-effect models we originally used for both and found very similar results for the Atri study but only moderately similar findings for the Locascio study. The divergence for the latter case may have occurred because time points were not quite as homogeneously separated as for the former, as assumed by the GEE procedure. For Atri et al. [29] about 2/3 of test-retest intervals were within 1 month of the median of 6 months, whereas for the Locascio study, only about 60% were. What may have been even more important is the fact that the focus of the latter study was to examine longitudinal effects of a variable that varied very incrementally across subjects (numeric biomarker levels) as opposed to testing large group effects in the Atri study, and as mentioned, random-effect models are meant to focus more on what is happening at the individual subject level.

Latent Growth Curve Models

LGCM [19, 33, 34] can be thought of as re-parameterizing random-effect models to specify latent variables that affect measures at time points in a kind of structural equation model (SEM) [35]. More commonly employed SEM show variables having predictive (or some say 'causal') effects on other variables (denoted by arrows) with coefficients indexing the strength

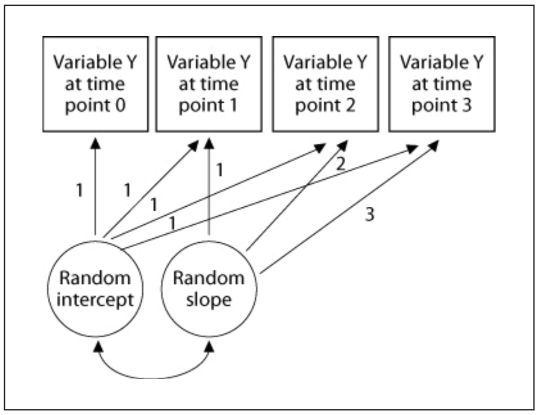
Variable Y at time point 0
Variable Y at time point 1
Variable Y at time point 2
Variable Y at time point 3
1
1
1
1
1
2
3
Random intercept
Random slope

Fig. 6. A simple LGCM illustrated as SEM. Circles denote latent random variables, squares are observed measures, straight arrows are predictive effects, and the double-headed curved arrow denotes a possible correlation between the random intercept and random slope latent variables. Numbers are coefficients applied to the predictors (intercept and slope). (Measurement error terms pointing at each observed measure are not shown for simplicity.)

and direction of the predictive relation. However, LGCM rearranges terms in a random-effect model so that what would have been coefficients in the classical SEM, e.g., intercepts and linear slopes, become the random variables in the LGCM and their impacts on measures across time are often prefixed. Figure 6 depicts a diagram of a simple linear LGCM.

Of the longitudinal methods discussed here, LGCM is the most recently developed and is still being further enhanced. Many procedures in the Mplus software package [2] are algorithms specifically for LGCM, and SAS is now introducing procedures explicitly for LGCM, e.g., the TCalis procedure, though some of its older methods [e.g., the Mixed and Calis (covariance analysis of linear structures) procedures] can be used to perform certain variations of it. (There is also a related SAS user-developed procedure, not supported by SAS, called Proc Traj.) One of the useful features of LGCM is that 'finite mixture models' [36] can be incorporated into them which, in the case of longitudinal analysis, basically look for underlying latent classes of subjects who have similar trajectories of change over time within each given class, but where trajectories differ across classes. A very useful application of this method might be in analyzing effects of medication (or other treatment intervention) on longitudinal change. If, among all medicated people, distinct clusters can be found which have significantly different mean trajectories of change, it might then be possible to explore characteristics on which the clusters differ and then use that information to decide which patients might benefit most from the drug. A profile of best responders might be developed across a number of demographic, clinical, and other variables. A different profile might apply to different drugs allowing the clinician to maximize the most beneficial drug for each patient. There is great flexibility in what LGCM can be used for. Perhaps the only drawback to this method is that it is relatively new with less available software, and it can be difficult to understand conceptually. Our group uses primarily random-effect models for longitudinal research though we are starting to experiment with LGCM. In any case, for very imbalanced 'messy' data, the LGCM provides similar or identical results to that of the random-effect analysis.

ANCOVA for the Special Case of Only Two Time Points per Subject

When there are only two time points of assessment for each subject, such as in a simple pre-/posttreatment or baseline to single follow-up design, ANCOVA, and variations of it, may be suitable to answer important research questions. For example, if the researcher wants to know if two medication groups with different mean symptom levels at baseline change differentially in their symptom levels from baseline to follow-up, ANCOVA indicates if they differ on the follow-up (the dependent variable), holding the baseline (covariate) constant, statistically, i.e., it is assessing differential group *change*. In some cases, a simple difference score ('change score') analyzed as a dependent variable without using the baseline as a covariate might be appropriate, but generally ANCOVA is superior in that the former can be thought of as a specific instance of ANCOVA where the linear regression of follow-up on baseline is assumed to have a slope of one, which may not be true. In fact, the baseline versus follow-up relation may even be nonlinear, which can be accommodated in an ANCOVA model. (But see Locascio and Cordray [37] for a situation in which ANCOVA and difference score analysis disagree and it is the standard ANCOVA solution that is wrong.) This kind of analysis assumes the time difference between baseline and follow-up is the same (or nearly so) for each subject or that it is irrelevant whether it is or not. If that is not the case, the time difference can be accounted for by introducing it as another covariate and looking for interactions in which the relation of follow-up and baseline differs depending on the time difference (perhaps the baseline to follow-up relation and correlation becomes increasingly attenuated as the time difference becomes greater). Classical ANCOVA may be too constraining in a given situation, and more flexible GLM or multiple regression variations of it might be employed instead. These kinds of analyses can test for group differences in the relation of baseline to follow-up (i.e., an interaction of group and baseline; conventional ANCOVA assumes this relation is homogeneous across groups), and/or how some predictors, like levels of a biomarker, predict change, and demographic variables (e.g., age, duration of illness, and years of education) can be covaried or entered as classification variables (e.g., gender) crossed and possibly interacting with a group factor, and so on. Scatterplots of follow-up versus baseline data within groups are very helpful, if not necessary, as part of the analysis. Predicted values from the model can be overlaid on the scatterplots to see more clearly the nature of the predicted model and get a visual sense of fit to actual scores.

Sometimes ANCOVA can be employed in the context of a 'regression discontinuity design'. This design can provide compelling evidence for a treatment intervention effect even in a nonrandomized study when ethical constraints require that the treatment, e.g., a medication rather than placebo, be given to those who seem to be in most immediate need of it at baseline, thus producing a confounding effect. This may be the case if it is felt that the treatment has a good chance of providing benefit, but the purpose of the study is to confirm this or assess the magnitude of the benefit, or in the case of a secondary analysis of a retrospective, observational study. See figure 7 for an illustration of a regression discontinuity design.

Other Related Methods and Designs

Space limitations prevent us from treating the methods below in greater detail. There is an abundance of literature on them in textbooks and journal articles.

Event History Methods

When a study is 'longitudinal' in the sense that the dependent variable is time to some event of interest, e.g., time to death, to recurrence of a disease, to institutionalization for dementia, or until diagnosis of dementia for patients who originally had only 'minimal cognitive impairment', event history methods are often employed [38, 39]. For example, a researcher might study whether a new medication treatment for memory impairment given to a group of patients admitted to a memory disorder clinic is associated with a longer time from entry into the clinic up until a formal diagnosis of dementia, as compared to what happens to patients given a standard medication. Such data cannot usually be handled with more conventional methods of analysis like ANOVA, primarily because of a problem termed 'censored data.' This problem occurs when, as is usually the case, subjects are not all entered into the study at the same time and not all of them are followed all the way until the event of interest takes place – some subjects drop out before that for various reasons, or the study must end before they have

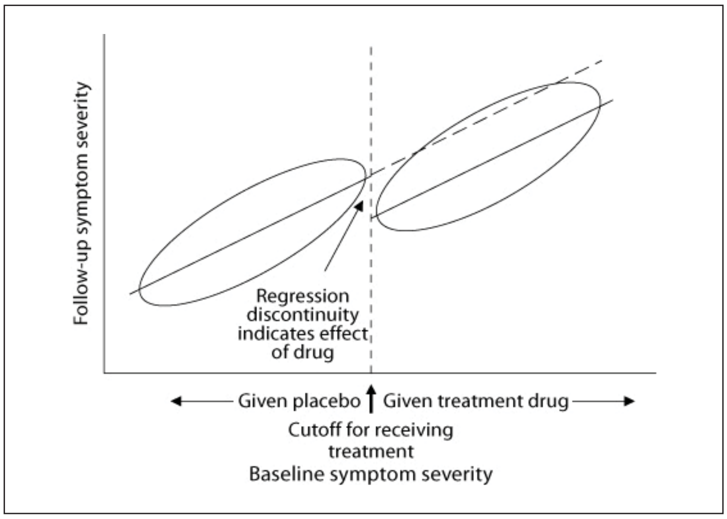
Follow-up symptom severity
Regression discontinuity indicates effect of drug
Given placebo
Given treatment drug
Cutoff for receiving treatment
Baseline symptom severity

Fig. 7. Regression discontinuity design (ANCOVA) – illustrative example. Ellipses denote swarms of data points for two respective groups (e.g., medication-treated and placebo groups) in a scatterplot of follow-up symptom severity versus baseline symptom severity scores (higher numbers = more severe). Solid diagonal lines are regression lines for the groups. Here the slope of the regression lines, and orientation and shape of the ellipses indicate an expected strongly positive correlation of follow-up symptom severity to baseline symptom severity within each group. For ethical reasons, the medication treatment is given to anyone above a cutoff on symptom severity at baseline (this may especially be the case if the treatment is in limited supply or very expensive). (Perhaps the treatment, if shown effective, might be given to the placebo group at a later time.) Note the treated group has worse mean symptoms than the placebo group at baseline as well as at follow-up. However, the discontinuous drop in the regression line in moving from the placebo to the treated group strongly suggests a beneficial effect of the treatment which, if strong enough, would be reflected in a significant group effect on follow-up symptom severity in ANCOVA with the baseline symptom severity as the linear covariate.

reached the event perhaps because they were one of the last ones entered into the study. The actual time to the event for these subjects is unknown and called ‘censored,’ although partial information is available in that it is at least known that they lasted up until the last assessment for them for whatever length of time that was into the study for them. (There is a distinction between ‘right’ and ‘left censored’ data; we discuss only the more common ‘right censored’ here.) It is undesirable to waste data, reduce power to detect effects, and possibly bias results by simply removing censored data from the analysis. Event history analysis tries to incorporate the partial information of the censored data together with the full information of subjects who reached the event of interest, to make optimal estimates of relations and effects of interest.

The oldest, classical form of event history analysis is ‘survival analysis’ in which the event being studied is death, as the name implies, but this label is also generally used for analyses that look at whether two or more groups differ in their mean time to events other than death [40]. ‘Kaplan-Meier product-limit’ survival curves are commonly used to graph predicted survival patterns; they display declining step functions of the estimated percent of subjects surviving across time for a given group (incorporating information from both complete and censored data). In 1972, David Cox [41] introduced a semi-parametric method of analysis usually called the Cox proportional hazard model or simply Cox regression model, which analyzes event history data in a much more flexible and general manner than traditional survival analysis does. Predictors in the model are related to a hazard function, which is the instantaneous risk of the event occurring among those who have still not experienced it as yet. Comparing classic survival analysis to a Cox regression is like comparing a t test or ANOVA to multiple regression or GLM, the former being specific instances of the latter, applicable only to restricted situations. The Cox model easily incorporates subject level and time-varying covariates, as well as group terms and the full array of interaction and other complicated predictor terms. Allison [38], Cox and Oakes [41], and Lee [40] cover the Cox model well. The SAS Phreg (proportional hazard regression) procedure performs Cox model regressions, and the SAS Lifetest procedure does more traditional survival analysis and plots Kaplan-Meier graphs.

Box-Jenkins Time Series Analysis

These methods [42–45] are appropriate for situations in which there are very many time points of observations (at least 50 are recommended as a rule of thumb) on just a few or even one subject or entity (e.g., a group). An applicable example might be the number of newly diagnosed cases of AD in a large geographical region, recorded each week for 2 years (see Locascio et al. [46] for an introduction to time series analysis and an application in functional MRI). Not surprisingly, these time series methods originated in econometrics, decades ago, where the applications were usually that of trying to forecast stock market prices. In a variation commonly employed in behavioral and medical research, the 'interrupted time series design', some sort of intervention has typically occurred about midway in the time series and the researcher is interested in whether this intervention has had any effect on the recorded dependent variable (usually on the mean level) after the point of intervention. To continue with the previous example of the AD study, suppose a region-wide program to improve dieting and exercise of vulnerable people is initiated at the beginning of the 2nd year of the study and the research question is whether this has had a beneficial effect on reducing the frequency of diagnosed cases of AD from the point of initiation of the program onward. It might appear that a simple t test or nonparametric test comparing the 52 weekly counts of AD before versus the 52 after the intervention would answer the question; however, a fundamental assumption of these significance tests (and most statistical significance tests) is violated in that the observations are almost certainly not independent of each other (within the pre-/postintervention epochs). A random error deviation at one time point may very well carry over to some extent onto the next few time points, i.e., the observations are serially correlated, usually positively so, with decreasing correlation the further apart two given observations are temporally. Although this lack of independence may not bias the estimated effect of interest, it will bias error variance estimates and consequently the results of a test of statistical significance of the effect (the p value), too, usually making it appear more significant than is truly the case.

Another related method is a 'bivariate time series analysis' in which the relation between two time series is examined to test, for example, if one of them might be causing or at least predicting the other one and if so, to estimate the approximate time lag across which this is occurring. Basically, the values of the two time series are paired by concurrent time points and a correlation is computed, then recomputed after lagging one of the time series forward and backward more and more with reference to the other series until hopefully an optimal correlation is reached that suggests the direction of prediction and at what lag. However, the significance levels for such an analysis are biased for the same reason as for the interrupted time series, i.e., the observations are not independent within each of the time series.

The solution employed by time series analysis to deal with the problem of serial correlation is to try to model, estimate, and then remove the nuisance autocorrelation component, so that the effects of primary interest can be validly tested for statistical significance. Two common models are used for this purpose – the autoregressive model, in which each value of a time series variable is assumed to be a linear function of one or more values of the same variable lagged back in time, and the moving average model, in which a weighted average of past error terms across a specific shifting window of time is postulated as contributing to the value of each time series variable score. Evidence as to which of these models fits best and its specific nature (how far back in time past observations impact the current one and via what set of weights) is obtained by examining the pattern of time point to time point correlations and (partial correlations) of the given time series as it is shifted numerous lags forward and backward relative to itself (the pattern of correlations is displayed in a 'correlogram' graph). Once the autocorrelation components are assessed, hopefully identified, estimated, and subtracted out, and a test confirms that the remaining residuals are not significantly different from independent, uncorrelated 'white noise', essentially a traditional significance test procedure can be applied (e.g., t tests, ANOVA, regression, or correlation) to the residuals to answer the research question of primary interest.

Besides often being termed dynamic regression models or Box-Jenkins Time Series Analysis according to its early developers [42], these types of time series analyses are sometimes referred to as time series analyses 'in the time domain' to distinguish them from time series analyses 'in the frequency domain', also called 'spectral analysis', which involves methods that try to decompose a time series into component frequencies of oscillation that sum to the fluctuations that are observed [47]. Spectral analysis is usually less applicable to the type of research situations in neurology which we have been discussing here, but may be useful, e.g., in some areas of analysis of image data taken over time.

SAS provides two procedures for time series analyses in the time domain, such as we have described – the Arima procedure and the Autoreg procedure, both of which are in a separate battery of SAS algorithms called Econometric Time Series programs [48]. (The Spectra procedure in the same battery may be used for spectral analysis.)

Panel Data

By panel data, we loosely mean situations in which two, occasionally more, variables are measured at approximately the same time at two or more time points, but usually no more than three or four, for all subjects, without missing values, and the researcher is interested in causality between the variables or at least reliable antecedent/subsequent predictive relations simultaneously and at various lags. The data are structured as a kind of series of cross-sectional sets [49]. There is a great variety of methods to analyze such data (e.g., the cross-lagged correlation technique [50]) usually involving some kind of multiple regression and/or partial correlation approach. They can be handled generally with 'path analysis' or SEM [35, 51–55], e.g., using the SAS Calis procedure. Panel design data and the techniques used to analyze them have a developed, sound methodology, but are limited in terms of situations to which they apply and questions they answer.

Miscellany

Missing Values

Some of the methods above, as noted, allow for small-to-moderate numbers of missing values, i.e., it is not necessary to remove all of a subject's data from the analysis because of a missing value at one or a few time points for that person ('listwise deletion'). For example, random-effect models allow the use of whatever dependent variable data are available for each given subject without the need for balanced data, equal intervals within and across subjects, or equal numbers of observations per subject. However, all of the methods above assume, as do most other analyses, that missingness of a dependent variable value is *not* related to what that value would have been if present. If it is, serious bias can result if nonmissing data are then analyzed and interpreted as if they are fully representative of referent populations. It is usually permissible for the missingness in the dependent variable to be related to values of *predictor* variables. This distinction is often made by employing the terms MCAR (missing completely at random), for situations in which the missingness is not related to any variable, as contrasted to MAR (missing at random), where it may be related to independent/predictor variables though not to the dependent variable. If missing values satisfy neither of these conditions, they are termed 'nonignorable' and although there are some newly developed methods for dealing with nonignorable missingness, these techniques have strong assumptions and must be used with caution [56].

If missing data are MCAR or MAR, new methods of maximum likelihood estimation and especially 'multiple imputation' (MI) can be usefully employed to provide very reasonable estimates of the missing values, thereby allowing the analysis to have greater power [56]. The MI methods cleverly use all available data and their observed interrelations to impute what all the missing values are likely to be, while at the same time introducing some random variation to mimic the uncertain element of the data and avoid attenuating error variance estimates. Multiple data sets are produced (only 5–10 seem to be necessary in most cases), each with possibly slightly different imputed values. After separate analyses of these data sets, separate estimates of relevant parameters are combined in an appropriate algorithm producing single values with associated valid significance tests. For longitudinal data with reasonably similar intervals for each subject, missing values can be imputed by structuring the data set so that different time points appear as columns like different variables would; then the MI algorithm uses information on the intercorrelations among time points to estimate the missing values at various time points.

We should also mention that when data are known to be missing for reasons that suggest meaningful approximate substitute values, for example, the condition of the patient was known to be too severe for him/her to be able to take the test measuring a variable (e.g., too demented to follow instructions for a cognitive test), that fact might be incorporated into the imputation algorithm in some way to obtain more reasonable estimates (see Locascio et al. [30], for an example of this). Care should be taken though that such estimated values are not constants carried forward repeatedly in longitudinal data to create an artificial plateau that may give a misleading impression of lack of progression, or at least that such results be properly interpreted for what they are.

The SAS MI procedure performs MI as the name implies. The multiple data sets produced are then analyzed separately by the relevant analysis procedure as 'by groups', and finally the MIAnalyze procedure combines the resulting separate output estimates for the relevant parameter in an appropriate way and provides a single estimate and a significance test for it.

Lastly, we have been experimenting recently with using empirical Bayesian estimated random trajectories (see Random-Coefficient Models) for each subject in a longitudinal analysis based on available nonmissing data, to obtain reasonable estimates for interim and extrapolated missing values for a given subject. Such a method might be useful for dealing with problems of study attrition also. Results seem promising but are as yet unclear.

Power Analysis

Power analysis is well developed and software available for more traditional analyses like ANCOVA and multiple regression, however, it is harder to find or is nonexistent for specialized longitudinal analysis methods. There are some packages that provide power for repeated-measure ANOVA [57]. SAS has Power and GLMPower procedures for an array of cross-sectional kinds of analyses, but not as yet for repeated-measure ANOVA. The power methods for cross-sectional data might provide rough estimates for the longitudinal case if conservative parameters are used. Cohen and Cohen [58] discussed power computation for within-subject designs.

Fortunately, powerful and fast modern computers permit a reasonable method of estimating power for virtually any complex analysis method, including random-effect models, using 'Monte Carlo methods'. One simply writes a computer algorithm that creates random simulation data with an embedded target effect size of interest, random variation of the kind and degree expected, perhaps based on past research, and a large number of randomly varying replications (at least 100, but better 1,000 or more) of the data set, each of which is analyzed with the longitudinal method at issue. The proportion of computed p values less than or equal to α is the estimated power level. The Mplus software has programs that perform Monte Carlo-type power analysis for different kinds of analyses, but it is not difficult to manually program the necessary algorithms, e.g., using SAS, to produce the simulated data and carry out the Monte Carlo tests with the SAS Mixed or NLMixed procedures.

Software

We have emphasized SAS software through most of the above because of its great breadth and depth of techniques, wide recognition of its reliability, and frankly because we are most familiar with it. The SAS Mixed procedure for random-effect models, the Genmod procedure for GEE models, the GLM procedure for ANCOVA and repeated-measure ANOVA, and the SAS Gplot procedure in SAS Graph software for graphing longitudinal data have been powerful work horses for us. We have also begun to experiment with LGCM via the SAS TCalis and Calis procedures. We have suggested many other SAS procedures above in the contexts where they were relevant. Online documentation for SAS can be found at http://support.sas.com/onlinedoc/913/docMainpage.jsp. Descriptions of SAS procedures always include at least a brief theoretical introduction to the statistical methods involved in the procedure. Information and downloading of the non-SAS supported Proc Traj for LGCM can be found at http://www.andrew.cmu.edu/user/bjones/.

Mplus is very good software for LGCM and many other methods. SPSS is an excellent general statistical battery which does mixed-effect models, and LISREL and EQS are especially intended for SEMs. Random-effect models are being increasingly incorporated into other statistical software products.

Summary and Conclusions

The purpose of this article was to present clinical researchers in neurology with an overview and practical guide to data analysis methods for longitudinal research. Older, traditional methods were covered and methods that are closely related to what are conventionally considered longitudinal methods, but emphasis was on more recently developed, advanced methods for analyzing data on a numeric, continuous, interval scale variable collected on a moderate-to-large number of subjects (10 to hundreds) with a small-to-moderate number of repeated assessments on each (2–10, 20 or 30). We generally do not recommend older methods such as: (1) simple regression of the dependent variable on the time measure, (2) analyzing a single summary number that indexes change for each subject, or (3) a GLM approach with a fixed-subject effect. We recommend the following, though only under restrictive situations: (1) repeated-measure ANCOVA, (2) ANCOVA for two time points, (3) GEE, and (4) latent curve growth models. In more general cases, we advise using (5) random-effect models.

Acknowledgments

Thanks go to Drs. Richard Jones, Rebecca Betensky, John Growdon, and David A. Schoenfeld for criticism and suggestions. Funded support: NIA K23AG027171 (Atri).

Appendix

Illustrative SAS Program Code to Run a Random-Effect Longitudinal Analysis

The SAS program below produces a preliminary raw data 'spaghetti longitudinal scatterplot', and then runs a random-effect analysis to test if the relation of a dependent variable to time (years) in the study is different for a group of patients treated with an experimental medication versus those given only a placebo. The underlying progression is tested as to whether it is also nonlinear (follows a quadratic function). Nonsignificant terms are to be removed in backward elimination (see SAS documentation for details).

Assume in the Program Below:
Group Level Variables:
Group = the treatment group classification variable
(either 'medication' or 'placebo')
Subject Level Variables:
Subject_ID = a unique subject identifier (a character variable)
Education = years of education
Age_Baseline = age of subject at initial visit
Dur_Baseline = duration of illness (in years) at initial visit
Visit Level Variables:
Years_in_Study = years in the study (the random time variable)
Years_in_Study_Sq = the square of years in the study (to determine
quadratic curvilinear effects)
Dep_Var = the dependent variable

Therefore, the rectangular data set to be analyzed, sorted by Visit Date within Subject_ID within Group, would look like the below with one row for each visit for each subject:

Group	Subject ID	Edu-cation	Age base	Duration base	Year study	Year study square	Dependent variable	Other variables (visit date, sex, etc.)
Medication	1	12	65	3	0	0	12	
Medication	1	12	65	3	1	1	14	
Medication	1	12	65	3	2	4	22	
Medication	3	20	81	2	0	0	3	
Medication	3	20	81	2	2	4	5	
Medication	3	20	81	2	4	16	3	
Medication	3	20	81	2	7	49	7	
etc.								
Placebo	2	16	85	5	0	0	9	
Placebo	2	16	85	5	1	1	11	
Placebo	4	8	72	1	0	0	5	
Placebo	4	8	72	1	2	4	(missing)	
Placebo	4	8	72	1	3	9	7	
etc.								

SAS Program:

```
*First make a “spaghetti plot” of the raw data vs Years in
Study. (You could also make ones vs. Age at Visit or Duration
of Illness at Visit);

Goptions dev=emf   ftext='Arial'  htext=1  gsfname=grafout
nofileonly  hsize=6 in  vsize= 5 in ;

Filename grafout ‘(the path to the file directory goes here)' ;

Proc Sort;
  By  Subject_ID  Years_in_Study ;

Proc gplot data= (SAS data set name goes here) ;
  Plot  Dep_Var*Years_in_Study=Subject_ID  / nolegend
     haxis= … to …  by …     vaxis= … to …  by …  ;
  Symbol  value=circle  interpol=join   repeat=5000(any arbitrarily
high number);
  Title ‘Spaghetti Plot of Raw Longitudinal Data……………….';

*Run the Random effects analysis;

*The terms with asterisks are interactions of Group with the
linear and quadratic terms for years in the study;
*Both the linear and quadratic terms for years in study are
indicated as random effects in the Random statement. If the
variance of the quadratic term is nonsignificant, it will be
dropped from the random statement, and subsequently also if
the linear term is not significant;
*Type=un specifies an “unstructured” covariance matrix of the
random terms. If the covariances of the random terms are not
significant, an uncorrelated covariance matrix will be
specified with Type=vc (variance components);

Proc Mixed covtest noclprint data=(data set name goes here) ;
  Class  Subject_ID  Group ;
  Model  Dep_Var =
           Education
           Age_Baseline  Dur_Baseline
           Years_in_Study  Years_in_Study_Sq
           Group
           Group*Years_in_Study  Group*Years_in_Study _Sq / s ;
  Random  Intercept    Years_in_Study   Years_in_Study_Sq
                /  Subject=Subject_ID(Group)  Type=un  Gcorr ;
  Title ‘Random Effects Longitudinal Analysis…………………………..';
```

DOI: 10.1159/000330228
Published online: October 26, 2011

References

1 SAS/Stat User's Guide, version 9.2. Cary, SAS Institute, 2011.

2 Mplus Statistical Software. Los Angeles, Muthen & Muthen, 2011.

3 SPSS Software. Chicago, Illinois, SPSS, 2011.

4 JMP Software. Cary, SAS, 2011.

5 Zeger SL, Liang KY: An overview of methods for the analysis of longitudinal data. Stat Med 1992;11: 1825–1839.

6 Palta M, Lin CY: Latent variables, measurement error, and methods for analyzing longitudinal binary and ordinal data. Stat Med 1999;18:385–396.

7 Zeger SL, Liang KY, Albert PS: Models for longitudinal data: a generalized estimating equation approach. Biometrics 1988;44:1049–1060.

8 Rutter CM, Elashoff RM: Analysis of longitudinal data: random coefficient regression modeling. Stat Med 1994;13:1211–1231.

9 Edwards LJ: Modern statistical techniques for the analysis of longitudinal data in biomedical research. Pediatr Pulmonol 2000;30:330–344.

10 Twisk JW: Longitudinal data analysis: a comparison between generalized estimating equations and random coefficient analysis. Eur J Epidemiol 2004;19:769–776.

11 Adamis D: Statistical methods for analyzing longitudinal data in delirium studies. Int Rev Psychiatry 2009;21:74–85.

12 Petkova E, Teresi J: Some statistical issues in the analyses of data from longitudinal studies of elderly chronic care populations. Psychosom Med 2002;64:531–547.

13 Gibbons RD, Hedeker D, Elkin I, Waternaux C, Kraemer HC, Greenhouse JB, Shea MT, Imber SD, Sotsky SM, Watkins JT: Some conceptual and statistical issues in analysis of longitudinal psychiatric data: application to the NIMH treatment of Depression Collaborative Research Program dataset. Arch Gen Psychiatry 1993;50:739–750.

14 Diggle PJ, Heagerty P, Kung-Yee L, Zeger SL: Analysis of Longitudinal Data, ed 2. New York, Oxford University Press, 2002.

15 Fitzmaurice GM, Laird NM, Ware JH: Applied Longitudinal Analysis. Hoboken, Wiley, 2004.

16 Azar B: APA statistics task force prepares to release recommendations for public comment. APA Monitor Online, May 1999, vol 30.

17 Wilkinson L, APA Task Force on Statistical Inference: Statistical methods in psychology journals: guidelines and explanations. Am Psychol 1999;54:594–604.

18 Brooks JO, Kraemer HC, Tanke ED, Yesavage JA: The methodology of studying decline in Alzheimer's disease. J Am Geriatr Soc 1993;41:623–628.

19 Muthén B: Latent variable analysis: growth mixture modeling and related techniques for longitudinal data; in Kaplan D (ed): Handbook of Quantitative Methodology for the Social Sciences. Newbury Park, Sage, 2004, pp 345–368.

20 Myers JL: Fundamentals of Experimental Design, ed 3. Boston, Allyn & Bacon, 1979.

21 Girden ER: ANOVA: Repeated Measures; in: Quantitative Applications in the Social Sciences. Thousand Oaks, Sage, 1992, vol 84.

22 Greenhouse SW, Geisser S: On methods in the analysis of profile data. Psychometrika 1959;24:95–112.

23 Huynh H, Feldt LS: Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. J Educ Stat 1976;1:69–82.

24 Longford NT: Random Coefficient Models. New York, Oxford University Press, 1993.

25 Bryk AS, Raudenbush SW: Hierarchical Linear Models: Applications and Data Analysis Methods. Advanced Quantitative Techniques in the Social Sciences Series. Thousand Oaks, Sage, 1992.

26 Kreft I, Leeuw JD: Introducing Multilevel Modeling. Thousand Oaks, Sage, 1998.

27 Singer JD: Using SAS Proc Mixed to fit multilevel models, hierarchical models, and individual growth models. J Educ Behav Stat 1998; 24:323–355.

28 Atri A, Dodd JS, Yang FM, Locascio J: Presence of extrapyramidal motor signs predict worse cognitive and functional trajectory of decline in mild to moderate AD. Alzheimers Dement 2010;6:S350 (P2–125).

29 Atri A, Shaughnessy LW, Locascio JJ, Growdon JH: Long-term course and effectiveness of combination therapy in Alzheimer's disease. Alzheimer Dis Assoc Disord 2008;22:209–221.

30 Locascio JJ, Fukumoto H, Yap L, Bottiglieri T, Growdon JH, Hyman BT, Irizarry MC: Plasma amyloid protein and C-reactive protein in relation to rate of progression of Alzheimer disease. Arch Neurol 2008;65:776–785.
31 Hardin JW, Hilbe JM: Generalized Estimating Equations. Boca Raton, Chapman & Hall/CRC, 2003.
32 Pan W: Akaike's information criterion in generalized estimating equations. Biometrics 2001;57:120–125.
33 Muthén B: Statistical and substantive checking in growth mixture modeling. Psychol Methods 2003;8:369–377.
34 Mehta PD, Neale MC: People are variables too: multilevel structural equations modeling. Psychol Methods 2005;10:259–284.
35 Bollen KA: Structural Equations with Latent Variables. New York, Wiley, 1989.
36 Gibbons RD: Identifying biological subtypes in psychiatric research; in Gibbons RD, Dysken MW (eds): Statistical and Methodological Advances in Psychiatric Research. New York, Spectrum, 1983.
37 Locascio JJ, Cordray DS: A reanalysis of 'Lord's paradox'. Educ Psychol Meas 1983;43:115–126.
38 Allison PD: Event History Analysis; in: Quantitative Applications in the Social Sciences. Thousand Oaks, Sage, 1984, vol 46.
39 Yamaguchi K: Event History Analysis; in : Applied Social Research Methods. Thousand Oaks, Sage, 1991, vol 28.
40 Lee ET: Statistical Methods for Survival Data Analysis. Belmont, Lifetime Learning Publications, 1980.
41 Cox DR, Oakes D: Analysis of Survival Data. New York, Chapman & Hall, 1984.
42 Box G, Jenkins GM, Reinsel G: Time Series Analysis: Forecasting and Control, ed 3. Upper Saddle River, Prentice-Hall, 1994.
43 McCleary R, Hay RA: Applied Time Series Analysis for the Social Sciences. Thousand Oaks, Sage, 1980.
44 McDowall R, McCleary R, Meidinger EE, Hay RA: Interrupted Time Series Analysis; in: Quantitative Applications in the Social Sciences. Newbury Park, Sage, 1980, vol 21.
45 Cook TD, Campbell DT: Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago, Rand McNally, 1979.
46 Locascio JJ, Jennings PJ, Moore CI, Corkin S: Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. Hum Brain Mapp 1997;5:168–193.
47 Gottman JM: Time-Series Analysis: A Comprehensive Introduction for Social Scientists. New York, Cambridge University Press, 1981.
48 SAS/Econometric Time Series (ETS) Analysis, version 9.2. Cary, SAS Institute, 2011.
49 Markus GB: Analyzing Panel Data; in: Quantitative Applications in the Social Sciences. Newbury Park, Sage, 1979, vol 18.
50 Locascio JJ: The cross-lagged correlation technique: reconsideration in terms of exploratory utility, assumption specification and robustness. Educ Psychol Meas 1982;42:1023–1036.
51 Cohen J, Cohen P, West SG, Aiken LS: Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Mahwah, Erlbaum, 2002.
52 Duncan OD: Introduction to Structural Equation Models. New York, Academic Press, 1975.
53 Hatcher L: A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling. Cary, SAS, 1994.
54 Kenny DA: Correlation and Causality. New York, Wiley, 1979.
55 Locascio JJ, Lee J, Meltzer HY: The importance of adjusting for correlated concomitant variables in psychiatric research. Psychiatry Res 1988;23:311–327.
56 Allison PD: Missing Data; in: Quantitative Applications in the Social Sciences. Newbury Park, Sage, 2002, vol 136.
57 NCSS Statistical and Power Analysis Software. Kaysville, NCSS, 2007.
58 Cohen J, Cohen P: Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, ed 2. Mahwah, Erlbaum, 1983.