Published in final edited form as:

Stat Med. 2012 September 28; 31(22): 2457–2460. doi:10.1002/sim.5500.

Novel Statistical Methodology for Analyzing Longitudinal Biomarker Data

Paul S. Albert^a and Enrique F. Schisterman^b

Paul S. Albert: albertp@mail.nih.gov; Enrique F. Schisterman: schistee@mail.nih.gov

^aBiostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics, and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Boulevard Room 7B05, Bethesda, Maryland 20892-7510

^bEpidemiology Branch, Division of Epidemiology, Statistics, and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Boulevard Room 7B03, Bethesda, Maryland 20892-7510

Introduction

Longitudinal designs for biomarker traits are very appealing for a number of reasons. First, the collection of repeated biomarker measurements over time allows investigators to characterize both the overall and individual trajectories in these measurements. This helps investigators to understand overall effects as well as how individuals vary around these average effects. Distinguishing individual heterogeneity from average trend is impossible without longitudinal data. This distinction is important for understanding individual between-subject in its own right, as well as in developing predictors of disease outcome from longitudinal marker data. There are many statistical issues that are unique to longitudinal data analysis ranging from correctly specifying realistic models which characterize average changes over time and individual variation, accounting for observations that are taken at unequally spaced observation times which may differ between individuals, to accounting for missing data due to occasionally missed measurements or loss to follow-up. When the reason for missing data is not related to the underlying observation process, most longitudinal data analysis methods can be routinely applied. However, complex modelling is necessary when the reason for missing an observation or dropout is directly related to the underlying outcome process [1–2]. Statistical techniques that appropriately analyze continuous and discrete longitudinal data have been discussed by many authors. Diggle, et al. [3] provides a good summary of these techniques for the practitioner, while Fitzmaurice et al. [4] discusses recent advances from a more technical perspective.

The three papers in this section present innovative approaches for solving important methodological problems in biomarker studies. Most standard longitudinal data analysis techniques assume that the measurement process is independent of the outcome process. However, when the outcome is related to the underlying observation process, failure to properly account for the observation process can result in bias [5]. Schildcrout et al. [6] proposed an approach for regression modelling for longitudinal binary data where the outcome is measured with probabilities determined by an auxiliary variable that is related, but not identical to the binary outcome. Modelling multivariate longitudinal biomarker data is an increasingly common issue with the advent of an increasing number of important biomarker techniques (e.g., multiple cytokines in a multiplex assay or multiple reproductive hormones that describe complex biological phenomenon). Although various authors have proposed techniques for analyzing multivariate longitudinal data [7–8 and others], little work has focused on estimating multivariate linear mixed models under known constraints in the relationship between the multivariate outcomes. Roy et al [9] proposes such an

approach for analyzing multivariate reproductive hormones over the menstrual cycle, where established relationships have solid biological justification. Finally, the collection of longitudinal biomarker data as disease predictors is becoming an increasingly important design in diagnostic medicine. Pfeiffer et al. [10] proposes flexible non-parametric methodology for combining biomarkers that exploits the longitudinal structure in formulating a score that is used for predicting a binary outcome.

Outcome-Dependent Sampling

Schildcrout et al. proposed a regression-based approach for binary longitudinal data when the sampling is related to an auxiliary variable, which is related to the binary outcome of interest. The methodology development was nicely motivated by the BioCycle study [11] where interest was on examining the relationship between dietary fiber intake and a marker of ovulation such as a surge in luteinizing hormone (LH>20 IU/L). Based on an at-home fertility monitor that required daily urine samples, results of the monitor were used to choose days for serum draws that were likely to result in an LH surge. Thus, by design, the measurement process was related to the outcome, resulting in a potential bias if standard regression methods are employed. The essence of this approach is the recognition that the marginal regression model of interest can be related to a regression model conditional on the observation process by a simple unknown off-set term. Further, this off-set term can be estimated by modeling the binary auxiliary variable as a function of the outcome and shared covariates with the model of interest. A joint estimating equation for parameter estimation is proposed with a robust sandwich estimator for variance estimation.

The approach of Schildcrout et al. is quite elegant and very useful for the situation in which both the auxiliary and outcome variables are both binary and the link function relating these variables to key covariates is the logistic link. However, when this is not the case, we need to find alternative approaches. For example, of interest in the Biocycle study is not only the effect of fiber intake on the occurrence of an LH surge, but also on the effect of fiber on other more continuously defined hormones such as estrogen, progesterone, and follicle-stimulating hormone (FSH). Further, we may be interested in understanding the effects of environmental factors of environmental exposures such as heavy metals on menstrual cycle patterns of these reproductive hormones [12]. Schildcrout et al. does discuss an inverse probability weighting generalized estimating equations (IPW-GEE) approach as an alternative, showing that it has marked efficiency loss relative to the proposed approach. Although, the inefficiency is a clear disadvantage, the IPW-GEE extends quite simply to cases where the outcome and auxiliary variables are not binary and where the effect of covariates on complex hormonal patterns may be examined.

An alternative approach to Schildcrout's proposed approach which may be more efficient than the IPW-GEE approach is a shared random parameter model (see chapter 19 in [4]) where the dependence between the observed outcome Y_{ij} , auxiliary variable, Z_{ij} , and the sampling variable S_{ij} conditional on the auxiliary variable is introduced through a shared set of random effects. Suppose that the shared random effects are denoted as b_i , then the individual contribution of the likelihood can be written as $P(y_i, z_i, s_i) = \int b g(y_i|b)h(z_i|b)k(s_i|z_i, b)m(b)db$, where g is the density of the vector of observed responses y_i given the random effects, h is the density of the auxiliary variable z_i given the random effects, k is the density of the sampling outcomes given the auxiliary variable and random effects, and m is the density of the random effects. Assuming that the random effects b are multivariate normal with a small dimension, the integration can be conducted using Gaussian quadrature. Marginal inference on y can be conducted exactly by assuming a generalized linear mixed model with a probit link or approximately using a logit link [13]. The above approach simplifies some for the BioCycle application since it is reasonable to assume that the

probability of sampling only depends on the auxiliary variable (by design) and not on the latent random effects b . The implementation and comparison of the above approaches will be the subject of future research.

Modelling Hormonal Dynamics During the Menstrual Cycle

There is work in the statistics literature on modeling menstrual cycle patterns. For example, Zhang, et al. [14–15] propose stochastic mixed models that model the menstrual cycle with a flexible periodic non-parametric curve for the mean structure, random effects which allow for differences between individuals, and an Ornstein-Uhlenbeck process to allow for serial correlation. However, little research has focused on modeling the complex multivariate relationships between multiple reproductive hormones. Meyer, et al [16] have proposed functional data analysis techniques that allows investigators to characterize the complex inter-relationships between multiple reproductive hormones. Roy et al., provide a nice contribution to the literature where they exploit known relationships between reproductive hormones in order to more efficiently estimate the relationships between F2-isoprostanes levels and hormonal levels.

Roy et al. propose a multivariate linear mixed model where constraints are placed on the mean structure that is based on well known established biological inter-relationships between reproductive hormones. These constraints are placed on subject-specific means, which are scientifically sensible in this application. In the application of reproductive hormones, for example, it is well known that on a subject specific level, the peak levels in estrogen are followed by a surge in LH values, which in turn, are followed by a rising progesterone level. Incorporating these multivariate constraints on mean values as well as constraints on how individual reproductive hormones vary across the menstrual cycle may improve efficiency in estimation.

For the application, Roy et al. [9] mentions that the results for the constrained analysis were in fact very similar to those for the unconstrained analysis, leading us to wonder under what situations this constrained analysis would help with efficiency. Our intuition is that this payoff will be largest in situations where the residual variation σ_e^2 is large.

In this approach, a cleaver and efficient modeling is proposed by exploiting either firm or soft constraints on the relationships between hormones and within a hormone over time. For the menstrual cycle function application, these relationships are very well established in most cases. However, it still remains interesting to ask the question, how wrong will your inferences be if in fact these assumptions are not met. Examining this issue would be interesting further work.

Prediction from longitudinally collected measurements

Predicting disease outcomes from longitudinal biomarker data is an increasingly important problem in many medical areas. For example, much of the early work in joint modeling of longitudinal and survival data was motivated by trying to understand the relationship between longitudinal CD4 counts and survival in patients with AIDS [17–18, among others] and PSA velocity and prostate cancer [19, among others]. An important feature in these joint models is that the longitudinal process has informative censored in the sense that those who have early disease occurrence have fewer longitudinal observations. Not properly accounting for this informative censoring with joint modeling can result in biased estimation of the relationship between the two processes as well as poor prediction of the disease outcome from the longitudinal measurements.

Pfeiffer et al. [10] propose a very flexible approach for prediction when this informative censoring is not of concern (i.e., the number of longitudinal observations is not related to the binary disease outcome). Assuming that the first two moments of the predictors can be decomposed into a time and biomarker component with a Kronecker product structure that exploits the longitudinal nature of the predictors, they propose an optimal strategy for developing a score that can be used for classification. Further, the authors show marked improvements in prediction compared with a similar non-parametric approach, which does not exploit this longitudinal structure. Pfeiffer et al.'s approach is particularly appealing since no parametric assumptions on the longitudinal process are made. However, with the flexibility and robustness of this approach come some potential complications. For example, it is not clear how to extend the model to allow for irregularly spaced measurements or missing data. Extensions of this methodology which allows for this real life complication would be interesting.

Pfeiffer et al. illustrate their methodology on data from an interesting case-control study where there are four longitudinal measurements. Particularly with the increasing occurrence of high dimensional longitudinal data (e.g. Longitudinal “omics” data), it would be interesting to examine the performance of this methodology when the number of predictors is large.

Conclusion

The three papers in this section present interesting and novel approaches for analyzing complex longitudinal biomarker data. Examined in total, these papers nicely address some of the fundamental analytical issues in longitudinal data; namely, complex modeling, accounting for outcome dependent sampling (or the related missing data problem), and prediction of disease outcome from longitudinal data. Namely, innovative new approaches motivated through the analytic complications in “real life” biomarker studies.

Acknowledgments

This research was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Maryland, USA.

References

1. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*. 1988; 44:175–188.
2. Albert PS, Follmann DA, Wang SA, Suh EB. A latent autoregressive model for longitudinal binary data subject to informative missingness. *Biometrics*. 2002; 58:631–642. [PubMed: 12229998]
3. Diggle, PJ.; Heagerty, P.; Liang, KY.; Zeger, SL. *Analysis of Longitudinal Data*. Oxford University Press; Oxford: 2002.
4. Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G., editors. *Longitudinal Data Analysis*. Chapman & Hall/CRC; Boca Raton: 2009.
5. Schildcrout JS, Heagerty PJ. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*. 2008; 9:735–749. [PubMed: 18372397]
6. Schildcrout JS, Mumford SL, Chen Z, Heagerty PJ, Rathouz PJ. Outcome-dependent sampling for longitudinal binary response data based on a time-varying auxiliary variable. *Statistics in Medicine*. On-line 2011.
7. Dubin JA, Muller HG. Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*. 2005; 100:872–881.
8. Fieuws S, Verbeke G. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*. 2006; 62:424–431. [PubMed: 16918906]

9. Roy A, Danaher MR, Mumford SL, Chen Z. A Bayesian order restricted model for hormonal dynamics during menstrual cycles of healthy women. *Statistics in Medicine*. On-line 2011.
10. Pfeiffer RM, Forzani L, Bura E. Sufficient dimension reduction for longitudinally measured predictors. *Statistics in Medicine*. On-line 2011.
11. Gaskins AJ, Mumford SL, Zhang C, Wactawski-Wende Hovey KM, Whitcomb B, Howards PP, Perkins NJ, Yeung E, Schisterman EF. for the BioCycle Study Group. Effect of daily fiber intake on reproductive function: the BiocCycle Study. *American Journal of Clinical Nutrition*. 2009; 90:1061–1069. [PubMed: 19692496]
12. Pollack AZ, Schisterman EF, Goldman LR, Mumford SL, Albert PS, Jones RL, Wactawski-Wende J. Cadmium, lead, and mercury in relation to reproductive hormones and anovulation in premenopausal women. *Environmental Health Perspectives*. 2011; 119:1156–61. [PubMed: 21543284]
13. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*. 1988; 44:1049–1060. [PubMed: 3233245]
14. Zhang D, Lin X, Raz J, Sowers M. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*. 1998; 93:710–719.
15. Zhang D, Lin X, Sowers M. Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles. *Biometrics*. 2000; 56:31–39. [PubMed: 10783774]
16. Meyer PM, Zeger SL, Harlow SD, Sowers M, Crawford S, Luborsky JL, McConnell DS, Randolph JF, Weiss G. Characterizing daily urinary hormone profiles for women at midlife using functional data analysis. *American Journal of Epidemiology*. 2007; 165:936–945. [PubMed: 17267418]
17. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modelling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*. 1995; 90:27–37.
18. Wang Y, Taylor JMG. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*. 2001; 96:895–905.
19. Yu M, Law NJ, Taylor JMG, Sandler HM. Joint longitudinal survival cure models and their application to prostate cancer. *Statistica Sinica*. 2004; 14:835–862.