# Emerging Issues in the Analysis of Longitudinal Data

Charmaine Dean (Simon Fraser University),
Xihong Lin (Harvard University),
John Neuhaus (University of California at San Francisco),
Liqun Wang (University of Manitoba),
Lang Wu (University of British Columbia),
Grace Yi (University of Waterloo)

August 16 - August 21, 2009

## 1 Overview of the Field

Longitudinal data arise frequently in practise, either in observational studies or in experimental studies. In a longitudinal study, individuals in the study are followed over a period of time and, for each individual, data are collected at multiple time points. That is, the defining feature of a longitudinal study is that multiple or repeated measurements of the same variables are made for each individual in the study over a period of time. A key characteristic of longitudinal data is that observations within the same individual or cluster may be correlated, which motivates most of the statistical methods for the analysis of longitudinal data. Although there have been extensive methodological developments for the analysis of longitudinal data, there are still many emerging issues arising in practice which motivates further research in this area. In particular, missing data, dropouts, and measurement errors are very common in longitudinal studies, and many of these issues need to be addressed simultaneously in order to draw reliable conclusions from the data. Moreover, longitudinal trajectories of observed data are often very complex. Parametric statistical models may not be flexible enough to capture the main features of the longitudinal profiles, so semiparametric or nonparametric statistical models are particularly attractive. Therefore, statistical analyses of complex longitudinal data can be very challenging, and much research remains to be done.

Specifically, the following problems are common in longitudinal studies: (i) longitudinal data may either be continuous or categorical or a mixture of both; (ii) longitudinal data trajectories may be highly complicated, and there may be large variations between individuals; (iii) there are often missing data or dropouts; (iv) some variables may be measured with errors; (v) longitudinal data may be associated with time-to-event data, and joint modelling may be necessary; and (vi) in some studies the number of variables may be large while the sample sizes may be small. In longitudinal data analysis, new statistical methods are required to address one or more of the above problems since standard methods are not directly applicable. For example, missing data or dropouts are almost inevitable in many longitudinal studies, and ignoring missing data or measurement errors or the use of naive methods may lead to severely biased or misleading results (Little and Rubin, 2002; Carroll et al. 2006).

There has been extensive research for the analysis of longitudinal data in the last few decades. Diggle et al. (2002) and Fitzmaurice et al. (2008) provided a comprehensive overview of various models and methods for the analysis of longitudinal data, among others. Commonly used models for longitudinal data include:

- *mixed effects models*: in these models random effects are introduced to incorporate the between-individual variation and the within-individual correlation in longitudinal data;

- *generalized estimating equations (GEE) models*: in these models the mean structure and the correlation structure are modelled separately without distributional assumptions for the data;

- *transitional models*, in these models the within-individual correlation is modelled via Markov structures.

- *nonparametric models* and *semiparametric models*: in these models the mean structures are modelled semiparametrically or nonparametrically or the distributional assumptions are assumed to be nonparametric, so these models are more flexible than parametric longitudinal models.

- *Bayesian models*: prior information or information from similar studies are incorporated for Bayesian inference, and the advance of Markov chain Monte Carlo (MCMC) methods has led to rapid developments of Bayesian methods.

Each of these modelling approaches offers its own advantages and disadvantages. For example, mixed effects models allow for individual-specific or subject-specific inference but require distributional assumptions, and GEE models are robust to distributional assumptions but may be less efficient. Moreover, transitional models may be particularly attractive for discrete data, Bayesian models borrow information from previous or similar studies, and nonparametric or semiparametric models are appealing for complex longitudinal data.

There is also an extensive literature on missing data and measurement errors. For missing data problems, Little and Rubin (2002) and Molenberghs and Kenward (2007) provided an overview of general models and methods. It is known that naive methods such as the complete-case method and the last-value-carried-forward method for missing data problems often lead to biased or misleading results. Formal methods for missing data include

- multiple imputation methods,

- likelihood inference using EM algorithms,

- single imputation methods with variance adjustments,

- weighted GEE methods,

- Bayesian methods.

These formal missing data methods can incorporate missing data mechanisms and provide valid statistical inference. For measurement error problems, Carroll et al. (2006) provided a overview of commonly used models and methods. It is known that naive methods which ignore measurement errors may lead to biased results and appropriate methods must be used for reliable inference. To formally address measurement errors, two general approaches are often considered:

- functional modelling approach, where no distributional assumptions are made for the true but unobserved covariates. For this approach, commonly used methods include regression calibration and simulation extrapolation (SIMEX).

- structural modelling approach, where models or distributions are typically assumed for the true but unobserved covariates. For this approach, commonly used methods include likelihood methods and Bayesian methods.

These formal methods may correct measurement errors and produce reliable statistical inference. For complex longitudinal data with missing values or measurement errors, further research is needed to extend the general ideas and methods.

In summary, analysis of longitudinal data has received much attention, especially in recent years. Although there have been extensive developments of statistical models and methods for the analysis of longitudinal data, there are many complex issues and problems to be addressed or solved, due to the complexities of longitudinal data in practice. Highly complicated longitudinal data arising in practice are challenging for statisticians, but they also provide great opportunities for research and advance of this important subject.

## 2 Recent Developments and Open Problems

There has been extensive research in statistical methods for longitudinal data or clustered data. Recent developments are reviewed in Carroll, Ruppert, Stefanski, and Crainiceanu (2006), Fitzmaurice, Davidian, Molenberghs, and Verbeke (2008), McCulloch, Searle, and Neuhaus (2008), Molenberghs and Kenward (2007), and Wu (2009), among others. It is difficult to give a complete list of recent developments due to the massive literature. In the following, we discuss some of the recent developments and some open problems.

In the analysis of longitudinal data, three types of models are commonly used: mixed effects models, GEE models, and transitional models. We first discuss some recent developments for mixed effects models. Mixed effects models introduce random effects in classical regression models for cross-sectional data to account for within-individual correlation and between-individual variation in longitudinal data. Distributional assumptions are often made for the within-individual random errors and for the random effects in the mixed effects models. In practice, the distributional assumptions for random effects may be difficult to check since the random effects are unobservable. Professors Charles McCulloch and John Neuhaus at the University of California at San Francisco are currently studying how mis-specifications of the random effects distributions may affect estimation and inference for generalized mixed effects models. Lai and Shih (2003) considered nonparametric distributions for the random effects, in which the distributions of the random effects in nonlinear mixed effects models are completely unspecified. Lai, Shih, and Wong (2006) proposed a hybrid estimation method for mixed effects models. However, nonparametric methods often require rich within-individual data. Professor Xihong Lin at Harvard University has been investigating semiparametric models for longitudinal data with measurement errors and missing data.

Computational challenges for generalized linear mixed effects models and nonlinear mixed effects models still require further investigation. The likelihood method is a standard estimation approach for generalized linear and nonlinear mixed effects models, but the likelihood functions typically involve multi-dimensional and intractable integrations. Numerical or Monte-Carlo methods may be computationally intensive and may even offer computational difficulties if the dimensions of random effects are not small. Computation may become a major challenge in the presence of missing data and measurement errors. Approximated methods, such as that based on Taylor approximations or Laplace approximations, have been proposed and widely used. However, Lin and Breslow (1996) showed that these approximate methods may be biased for generalized linear mixed models with binary responses. More recently, Joe (2008) showed that the accuracy of these approximate methods may be poor for mixed effects models with binary or count responses, such as logistic regression models with random effects. Lee, Nelder, and Pawitan (2006) have proposed higher order Laplace approximations for generalized linear mixed models. A comprehensive evaluation of these approximate methods is still required. The performance of the foregoing methods for mixed effects models with missing data and measurement errors remains to be investigated.

Generalized estimation equation (GEE) methods are another popular approach for the analysis of longitudinal data. GEE methods only assume the first two moments without specific distributional assumptions, so they are robust against distributional assumptions. For longitudinal data, a working correlation matrix is typically assumed in a GEE method. GEE methods enable one to estimate regression parameters consistently even when the correlation structure is misspecified. For generalized linear mixed effects models, however, Chaganty and Joe (2004) showed that the choices of valid working correlation matrices can be very limited, and inappropriate choices of the working correlation matrices may lead to misleading results. Moreover, under mis-specifications of working correlation matrices, the estimators of the regression parameters can be inefficient. Qu, Lindsay, and Li (2000) and a series of articles thereafter introduced a method of quadratic inference functions that does not involve direct estimation of the correlation parameters, and that remains optimal even if the working correlation structure is misspecified. The idea is to represent the inverse of the working correlation matrix by the linear combination of basis matrices. They showed that under misspecified working correlation assumptions these estimators are more efficient than GEE estimators. This method may be applied to a wide variety of problems, including longitudinal data with missing values and measurement errors, so much research remains to be done. Yi and Cook (2002) and a series of articles thereafter studied GEE methods for clustered longitudinal data with missing values in which the correlation within clusters is incorporated, in addition to correlation between longitudinal measurements.

Transitional models assume Markov structures for longitudinal dependence. Nathoo and Dean (2008) considered multistate transitional models in which at any time point individuals may be said to occupy one of a discrete set of states and interest centers on the transition process between states. They developed a hierarchical modeling framework in which the processes corresponding to different subjects may be correlated spatially over a region and continuous-time Markov chains incorporating spatially correlated random effects are introduced. Yi and Cook (2002), Cook et al. (2004), and their recent work considered marginal models for incomplete longitudinal data arising in clusters. They used odds ratio and Markov structures to model dependence between multivariate discrete longitudinal data, incorporating missing data. Modeling clustered or multivariate longitudinal data with missing values or measurement errors can be challenging, both mathematically and computationally.

Joint modelling longitudinal data and survival data has received much attention in recent years. Such joint models are required in survival models with measurement errors in time-dependent covariates, longitudinal models with dropouts or certain events of interest, and many other situations in longitudinal studies. Professor Joseph Ibrahim is working on diagnostic measures for assessing the influence of observations and model misspecification for joint models of longitudinal and survival data, in the presence of missing data. Professor Jeremy Taylor is studying individual predictions of future event times for censored subjects using joint models. Professor Bin Nan is investigating joint modeling of longitudinal and survival data when the event time is a covariate. Professors Charmaine Dean and Farouk Nathoo are considering longitudinal studies in forestry where trees are subject to recurrent infection and the hazard of infection depends on tree growth over time. They have developed a joint model for infection and growth where a mixed non-homogeneous Poisson process is linked with a spatially dynamic nonlinear model representing the underlying height growth trajectories. Much work remains to be done for joint models with missing data and measurement errors. In particular, during the workshop many people emphasized the importance of software developments for joint models so that these models may be more widely used in practice.

A characteristic of longitudinal studies is that there are often missing data, dropouts, and measurement errors. Thus, in practice when modeling longitudinal data one often also needs to address missing data, dropouts, and measurement errors. Formal approaches for addressing missing data may require modeling of the missing data or dropout processes. As pointed out by Professor Roderick Little at the University of Michigan, it is important to check model assumptions and provide model justifications. When the missing data is nonignorable in the sense that the missingness may depend on the missing values, a missing data mechanism must be assumed and be incorporated in likelihood inference. However, such assumed missing data models are not testable based on observed data. Thus, sensitivity analyses under different missing data mechanisms is required. Professor James Carpenter and Mike Kenward at the University of London organized and led a discussion session in the workshop on sensitivity analysis for missing data problems. There have also been substantial developments in measurement error problems, as reviewed in Carroll, et al. (2006). Professor Xihong Lin has been working on measurement error problems in semiparametric models. Recently, there are interests in jointly modeling missing data and measurement errors. Wang (2004) proposed moment-based methods for nonlinear models with Berkson measurement errors, where only the first two moments are required for estimation and inference without distributional assumptions.

# 3 Presentation Highlights

## 3.1 Monday, August 17

The workshop began on Monday, August 17. Professor Charmaine Dean from Simon Fraser University chaired the sessions on Monday morning, and Professor Grace Yi from University of Waterloo chaired the sessions on Monday afternoon. The focus on Monday's talks is on missing data and measurement error problems in longitudinal studies.

The workshop began with an excellent presentation by Professor Roderick Little from University of Michigan. He provided an overview of missing data problems in longitudinal studies. Missing data are very common in longitudinal studies because of attrition, missed visits, dropouts, and other problems. Professor Little emphasized the importance of model assumptions and model justifications, pointed out that clever estimation methods will not help, and one should keep models simple and carefully design the studies to avoid missing data. He discussed the pros and cons of different forms of likelihood inference, Bayes and multiple imputation, robust estimation, GEE methods, the complete-case method, and other ad-hoc methods. He also reviewed selection models and pattern-mixture models for missing data problems, as well as sensitivity analysis. Following Little's presentation, Professor Joan Hu from Simon Fraser University presented a talk on Cox proportional hazards models with non-random missing covariates using a likelihood-based estimation method. Her research is motivated by a study for disease control. Following Hu's talk, Professor Annie Qu from University of Illinois at Urbana-Champaign gave a presentation on analysis of longitudinal data with data missing at random using an estimating function approach. Their approach differs from inverse weighted estimating equations and the imputation methods in that it does not require estimating the probability of missing data or imputing the missing data based on assumed models, and it is based on an aggregate unbiased estimating function which does not require the likelihood function.

Professor Mike Kenward from University of London discussed double robust estimators based on a multiple imputation method for missing data. Their method is based on Bang and Robins (2005) who showed how doubly robust estimators for monotone incomplete data problems can be obtained through a sequence of regressions, and they showed how reformulation of Bang and Robins (2005) approach in a multiple imputation framework leads to very convenient route for calculating doubly robust estimators, while at the same time providing an explicit and easily calculable variance estimator. They also extend the method to non-monotone missing value settings. Following Kenward's talk, Professor James Carpenter from University of London considered a class of models for multivariate mixtures of Gaussian, ordered or unordered categorical responses and continuous distributions that are not Gaussian, each of which can be defined at any level of a multilevel data hierarchy. He described a MCMC algorithm for fitting such models, and shows how this approach can be used to implement multilevel multiple imputation (assuming data are missing at random) and extended to allow imputation of missing data that is congenial/consistent with a complex multilevel model.

Professor Richard Cook from University of Waterloo presented marginal models for estimating treatment effects in cluster randomized longitudinal studies with incomplete responses and non-compliance. He proposed inverse weighted generalized estimating equation methods to address incomplete compliance data in a model for the compliance process and used a mean-scored approach to deal with the missing compliance data in the response model. Following Cook's talk, Dr. Baojiang Chen from University of Washington discussed models for longitudinal data where both the response and the covariates may be missing. A method based on inverse probability weighted generalized estimating equations was proposed, which incorporates the association between the missing data process and the response process.

Professor James Carpenter and Mike Kenward organized and led a discussion session on sensitivity analysis for missing data models. The speakers in the discussion session include Professor Andrea Rotnitky from Harvard University, Professor Joe Ibrahim from the University of North Carolina at Chapel Hill, and Professor Ray Carroll from Texas A & M University. Rotnitky discussed some issues of sensitivity analyses for inference in causal models. Ibrahim discussed Bayesian sensitivity analysis, proposed a perturbation model to simultaneously perturb the data, the prior, and the sampling distribution, and developed a Bayesian perturbation manifold to measure each perturbation in the perturbation model and applied the method to a wide variety of statistical models, allowing for missing data. Professor Raymond Carroll summarized the presentations and discussions in Monday's sessions and raised many interesting questions. For example, are we torturing investigators by doing fancy sensitivity analyses? What about just doing a few different approaches such as multiple imputations, augmented inverse probability weighting methods, and full model-based methods? He also pointed out that the last-observation-carried-forward method or the baseline-observation-carried-forward method may lead to severely biased results.

## 3.2 Tuesday, August 18

The presentations on Tuesday, August 18, focus on functional data, mixed effects models, and estimating equations. Professor Xihong Lin from Harvard University chaired the sessions on Tuesday morning, and Professor John Neuhaus from University of California at San Francisco chaired the sessions on Tuesday afternoons.

Tuesday's sessions began with a presentation by Professor Raymond Carroll from Texas A & M University. He discussed an efficient inference approach for additive models with repeated measures, where the additive model consists of a parametric component and an additive nonparametric component, in the presence of interactions. He derived efficient estimates using smooth backfitting and a Tukey-type 1-degree of freedom formulation, and derived a general profile-type score statistic which involves circumventing the need to solve an integral equation. He also proposed the "Carroll's law of nonparametric regression": "If things work out seamlessly for efficient kernel approaches, then they will work out for efficient spline methods. Thus, one can do theory for kernels and do practice for splines". Following Carroll's talk, Professor Hua Liang from University of Rochester discussed variable selections for semiparametric models with measurement errors. He explored variable selections for partially linear models when the covariates are measured with additive errors, and proposed two classes of variable selection procedures, penalized least squares and penalized quantile regressions, using the nonconvex penalized principle. He showed that the first procedure corrects the bias in the loss function caused by the measurement error by applying the so-called correction-for-attenuation approach, whereas the second procedure corrects the bias by using orthogonal residuals. Following Liang's talk, Professor Lu Wang from University of Michigan considered nonparametric regressions in longitudinal studies with dropout at random. She proposed inverse probability weighted (IPW) kernel generalized estimating equations (GEEs) and IPW seemingly unrelated (SUR) kernel estimating equations using either complete cases or all available cases, and showed that all these IPW kernel estimators are consistent when the probability of dropout is known by design or is estimated using a correctly specified parametric model. She also showed that the most efficient IPW kernel GEE estimator is obtained by ignoring the within-subject correlation while in contrast the most efficient IPW SUR kernel estimator is obtained by accounting for the within-subject correlation and is more efficient than the most efficient IPW kernel GEE counterpart.

Professor Naisyin Wang from University of Michigan presented functional latent feature regression models for data with longitudinal covariate process. She considered a joint model approach to study the association of nonparametric latent features of multiple longitudinal processes with a primary endpoint. She proposed estimation procedures and the corresponding supportive theory that allows one to perform investigation without making distributional assumptions of the latent features, and investigated the practical implications behind certain theoretical assumptions which aim at having a better understanding of where the estimation variation lies. Following Wang's talk, Professor Wenqing He from the University of Western Ontario discussed local linear regressions for clustered censored data. He presented a local linear regression method for the estimation of the relationship between censored response and covariates by considering a transformation of the censored response, and used simulation to assess the performance of the proposed method.

Professor Charles McCulloch from University of California at San Francisco discussed estimation efficiency problems in generalized linear mixed models under misspecified random effects distributions. Previous work has shown that incorrect specification of the random effect distribution typically produces little bias in estimates of covariate effects and very modest inaccuracy in predicted random effects, but few studies have assessed the effect of misspecification on standard errors and statistical tests. Professors McCulloch and Neuhaus examined the impact of a misspecified random effects distribution on estimation efficiency. They showed that linear mixed models are well-behaved in the sense that the random effects influence only the variance-covariance structure, not estimation of the fixed effects. For logistic regression models with random intercepts, they showed that (i) within cluster covariates show no loss of efficiency; (ii) between cluster covariates show loss of efficiency comparable to or better than a linear regression with assumed normal errors, so quite robust; (iii) fitting flexible distributional shapes is an easy way to check sensitivity of the results. Following McCulloch's talk, Mr. Daniel Li from the University of Manitoba presented a second-order least squares estimation method for mixed effects models. Li and Wang applied the second-order least squares method to estimate generalized linear mixed effects models where the distributions of the regression errors are nonparametric while those of random effects are parametric but not necessarily normal. They presented simulation studies of finite sample properties of the second-order least squares estimators and compared them with the maximum likelihood estimators.

Professor Peter Song from University of Michigan discussed analyzing unequally spaced longitudinal data with quadratic inference functions. Quadratic Inference Function (QIF) is getting increasingly popular, as an alternative to the well-known GEE method, to estimate parameters in the marginal models for longitudinal data. One limitation with the QIF is that it is currently only applicable for longitudinal data with equally spaced times. In his talk, Peter Song presented a generalized QIF method to relax this limitation. Following Song's talk, Professor Youngjo Lee from the Seoul National University discussed hierarchical generalized linear models (HGLMs) and variable selection. HGLMs provide a flexible and efficient framework for modeling non-Normal data when there are several sources of error variation, and they extend the familiar generalized linear models (GLMs) to include additional random terms in the linear predictor. Thus HGLMs bring a wide range of models together within a single framework and they also facilitate the joint modeling of mean and dispersion. He also showed how HGLM can be used for variable selection and showed how LASSO and its extension can be obtained via random-effect models.

## 3.3 Wednesday, August 19

The presentations on Wednesday, August 19, focus on joint models for longitudinal data and survival data. Professor Wei Liu from York University chaired the first session, and Professor Charmaine Dean from Simon Fraser University organized and chaired a discussion session on joint models. Wednesday afternoon is free, without formal scientific activities. In Wednesday evening Professor Andrea Rtonitzky from Harvard University offered a three-hour lecture on causal inference.

Professor Joseph Ibrahim from University of North Carolina at Chapel Hill began Wednesday's sessions with a presentation on local influence for joint models for longitudinal and survival data. He discussed diagnostic measures for assessing the influence of observations and model misspecification for longitudinal models and for joint models of longitudinal and survival data, in the presence of missing data. He proposed a local influence approach and examined various perturbation schemes for perturbing the models in this setting, and developed a perturbation manifold and various local influence measures to identify influential points and test model misspecification. Following Ibrahim's talk, Professor Jeremy Taylor from University of Michigan discussed using joint models for longitudinal and survival data to give individual predictions. He considered using a joint model to assist with individual prediction of future event times for censored subjects. The model and methods are developed in the context of a prostate cancer application where the longitudinal variable is PSA and the event time is recurrence of the cancer following treatment with radiation therapy. Estimates of the parameters in the model are obtained by MCMC techniques, and an efficient algorithm is developed to give individual predictions for subjects who were not part of the original data from which the model was developed. Many important statistical issues were discussed in his presentation. Following Taylor's talk, Professor Bin Nan from University of Michigan discussed joint modeling of longitudinal and survival data when the event time is a covariate. His research is motivated from estimation of the hormone profile, such as serum estradiol or follicle stimulating hormone, during menopausal transition. Due to limited follow up time, the age at the final menstrual period for many women in a study cohort is censored. He proposed a two-stage pseudo likelihood approach to estimate the hormone profile during menopausal transition using a nonparametric stochastic mixed model.

Professor Charmaine Dean from Simon Fraser University organized and chaired a discussion session on joint models of longitudinal data and survival data. Various important issues were raised and discussed, such as the current challenges in joint modeling, illustrations of various applications, and benefits and drawbacks of various approaches. Professor Farouk Nathoo from University of Victoria showed an interesting application of joint modeling in spatial statistics where the growth of trees is modelled using nonlinear mixed effects models. He also proposed various approaches for estimation and inference. An important issue that received much discussions is software developments for joint models. Currently, existing software for joint models is very limited. Developments of software, such as R packages, allow joint modelling methods to be more widely used in practice.

On Wednesday evening, Professor Andrea Rtonitzky from Harvard University offered a three-hour lecture on causal inference. The lecture was well received, with many live and interesting discussions.

## 3.4 Thursday, August 20

The presentations on Thursday, August 20, focus on important applications, binary, and count data. Professor Liqun Wang from University of Manitoba chaired the sessions on Thursday morning, and Professor Jiayang Sun from Case Western Reserve University chaired the sessions on Thursday afternoon.

Professor John Petkau from the University of British Columbia began Thursday's sessions with a presentation on an interesting application of neutralizing antibodies and the efficacy of interferon Beta-1b in multiple sclerosis clinical trials. He discussed the question of whether neutralizing antibodies (NAbs) impact on the efficacy of Type I interferons treatments, which is an unresolved scientific issue directly related to the question of how multiple sclerosis (MS) patients should be treated. He also described the initial analyses which raised the concern, and the analyses they have carried out to try to resolve this issue. A fascinating part of their project has been attempting to persuade the neurological community of the need for more detailed analyses of the clinical trial data than is customary in the field to fully address this issue. Following Petkau's talk, Professor Andrea Rtonitzky from Harvard University discussed estimation and extrapolation of optimal dynamic treatments and testing strategies from observational longitudinal data. They considered methods for using the data obtained from an observational database in one health care system to determine the optimal treatment regime for biologically similar subjects in a second health care system when, for cultural, logistical, and financial reasons, the two health care systems differ (and will continue to differ) in the frequency of, and reasons for, both laboratory tests and physician visits. Professor Tze Leung Lai from Stanford University discussed a dynamic empirical Bayes approach to econometric panel data via generalized linear mixed models. He first gave a brief review of the literature on credibility rate-making in insurance and default modeling of corporate loans in finance, particularly on the econometric models used to analyze the associated panel data. Then they proposed a new, unified class of dynamic empirical Bayes models for these longitudinal data and their subject-matter applications. The advantages of these models and their connections to generalized linear mixed models were also discussed and illustrated.

Dr. Taraneh Abarin from Samuel Lunenfeld Research Institute gave a talk on instrumental variable approach to covariate measurement errors in generalized linear models. They proposed a method of moments estimation for generalized linear measurement error models using the instrumental variable approach. They also proposed simulation-based estimators for the situation where the closed forms of the moments are not available. Following Abarin's talk, Mr. Zhijian Chen from University of Waterloo gave a talk on a marginal method for correlated binary data with misclassified responses. Much research in the literature has been directed to problems concerning error-prone covariates, and there is relatively little work on measurement error or misclassification in the response variable. They proposed a marginal analysis method to handle binary response which is subject to misclassification. Numerical studies were presented to assess the performance of the proposed methods.

Professor Renjun Ma from University of New Brunswick organized and chaired a discussion session on random effects modeling of longitudinal data with excessive zeros. He first described various applications of longitudinal data with excessive zeros in different subject areas and interesting datasets, and then he discussed different approaches to random effects modeling of longitudinal data with excessive zeros in the literature (models, estimation methods, etc.) and the relative advantages and limitations of these approaches. He also proposed new approaches to random effects modeling of data with excessive zeros.

Dr. Li Qin from Fred Hutchinson Cancer Center discussed a registration-based functional linear model for post-ART viral loads in patients with primary HIV infection. Traditionally, such data were analyzed by approximate parametric or dynamical models, but the parametric forms may be too restrictive while the dynamical models may be highly assumption dependent. The proposed model presents a trade-off between the parametric models and the flexible functional effects associated with the viral loads. L-splines are used to model the viral loads and account for the plausible monotonicity in the curves over time. Following Qin's talk, Professor Yang Zhao from University of Regina discussed likelihood methods for regression models with data missing at random. She extended the maximum likelihood methods to deal with missing data problems in longitudinal data analysis.

## 3.5 Friday, August 21

On Friday, August 21, Professors Xihong Lin and Grace Yi organized a discussion session on emerging issues of longitudinal data analysis. The session provided a brief summary of the presentations and discussions in the workshop, with discussions of some further issues. The formal session on Friday ended early since many participants needed to catch early shuttles and flights. Informal discussions continued until Friday afternoon. Professor Lang Wu from the University of British Columbia chaired the formal session on Friday.

Professor Xihong Lin from Harvard University discussed analysis of high-dimensional population-based "omics" data. Population-based "Omics" Studies are observational studies, including longitudinal studies, which contain a large number of subjects and massive high-throughout "omincs" data such as genomics, epigenomics, proteomics, and metabolomics. Genome-Wide Association Studies (GWAS) have recently become popular for identifying common gene variants for complex diseases, such as cancers. The goal is to identify genes or gene regions, gene-gene interactions that are associated with a disease or a phenotype. She discussed various approaches to analyze these data, including mixed effects models and GEE methods. Professor Grace Yi from the University of Waterloo provided a comprehensive overview of missing data problems and measurement error problems in longitudinal studies. She reviewed existing approaches, including likelihood methods, weighted GEE methods, selection models, pattern-mixture models, and shared-parameter models. She also discussed how to address measurement errors and missing data simultaneously and other important issues such as how to balance the complexity of modeling and interpretability of model parameters, model identifiability, model checking, sensitivity analysis, and computational issues. Professor James Carpenter from University of London provided a summary of the discussions on sensitivity analysis for longitudinal data with dropout. He argued that the most accessible way to frame discussion is to consider how post-withdrawal measures may differ from missing at random (MAR) predictions (i.e. a pattern mixture approach) and then estimation follows naturally by multiple imputations (MIs). He also summarized comments from Professors Andrea Rotnitsky and Joe Ibrahim on causal modelling ideas, inverse probability weighting methods, and local influence measures. The agreement is that sensitivity analyses should be accessible and relevant. Professor Jiayang Sun from Case Western Reserve University also provided a review on missing data and measurement error problems in longitudinal studies. She advocated alternative models and methods and considered mixed-effects selection and hybrid models, and presented two new non-Fourier density estimation procedures from data with measurement errors.

## 4 Outcome of the Meeting

All 42 participants attended the workshop. Among these 42 participants, there are 20 Canadians, 18 females, and 14 graduate students or junior researchers. The workshop also attracted a large number of well-known researchers. The workshop is a great success. Participants had extremely positive experiences with the workshop, and they were very satisfied with facilities, meals, and accommodation at the Banff Center. The workshop had a great impact on the graduate students and junior researchers with respect to their future career plannings. They find themselves greatly benefited from such a workshop. Senior researchers also find the workshop an excellent place for strengthening collaboration and communication. The organizers have received many very positive comments, e.g., “This is the best workshop I have ever been!”, “This is a wonderful workshop!”, “I really benefited a lot from the workshop”. The workshop provides an excellent opportunity for leading and young researchers in the field to discuss recent developments, emerging issues, and future directions in the analysis of longitudinal data.

One of the major goals of the workshop is to strengthen collaboration and communication among different research groups and consolidate existing ones. We have successfully achieved this goal. There were many interesting and active discussions throughout the 5-day workshop. Senior researchers offered their visions, suggestions, and guidance, and junior researchers learned many latest developments and exciting future research opportunities. Overall, the workshop is timely and provides a great platform for collaborative research and interactions between methodological and applied researchers. We find that BIRS is an ideal place for such communications, and we believe that we could not achieve the same results in a “classical” scientific meeting.

Finally, workshop participants have expressed great appreciation to BIRS and Banff Center staff members for the outstanding local arrangements and service. In particular, the workshop organizers would like to express their sincere appreciation to the BIRS Station Manager Brenda Williams and BIRS programme coordinator Wynne Fong for their extremely professional help. We understand that there is a large amount of work involved in the organization and local arrangements for the workshop. The wonderful BIRS staff team has made the workshop a very successful one!

## References

[1] H. Bang and J.M. Robins, Doubly robust estimation in missing data and causal inference models, *Biometrics* **61** (2005), 962–973.

[2] R.J. Carroll, D. Ruppert, L.A. Stefanski, and C. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edition, London: Chapman and Hall, 2006.

[3] R.J. Cook, G.Y. Yi, K.A. Lee, and D.D. Gladman, A conditional Markov model for clustered progressive multistate processes under incomplete observation, *Biometrics* **60** (2004), 436–443.

[4] N.R. Chaganty and H. Joe, Efficiency of the generalized estimating equations for binary response, *Journal of the Royal Statistical Society Series B* **66** (2004), 851–860.

[5] P. Diggle, P. Heagerty, K.Y. Liang, and S. Zeger, *Analysis of Longitudinal Data*, 2nd edition, Oxford, England: Oxford University Press, 2002.

[6] G. Fitzmaurice, M. Davidian, G. Molenberghs, and G. Verbeke, *Longitudinal Data Analysis*, Boca Raton, Florida: Chapman & Hall/CRC, 2008.

[7] H. Joe, Accuracy of Laplace approximation for discrete response mixed models, *Computational Statistics & Data Analysis* **52** (2008), 5066–5074.

[8] T.L. Lai and M.C. Shih, Nonparametric estimation in nonlinear mixed effects models, *Biometrika* **90** (2003), 1–13.

[9] T.L. Lai, M.C. Shih, and S.P.S. Wong, Flexible modeling via a hybrid estimation scheme in generalized mixed models for longitudinal data, *Biometrics* **62** (2006), 159–167.

[10] Y. Lee, J.A. Nelder, and Y. Pawitan, *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, London: Chapman & Hall/CRC, 2006.

[11] X. Lin and N.E. Breslow, Bias correction in generalized linear mixed models with multiple components of dispersion, *Journal of the American Statistical Association* **91** (1996), 1007–1016.

[12] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, 2nd ddition, New York: Wiley, 2002.

[13] C.E. McCulloch, S.R. Searle, and J.M. Neuhaus, *Generalized, Linear, and Mixed Models*, 2nd edition, New York: Wiley, 2008.

[14] G. Molenberghs and M.G. Kenward, *Missing Data in Clinical Studies*, Chichester, UK: Wiley, 2007.

[15] F.S. Nathoo and C.B. Dean, Spatial multi-state transitional models for longitudinal event data, *Biometrics* **64** (2008), 271–279.

[16] A. Qu, B.G. Lindsay, and B. Li, Improving generalized estimating equations using quadratic inference functions, *Biometrika* **87** (2000), 823–836.

[17] L. Wang, Estimation of nonlinear models with Berkson measurement errors, *The Annals of Statistics* **32** (2004), 2559–2579.

[18] L. Wu, *Mixed Effects Models For Complex Data*, Boca Raton, Florida: Chapman & Hall/CRC, 2009.

[19] G.Y. Yi and R.J. Cook, Marginal methods for incomplete longitudinal data arising in clusters, *Journal of the American Statistical Association* **97** (2002), 1071-01080.