Statistica Sinica **14**(2004), 615-629

DISCUSSION OF NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION

Summarized and contributed by: James S. Marron, University of North Carolina; Hans-Georg Müller, University of California at Davis; John Rice, University of California at Berkeley; Jane-Ling Wang, University of California at Davis; Naisyin Wang, Texas A&M University; Yuedong Wang, University of California at Santa Barbara.

Additional contributors at the conference: Brent Coull, Ludwig Fahrmeier, Wensheng Guo, Peter Hall, Jaroslav Harezlak, Jianhua Huang, Alois Kneip, Xihong Lin, Jeffrey Morris, Mohsen Pourmadi, John Staudenmayer, Colin Wu, Daowen Zhang, Heping Zhang and Yihua Zhao.

Longitudinal data and functional data are both data collected over a period of time on the same subject. They both depict the realization of a smooth underlying process at discrete time points. However, there are intrinsic differences between the two approaches, partly due to different sampling schemes. A comparison of the two perspectives and methods in functional data analysis and longitudinal data analysis is provided in the article by Rice (2004) in this special issue. The two fields have recently crossed paths due to challenges faced in each and this has led to the beginning of fruitful interactions. From the longitudinal data point of view, there is a need to pursue more flexible non- or semi-parametric frameworks that may better capture the complex data features that are present in many longitudinal studies. From the functional side, there is a need to provide techniques that work for "sparse" data commonly encountered in longitudinal studies. The following summarizes the discussions of two round-table discussions on this topic, which took place at Mt. Holyoke College in Summer 2002. Notes taken at each table were edited and additional input was solicited from several researchers.

Research topics that are important to both functional data analysis (FDA) and longitudinal data analysis (LDA) are highlighted. An area of application where both approaches have been used extensively is the study of growth curves or patterns. Exploring the impact of the functional viewpoint on established approaches in longitudinal data analysis was a main focus, and a good example of this is the growth curve study in Gasser et al. (1984), where a mid-growth spur was discovered for boys at around age seven through a nonparametric analysis of the derivatives of the growth curves. We believe that one domain which has attracted a certain level of attention but could still benefit from further study is the problem of *curve registration.* This is a well-established topic in FDA but has received much less attention from the LDA side.

Curve Registration

Taking the modeling of human growth again as an example, it is known that there is a common pattern shared by the growth velocities of different children. The growth velocity decreases sharply before and after birth until a certain age, after which there is a slightly increasing trend, the so-called mid-growth spurt, that is then followed by another decreasing trend that lasts until the onset of puberty. Even though this pattern is shared by all children, certain features and especially the timing of these features can vary from child to child. For example, the location of typical characteristic points, such as onset of puberty, can occur at different ages. Varying heights produce different growth amplitudes. Ignoring these differences among subjects could lead to inferior outcomes in data analysis or cause loss of information. Open research questions and problems include:

- **How to identify landmark features for registration and how to register?**
  Typically, there are both vertical and horizontal directions to be considered. The complexity of the analysis procedure increases and methods become unidentifiable when trying to account for both directions simultaneously without restrictions. How to identify the "typical" feature(s) from various curves as illustrated in Gasser and Kneip (1992) is another important issue.

- **How to perform inference after registration?**
  This problem is similar to that of how to analyze transformed data but it is more complex due to the more complicated nature of the registration problem. The transformation function considered in the transformation literature is usually monotone, and the problem can be viewed as an issue of scale-change. This is no longer the case here. How to account for the extra variation due to the estimated transformation in the main analysis warrants further investigation.

- **How to interpret the outcomes when the data analysis process involves registration**?
  Ideally, the advantage of using registered curves over using the original data is that the analysis could now focus on modeling or comparing the "features" that really matter. Interpretation of the outcomes should be specific to the field of application and may require further study.

How to Increase the Impact of Functional Viewpoint on Longitudinal Data Analysis?

A basic question is what should be emphasized when FDA researchers develop methods that would be viewed as useful and important in the LDA world of biological or medical research. In general, people seem to agree that there has to be a trade-off between simplicity versus flexibility when building a model and estimation procedures. Furthermore, efforts need to be put into explaining what has been developed. An understanding of what is needed in various fields of other disciplines could be rewarding. An issue that the FDA community needs to address much more in order to increase the relevance of FDA methodology is the case of highly irregular, sparse and missing data. Such data are prevalent in the field of LDA, but have not been a focus in FDA research. The approaches in Shi, Weiss and Taylor (1996), James, Sugar and Hastie (2000), Rice and Wu (2001) and recently in Yao, Müller and Wang (2003) provide some partial answers from the FDA viewpoint, but more attention is called for.

In addition to sparsity, longitudinal data in clinical trials or medical follow-up studies often involve missing data, and measurements are often not available after an event-time, such as death. The latter issue results in missing data that are informative, an issue that can be addressed by the "joint modeling" approach discussed in the next section. Briefly, this involves modeling the longitudinal data jointly with event-time data. Details can be found in two articles (Tsiatis and Davidian (2004) and Yu, Law, Taylor and Sandler (2004)) in this special issue. To address the general missing data issue, methods to handle missing data need to be developed for the functional approaches. This is largely unavailable in the literature.

One question raised was what are the open problems in the area of nonparametric regression. Powerful, flexible, and complicated models have been proposed in the literature. In smoothing splines, for example, the general smoothing spline regression models have been extended to smoothing spline ANOVA (SS ANOVA) regression models for multivariate functions, SS ANOVA regression models for correlated observations and data from exponential families, nonlinear nonparametric regression models, semi-parametric nonlinear regression models, and linear and nonlinear nonparametric mixed effects models. The research activities have emphasized methodologies for estimation while inferential methods have received relatively less attention. Bayesian and bootstrap confidence intervals are often used for inference on nonparametric functions. Care needs to be taken when interpreting these confidence intervals as to whether they have across-the-curve or pointwise properties. Hypothesis tests have been developed only for simple regression models. Inferential tools are important because one of the most useful aspects of the nonparametric methods is to check or suggest a parametric model.

When equipped with inferential tools, the above mentioned linear/nonlinear, nonparametric/semi-parametric, fixed/mixed models can be used to test common parametric models in LDA, such as nonlinear regression, linear/nonlinear mixed effects and generalized linear mixed effects models. Development of inference procedures with rigorous theory is essential in order for the nonparametric FDA approach to prevail in the biomedical community where longitudinal data routinely arise.

Another question raised was to what extent is the non- or semi-parametric approach to longitudinal non-Gaussian data a solved problem, particularly for longitudinal binary or categorical data. This is an area where more research is needed. The popular generalized linear mixed effects model has been extended to allow for both the fixed and random effects to be modeled nonparametrically (Lin and Zhang (1999) and Karcher and Wang (2002)). The estimation is challenging since the likelihood function does not have a closed form. Also, results can be sensitive to the distribution of random effects. The double penalized quasi-likelihood approximation of Lin and Zhang (1999) perform well for a good range of cases, but may lead to biases for sparse data, such as binary data in small clusters. Work is underway on MCMC approaches. Karcher and Wang (2002) used stochastic approximation with Markov chain Monte Carlo which guarantees convergence of the estimates to the expected fixed points. This approach is computationally intensive and its implementation is non-trivial. Thus, more research is necessary.

In the smoothing spline literature, fast ($O(n)$) algorithms exist only for special cases, and the computation of general splines is usually of the order $O(n^3)$. This computational burden (both speed and memory) limits the applicability of spline smoothing for the case of large data sets, particularly for smoothing spline based MCMC. This is less of a problem for other MCMC approaches. Comparison of several different MCMC methods is in order. There have been developments in Bayesian nonparametrics, but the area is challenging, both computationally and in theory, since Bayes procedures can be inconsistent in infinite dimensional settings and the Bernstein-von-Mises theorem may not hold. Interpretability is difficult for very high dimensional priors and it is difficult to assess the sensitivity of the results to the choice of the prior. Several authors proposed using a subset of knots (bases, representers) which leads to the P-spline literature (Ruppert, Wand and Carroll (2003)). More sophisticated methods have also been proposed: addition/deletion according to certain schemes and estimating the location using free-knot spline (DiMattea, Genovese and Kass (2001)). These more complicated methods can make the spline fit spatially adaptive, but at the expense of more computational time.

Besides basis function approaches, such as splines, local polynomials and kernel methods have been extended to longitudinal data (Lin and Carroll (2000) and

Wang (2003)). These methods have potentials for FDA implementations. These local smoothing methods gain their appeal from simplicity due to straightforward explicit representations which facilitates mathematical analysis and is a pre-requisite for many asymptotic results. These methods also allow straightforward extensions to quasi-likelihood models with binary or count responses, where link and variance functions are unknown. Extensions to generalized linear and quasi-likelihood models for functional data are of interest, including inference. Extensions to constrained estimation, such as monotonicity constraints for smoothing one-dimensional functions and symmetry and non-negative definiteness constraints for smoothing surfaces are also relevant for FDA.

Software development is an integral part of research. Advanced and powerful statistical methods are useful only when software is available. In particular, many programs are available for fitting various smoothing spline models. Old programs in Fortran are repackaged using more user friendly S language. Specifically, the S-Plus function smooth.spline fits cubic splines; FIELDS, a suite of S-Plus functions which can be downloaded from http://www.cgd.ucar.edu/stats/software.shtml, fits cubic and thin plate splines; smooth.Lspline, a S-Plus function which can be downloaded from ftp://ego.psych.mcgill.ca/pub/ramsay/Lspline, fits L-splines; `gss`, a suite of R functions which can be downloaded from cran.r-project.org/src/contrib/PACKAGES.html, fits general smoothing spline density, regression and hazard regression models; and ASSIST, a suite of S-Plus/R functions which can be downloaded from http://www.pstat.ucsb.edu/faculty/yuedong/research, fits many spline-based non-parametric/semi-parametric linear/non-linear fixed/mixed models. These are just a few links and by no means a complete listing. More software needs to be developed and enter mainstream software packages or libraries.

What Are the Interesting Open Problems?

- **Unifying theory for FDA and LDA:** Theoretical results are limited and most theoretical results in FDA assume that the entire curve is observed for each individual. This is unrealistic for longitudinal data which are often sparsely observed. Is it possible to develop a theory for this, and what are the mathematical tools needed? One complication is the non-invertibility property of many operators involved in FDA, such as the covariance operator. An encouraging aspect though is that often one does not need large sample size in terms of repetitions, but instead may rely on assumptions to allow "borrowing strength" from the data of other subjects.

- **Confidence bands for nonparametric estimates:** This is largely undeveloped, partly due to the lack of theoretical results in the FDA setting.

Bootstrap procedures are often used instead but there is no theoretical support to their validity, and their use in the FDA setting has not been systematically studied. It was noted that confidence bands with data driven smoothing parameters can be very slow to compute by Monte Carlo.

- **Bootstrap procedures for functional and longitudinal data:** How does bootstrap work for such data, and what are the theoretical justifications? We know very little about these. This topic deserves a lot more attention, especially because the asymptotic theory, even if it is available, is going to be complicated and difficult to be implemented in statistical inference. Thus, bootstrap procedures may well be a preferred approach to deal with inference.

- **Effective choice of smoothing parameters is a challenge:** This problem is hard enough for simple smoothing, and very challenging for longitudinal data. For instance, how should knots be selected for splines? Dense grids (e.g., with about 30 knots) may be an option for densely observed data, especially for linear splines. This is technically not consistent, but the consensus seems "small bias incurred is worth the simplicity". An alternative approach is to keep knots from getting too close to each other using "span restrictions". A new approach, hopefully useful in high dimensions, puts bounds on "maximal correlation". Adaptive knot choice and "hybrid splines" seem promising based on preliminary results. All of this requires further studies.

  Moreover, how should bandwidths be selected for kernel smoothers? Different bandwidths may be needed for different purposes, for example making inferences about covariate effects or smoothing individual curves. An adaptive choice is important, but the effects of adaptation must be taken into account in inference.

References

DiMattea, I., Genovese, C. R. and Kass, R. E. (2001). Bayesian curve fitting with free-knot splines. *Biometrika* **88**, 1055-1073.

Gasser, T. and Kneip, A. (1995). Search for structure in curve samples. *J. Amer. Statist. Assoc.* **90**, 1179-1188.

Gasser, T., Müller, H. G., Köhler, W., Molinari, L. and Prader, A. (1984). Nonparametric regression analysis of growth curves. *Ann. Statist.* **12**, 210-229.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587-602.

Kachar, P. and Wang, Y. (2002). Generalized nonparametric mixed effects models. *J. Comput. Graph. Statist.* **10**, 641-655.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Amer. Statist. Assoc.* **95**, 520-534.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *J. Roy. Statist. Soc. Ser. B* **61**, 381-400.

Rice, J. A. and Wu, C. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253-259.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge, New York.

Shi, M., Weiss, R. E. and Taylor, J. (1996). An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Appl. Statist.* **45**, 151-163.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90**, 42-58.

Yao, F., Müller, H. G. and Wang, J. L. (2003). Functional data analysis for sparse longitudinal data. Manuscript.

DISCUSSION OF JOINT MODELING LONGITUDINAL AND SURVIVAL DATA

Summarized and contributed by: Marie Davidian, North Carolina State University; Peter Diggle, Lancaster University; Dean Follmann, National Institute of Allergy and Infectious Diseases; Thomas A. Louis, Johns Hopkins University; Jeremy Taylor, University of Michigan; Scott Zeger, Johns Hopkins University.

Additional contributors at the conference: Laurel Beckett, Patrick Heagerty, Helene Jacquim-Gadda, Danyu Lin, John Rice, Jane-Ling Wang and Lee-Jen Wei.

The last 10 years has seen a number of articles on joint modelling of longitudinal and survival data. These articles are nicely reviewed in the papers by Tsiatis et al. and Yu et al. in this issue of Statistica Sinica. Here we will provide some general discussion and commentary, focusing on existing problems, challenges, open questions and future directions. These comments are based on the discussion which took place at the Mt Holyoke conference in 2002, together with solicitation of comments from various experts in the field.

The basic setup for a joint model is a study where repeated measurements are obtained along with survival data. The repeated measures are generally "internal" time-dependent covariates in the survival model, sometimes called biomarkers, i.e., they are generated internally by the subject and measure some aspect of the progression towards the event time, rather than being externally imposed. An example is HIV positive patients where it is common to collect serial CD4 counts as well as survival data. Another example concerns patients who undergo bone marrow transplantation where serial bilirubin levels are collected post-transplant along with the times of serious events.

A key question relates to the objectives of a joint model. There are many possible objectives, and how it is formulated will depend on the intended use. The most popular use of joint models has been to estimate the regression parameter in a time-dependent hazard model, where it provides a way to account for measurement error and infrequently measured values of the longitudinal variable. Another possible use is when the repeated measures are of primary interest and the event time is a cause of possible dependent censoring. However, there are other ways of analyzing longitudinal data with dependent drop-out, which might be appropriate to consider, especially if there are drop out mechanisms additional to the event time in the joint model. Current thinking suggests that it is important to undertake a sensitivity analysis when dependent drop-out is possible in the analysis of longitudinal data, although it is frequently not obvious how one does this. If one chooses to use a joint model and there are additional reasons for drop-out which might bias the main results of interest, then the joint model would need to be extended to include other drop-out mechanisms. In other applications both the longitudinal and the survival process may be of equal interest and a joint model with common parameters can result in more efficient inference than separate models. Other uses of joint models are to investigate whether the longitudinal variable might act as a surrogate endpoint, replacing the real survival endpoint, in a clinical trial, or whether it might be used as an auxiliary variable to assist in inference about the real endpoint. While joint models can be helpful to assess whether an early endpoint is a useful surrogate for the real endpoint in a particular completed study, they cannot address the bigger issue of whether the early endpoint will be a useful surrogate for the real endpoint in a future study with a different intervention. Another use of joint models could be for individual prediction of future longitudinal or survival data.

Up until now the majority of publications on joint models have focussed on a single Gaussian longitudinal variable with a time-dependent Cox model for the event time. There are obviously many ways in which this can be generalized, some of which have already been considered. For example, non-Gaussian or multivariate repeated measures, non-parametric longitudinal models, informative timing of events and recurrent events. One aspect which has been somewhat lacking in the literature has been methods for comparing models and assessing goodness-of-fit.

Maximum likelihood, Bayesian and other methods of parameter estimation have been developed. The fully Bayesian and likelihood methods tend to be computationally complex, which calls for easy to use software to make the methods accessible. A number of approximate methods and methods which focus on selected aspects of the model have been suggested. They are typically much less computationally intensive, however, their appropriateness will depend on the goals and context of the application. A popular simple method is first to do an analysis of the longitudinal data to get empirical Bayes estimates of the biomarker, which are then used as predictors in the survival model with adjustment for bias and variance estimation. While such an approach has been shown to be inferior statistically to joint modelling, it does offer a lot of flexibility and it may be adequate in many applications, or useful for preliminary model selection steps.

A major issue in joint models is model formulation. This is likely to be context specific, so it is hard to make any general recommendations. Another important aspect, is what is the question being asked. Some possible questions are described above, and how the model is formulated will depend crucially on this. The most common formulation is as a random effects model for the repeated measures data and a time-dependent Cox model for the hazard. A conceptual issue arises here, as in some sense, the repeated measures only make sense for patients who are alive. Thus perhaps the repeated measures model can only be viewed conditional on being alive. A desirable feature of a model would be that any joint model reduces to sensible marginal models for each of the longitudinal variable and the survival variable. This is obviously not the case for the usual survival model, although the marginal survival model can be obtained via integration. The complexity of the model might also depend on the quality of the data and what one believes about the underlying measurements. For example, one might postulate a stochastic process for the underlying longitudinal data instead of the usual random effects model. However, one would need a lot of data and a good rationale for thinking this might make a difference before pursuing this approach. In general the impact of misspecification of the longitudinal model on inferences from the joint model has not been explored enough. In the majority of applications the relationship between the longitudinal process and the hazard has been through the current value of the longitudinal process. Some analysts may have incorrectly inferred that this must be the case. It would be interesting to see more work done in contexts where other aspects of the longitudinal model, such as the slope or past history, are needed in the survival model.

In many applications of joint models, the repeated measurement variable has been something very central to the disease, such as CD4 counts and viral load in AIDS. Often in such cases there is a mechanistic flavor to the model, and this frequently lends to a causal interpretation to the results. Strictly speaking joint models are only capable of assessing associations, so causal interpretations must be undertaken with caution. One challenge is to clarify in the literature exactly what inferences are possible (or not) with these models. Frequently, in the absence of a mechanistic understanding of the underlying disease progression, the models used for each component are relatively simple and chosen for convenience.

They are designed to provide an empirical representation of the observed data, by distilling them down into a few prominent features. For example, random effects are frequently chosen for the longitudinal data. While such a model is not likely to be an accurate description of how the time-dependent variable develops over time, it is quite possible that it is adequate for some goals of the analysis. This will depend on the question being asked. If the goal is just assessing association, as represented by a regression parameter in a Cox model, then it may be adequate. Whereas if the goal is using the joint model to develop subject-specific predictions then it may be important to get an accurate representation of the stochastic development of the repeated measures variable.

Lastly, it is interesting to note a slight misnomer. One of the contributors to this summary indicated that we should use the term "longitudinal data" for data that arise in longitudinal (aka cohort) studies. These comprise both what we now mean by longitudinal data, namely repeated measurements and time to events.

DISCUSSION OF CAUSAL INFERENCE: WHAT AND HOW?

Summarized by: Els Goetghebeur, University of Ghent and Rod Little, University of Michigan.

Additional contributors at the conference: Steve Cole, Phil Dawid, Ming Ji, Paul Rathous, James Robins, Butch Tsiatis, Ravi Varadhan and Wei Wu.

We discussed two points, one fundamental and philosophical, the other much more practical and hence equally (or more) important.

1. (a) Can meaningful, non-ambiguous counterfactuals be defined to help describe well understood causal effects that are not (directly) observable? Or does everything stay possible in the science fiction world of counterfactuals?
   (b) Where counterfactuals can be useful? Can they still be avoided?
2. What can be done at the design stage to ultimately justify (believe) causal inference and its assumptions ?

1. Counterfactuals = Science Fiction?

We start by debating the useful existence of counterfactuals in the mind of practicioners and/or statisticians seeking to understand causal efects. Some (e.g., Dawid (2000)) argue that one can claim anything about 'what would happen if we intervened to change exposure' as long as no such specific (randomized) intervention is performed or can be designed. Such claims must therefore be void of meaning or arbitrary: while causal effects exist, counterfactuals do not.

Others (e.g., Robins and Greenland (2000)) find counterfactuals useful, indeed sometimes indispensable to communicate causal effects clearly. They are observing a world of selective exposures and make claims about 'what would happen if we intervened to change exposure' under stated assumptions. The following example seeks to illustrate that unambiguous meaning can be given even when the intervention is infeasible.

Example. Imagine a respiratory machine designed to improve lung function by week 6 through a specific biological mechanism. When using the test in a randomized study, unfortunately more deaths occur prior to week 6 in the treated group through a mechanical failure (say) of the machine that knocks people down periodically (the hammer effect). This mechanical side effect operates independently of the intended biological action of the machine but could be triggered by some characteristic of the patient (e.g., shivering). The question arises: what would have been the (average) values of lung function at week 6 in both groups had the hammer effect not existed/been removed and all other things stayed equal? This is a question cast in terms of counterfactuals with a well-understood and relevant meaning concerning the biological action of the machine.

Having agreed that such a question is meaningful, where/how/when can we find a reliable answer based on observed data?

1. When a small percentage of patients dies on each arm, say 2% on the treatment arm and 1% on the control arm, it is clear that useful bounds can be derived on the average causal effect on lung function at week 6. At some point the percentages of deaths will however get too large to allow for useful bounds.

2. Conditioning on baseline covariates, which predict a small chance of death, then allows the argument above to be exploited for causal inference restricted to this observable subset of the study population.

3. If the deaths 'by hammer' are unavoidable, it is useful to ask about the (expected) causal effect on lung function at week 6 in the subpopulation that would survive week 6 under either treatment assignment. This concerns an unobserved (but factual) stratum of the study population. To estimate such stratum-specific difference in average lung function between arms, additional assumptions must be made. For instance, the strong rank

preserving failure time assumption, which states that patients are on the same quantile position in the arm-specific survival distribution on either randomized arm.

Many other untestable, identifying assumptions are possible which must be argued or subjected to a sensitivity analysis. To have generalizable results, subject matter knowledge must help decide whether we are estimaring causal effects.

The discussion topic above is not new. It relates to the classical competing risks survival problem and runs parallel to the small pox vaccination problem addressed by Daniel Bernouilli in the 18th century. He tried to assess how good blood sucking leeches would be as a treatment for small pox if their own direct operational mortality risk could somehow be removed. Novelty lies however in structural models for ‘counterfactuals’ or ‘potential outcomes’ which have enabled a large body of recent theoretical results. Our next question relates to their implementation.

2. Observational Data and Time-Dependent Exposures

One approach to causal inference relies on the assumption of ‘sequential randomization’ or ‘no residual confounders’. Its justification in any practical data set depends crucially on the nature and number of recorded time-dependent covariates. Two questions are raised:

- What can be done to ensure the necessary data get collected? Often this is an expensive and hence unpopular proposition.

  Scientists have the duty to confront plausible biases and protect against them. It may be tempting however to ignore confounders (less time and effort, less cost, a higher chance of surprising results and hence of publication...) and indeed there seldom exist guarantees that we have covered them all. Today’s omni-presence of randomized clinical trials owes a lot to legislation that came via the FDA. Legislation on the measurement of potential confounders is however a complicated goal.

- One avenue towards high quality predictors of exposure in clinical trials is the design of a run-in period (on controls). Given the collective benefit causal inference should bring to future patient generations and the lack of harm done to the current patient population, we find such efforts carry the ethical benefit.

- How to avoid unstable causal inference due to data mining when a high dimensional predictor space allows for many different regression models to be fit? Here the ethical data-analyst will convince him/herself and others of the value of the analysis by due exploration of the major threats of instability.

In conclusion the issues tackled are of fundamental importance to our discipline and indeed science as a whole. They deserve to be studied and discussed more broadly.

References

Dawid, A. P. (2000). Causal inference without counterfactuals, *J. Amer. Statist. Asssoc.* **95**, 407-424.

Robins, J. M. and Greenland, S. (2000). Causal inference without counterfactuals – Comment. *J. Amer. Statist. Assoc.* **95**, 431-435.

DISCUSSION OF TWO IMPORTANT MISSING DATA ISSUES

Summarized by: Raymond J. Carroll, Texas A&M University.

Additional contributors at the conference: Marie Davidian, Joel Dubin, Garrett Fitzmaurice, Mike Kenward, Geert Mohlenberghs and Jason Roy.

1. Introduction

We comment on two important issues involving missing data in longitudinal data:

- The unsuitability of Last Observation Carried Forward (LOCF) as a missing data imputation scheme in longitudinal data and the need for regulatory agencies, e.g. the U.S. Food and Drug Administration (FDA), to recognize this unsuitability.
- The emerging difficult issue of sensitivity analyses for longitudinal studies in which data are missing not at random, and its interface with recent International Conference on Harmonization (ICH) guidelines for such sensitivity analyses.

2. Last Observation Carried Forward

Missing data, and in particular dropouts, are ubiquitous in longitudinal studies, and for the most part the missingness is not missing completely at random (MCAR). For this reason, it is well–known that such naive data imputation methods as completers analysis (use only those who do not drop out) lead to tests for treatment effects that have elevated levels and are at the same time lacking in statistical power.

In regulatory settings, it has become popular to use the last observation carried forward (LOCF) method to handling dropouts and missing data. This method imputes missing response data by using the most recently observed responses. Thus, in a 6–week study if a patient drops out after 3 weeks, the responses for weeks 4–6 are imputed to be the same as the response at the third week.

Our conclusion is that LOCF should (almost) never be used as the primary means of handling missing data in longitudinal studies.

LOCF is seemingly a more sophisticated means of handling missing data than is a completers analysis, but this sophistication is illusory. The method is only valid if the data are MCAR, and it leads to incorrect treatment test levels in the almost universal situation that missing data are either Missing at Random (MAR) or informative, i.e., Missing Not at Random (MNAR). The ethics of using a method for making decisions about human health when that method is well–known to be often seriously invalid in most settings seems to have been ignored.

Arguments have been made that LOCF is clinically relevant, because it involves describing patient outcomes up to the point that patients dropout. This viewpoint is forced, to say the least, and makes little statistical sense, because it is not at all clear what a LOCF analysis is actually making inference about when two drugs have different dropout rates.

Alternatives to LOCF are available, that, unlike LOCF, are supported by formal, underlying theory. MAR analyses are well–known, e.g., repeated measures analyses under likelihood assumptions, Horvitz–Thompson inverse weighting schemes, multiple imputation, etc. The simplest method for continuous data, repeated measures in the mixed model framework, has been available for many years, is part of the standard curriculum in graduate school for biostatisticians, and is available in common statistical packages such as SAS, Splus, SPSS, etc.

Our suggestion is that MAR methods be the base for the analysis of missing data in longitudinal studies, in particular repeated measures analyses for continuous data.

It is disappointing that regulatory agencies cling to LOCF, instead of using the many advances in missing data analyses that have been developed over the last 20 years.

3. Sensitivity Analyses

Notwithstanding our earlier comments, one of the major research problems in missing data for longitudinal studies is how to handle data that have informative dropouts, i.e., are MNAR. The topic has increased impetus because of recently proposed ICH guidelines that dictate a sensitivity analysis as part of any MNAR methodology.

A few things are clear in this area.

1. Handling informative missingness inevitably requires making assumptions that cannot be verified from data. Thus, sensitivity analyses must be done in conjunction with more formalized inferences. It is not enough to build a model that fits the observed data. While this is a necessary condition, it is not even close to sufficient because of the inherent nonidentifiability.

2. The emphasis should be on inference directed at the specific scientific context. In particular, sensitivity analysis should not be directed exclusively at point estimates, but should instead focus on key issues of the inferential decision process, including significance levels and interval estimation.

3. One should avoid the "*super–model fallacy*", i.e., the tendency to posit a complex highly flexible model incorporating informative missingness and then declaring that the model gives the correct answer. The lack of identifiability inherent in the problem dictates against the hubris of a "my method is the best" claim.

4. Sensitivity analyses should not be prescribed or proscribed. The ICH guidelines will inevitably lead to a desire to write down a single approach for each of a list of problems. We view this as both absurd and harmful. The area of sensitivity analyses for informatively missing data is in its infancy, and the danger is that old, not very good and often invalid methods will be enshrined (a good example of such fossilization is LOCF, see above).

5. The methods should be transparent, so that their assumptions can be understood and criticized.

6. One of the outstanding technical problems is to be able to simulate data that give reasonable extreme situations. It seems to us that "worst case" methodology is extremely limited in its applicability.