RESEARCH STATEMENT

Fei Jiang (jiang@hsph.harvard.edu)

In the biostatistical research, investigators often face three challenges: designing efficient and ethical clinical trials for new treatment evaluations; analyzing data with complex structures from large observation studies; providing robust estimation and inference procedures for analyzing the data from randomized clinical trials. Because of the increasing demand on new treatment, the growing complexity of data structures and randomization trial procedures, traditional methods are no longer sufficient to provide satisfactory results. To this end, I focus on developing novel adaptive clinical trial designs, robust semi-parametric approaches, and valid estimation and inference procedures in randomized clinical trials to address these issues.

Specifically, I developed utility-based clinical trial designs, which resulted in powerful trials with low sample size requirements. Further, I proposed robust semi-parametric methods which were capable of handling functional data from general observational studies, especially in the contexts of longitudinal and survival data analysis. Moreover, I identified the lack of robustness problem in the classical methods and introduced more accurate and efficient estimation and inference procedures for the analysis in randomized clinical trials.

Below, I describe in detail on my selected research topics and future research plans. Overall, my research projects were initiated by existing clinical problems and real data examples in practice.

Adaptive Clinical Trial Design

The Bayesian decision-theoretic method for clinical trial design (Jiang et al., 2013).

In clinical trial studies, investigators are naturally interested in knowing what are the expected losses (utilities) if a trial deviated from its target because of the wrong decisions made during the process. To resolve such an utility-based decision making issue, the Bayesian-decision theoretic approach has long been advocated. Although intuitively appealing, the approach has not been popularized in clinical research. The phenomenon is due to the fact that FDA's judgement only depends on the frequentist operating characteristics of the trial, while the regulation based on the expected losses has not been well established.

To balance the practical needs and FDA requirements, we proposed a novel Bayesian decision-theoretic phase II clinical trial design with sequential stopping to evaluate treatment efficacy. The design not only directly utilized the expected losses for the decision making but also achieved favorable frequentist operating characteristics. Furthermore, the design assigned more patients to better treatment arms by applying adaptive randomization procedures. To support the practical uses of the designs, we developed an efficient constrained backward induction and forward simulation algorithm to implement the design and a simulation routine to perform the frequentist inference procedures.

Randomization adapted to continuous and discrete covariates in clinical trials (Jiang et al., 2015b).

Covariate balance among different treatment arms is critical to clinical trials, as known and unknown confounding effects can be eliminated when patients in different arms are alike. To conduct covariate adaptive randomization when continuous covariates are considered, the current practice is to categorize the continuous covariates into several groups. However, there is no standard rule for the discretization, and the different ways of categorizing may lead to different imbalances as well as different covariate adaptive randomization procedures for patients.

To overcome this difficulty, we proposed a new kernel based dynamic allocation scheme to assign each patient to a treatment arm, such that the prognostic factors can be balanced across different arms. Our approach did not need to discretize continuous covariates into multiple categories, and it could handle both the continuous and discrete covariates naturally. The new design was theoretically grounded and easy to implement in practice.

Semi-parametric Approach for Longitudinal Data Analysis

Fused kernel-spline smoothing for repeatedly measured outcomes in a generalized partially linear model with functional single index (Jiang et al., 2015a).

In the cooperative huntingtons observational research trial (COHORT) study, patients cognitive scores, and disease indicators were collected repeatedly during the follow-up period in aiming of finding reliable prodromes to enable early detection of Huntington's disease (HD). Historical research showed that the joint effect of the cognitive scores on the odds of HD diagnosis changes with time. In addition, the relationship between the cognitive symptoms and the log-odds of the disease diagnosis was shown to be nonlinear (Paulsen et al., 2008). Our goal is to study the nonlinear time dependent cognitive effects so as to facilitate the early detection of HD.

Motivated from the COHORT data set, we proposed a generalized partially linear functional single index risk score model for repeatedly measured outcomes, which accounts for the time dependency of the cognitive effects as well as the nonlinear structure of the cognitive effects through a fused kernel spline structure. The study provided a sophisticated device to model the complex functional data in practice.

Semi-parametric Approach for Survival Analysis

A semi-parametric method for survival analysis, with application to an AIDS clinical trial study (Jiang et al., 2014).

In the AIDS A5175 clinical study, patients' time to adverse outcomes and their CD4 counts were measured and analyzed in order to establish the relationship between the time-to-event response and the covariate: CD4 counts. To model such relationship, classical approaches merely focused on describing the mean structure between the response and the covariate. However, initial analysis on the A5175 data described a clear second order dependency of the time-to-event endpoint on the CD4 counts. In this case, correctly modeling the variance structures would improve the estimation efficiency as well as provide additional information on the distribution of the outcomes (Wang and Leblanc, 2008).

This particular data structure motivated our research on applying the second order semi-parametric efficient method in the survival analysis, where the dependence of the mean and variance of the time on the covariates were specified through a restricted moment model. To handle the censored data, we combined a second-order score with a non-parametric imputation device for estimation. The imputation method provided a general strategy of extending the efficient score method that was developed for complete data in survival analysis. The proposed efficient and robust semi-parametric method could be widely applied under various survival settings.

A semi-parametric transformation frailty model for semi-competing risks survival data (Jiang et al., 2015c).

In biomedical studies, a patient may experience multiple types of failure. One specific setting is where interest lies primarily with some so-called *non-terminal* event, the observation of which is subject to some *terminal* event. A patient experienced a non-terminal event does not preclude the patient from the terminal event. In contrast, a terminal event precludes the patient from subsequently experiencing the non-terminal event. In the current practice, the two endpoints were either analyzed separately or combined into a single endpoint. However, these methods failed to consider the intrinsic competing relationship between the two endpoints, and thus lost the statistical and clinical meaningful information. This drawback gave rise to a branch of statistical methods, namely, the *semi-competing risk data analysis*, which focused on effectively modeling the non-terminal and terminal event.

In this area, we proposed a novel class of transformation models that permit the non-parametric specification of the frailty distribution under the semi-competing risk setting. The semi-parametric method was robust to the misspecification of frailty distributions. Furthermore the proposed semi-parametric transformation model were broadly applicable to any analysis of multivariate time-to-event outcomes in which a unit-specific shared frailty was used to account for correlation.

Estimation and Inference in Randomized Clinical Trials

Estimating the treatment difference with data from a comparative randomized clinical trial in presence of potential baseline-covariate imbalance (Jiang et al., 2015d).

In comparing two treatment groups, when the patient's potentially predictive baseline covariates vector are available, one may utilize an analysis of covariance procedure to obtain an efficient estimator for the parameter of interest. It is well-known, however, that the resulting estimator may not be consistent, when the ANCOVA model is nonlinear (e.g., a logistic or proportional hazard model) and not correctly specified (Owen and Froman, 1998).

In this case, a model free augmentation estimation procedure with covariate adjustment produced a consistent estimator for the parameter of interest, while the bias and standard error of the augmented estimator was markedly smaller than the simple two-sample estimator. We implemented the method on the data from a comparative clinical trial for evaluating two treatments on treating HIV infected patients. Overall, the augmentation method provided a less biased and more efficient way of estimating treatment effects in the presence of potential covariate imbalance.

Future research plan.

Short term plan

I am currently engaging in several projects as a leading researcher. These projects are in the early stage of the development, and could serve as my short term research direction. I provide a brief description for each project as follows.

- **Locally efficient semi-parametric Cox frailty model.** A kernel based locally efficient method was proposed for the estimation in the Cox proportional model when the frailty distribution is unspecified.
- **The heat wave effect on the hospitalization rate.** A semi-parametric approach was used to explore the heat wave effect on the patients' hospitalization rate for different diseases, which demonstrates both spatial correlations and lag-effect in the time domain.
- **Joint modeling of semi-competing risk data and longitudinal covariate.** A model was proposed to analyze the semi-competing risk data and longitudinal covariate jointly through introducing a latent process with unspecified distributions.
- **Stratified analysis, do we need this?** A long misunderstanding on the stratified analysis was identified by demonstrating that a weighted average of stratum-specific odds ratio does not provide valid inference on the true odds ratio in the target population. A simple averaging method was advocated to estimate the the underlying odds ratio of interest.

Long term plan

Statistical methodology development. In my future study, I will continue focusing on the semi-parametric methods and their application in the biostatistical area. Specifically, I will emphasize the joint use of kernel and spline regression to enhance the model robustness. This technique allows estimating nonlinear covariates effect characterized by time or spatial dependent coefficients. It renders great flexibility to analyze data of complex structure.

In parallel, for the clinical trial design development, I will extend my focus on single adaptive randomization procedure to the combinations of different adaptive randomization processes. One particular interesting topic is the combination of covariate and response adaptive randomization trial. This design could achieve the efficiency and ethics goals simultaneously. However, because of the dependency induced by the complicated randomization procedure, the asymptotic property of the estimator for the treatment effect is generally a challenge for the researchers, which prevents the popular use of the combined procedure in practice. I am planning to use the martingale technique, asymptotic analysis to tackle this difficult but important problem.

In the meanwhile, survival analysis will continue to be my focus in the long term, especially in the area of semi-competing risk survival data analysis. One potential research topic is to consider different censoring mechanisms, such as left truncation, interval censoring. Another direction is to consider different covariates structures, such as time dependent covariate, the covariate measured with error, etc. The semi-competing risk survival data analysis is a relative novel research area which offers many challenges as well as opportunities.

Collaborations. The biostatistical research often involves active collaborations with researchers from various heath areas. In my experience, I have collaborated with environmental scientist Dr. Loretta J. Mickley and her research group from Harvard School of Engineering and Applied Sciences to investigate the heat wave effect on diseases' onset. Moreover, I have worked with Dr. Rebecca Li and her research team from Harvard Global Health Institute to develop multi-regional clinical trials across countries. In the future, I will continue to collaborate with medical researchers from diverse backgrounds. This collaborations will grant me practical resource to initiate statistical research. And, I hope my professional statistical contribution could push medical research forward.

Moreover, the collaborations within the statistical field is crucial to conduct effective statistical research. My current collaborators include the professors from Harvard University, Stanford University, Columbia University, Rice University, The University of M.D. Anderson Cancer Center, The University of South Carolina, and Hong Kong University. I will continue to extend my connections to gather the valuable intellectual resources for my future statistical research.

References

Jiang, F., Lee, J. J., and Müller, P. (2013), "A Bayesian decision-theoretic sequential response-adaptive randomization design," *Statistics in Medicine*, 32, 1975–1994.

Jiang, F., Ma, Y., and Lee, J. J. (2014), "A semi-parametric method for survival analysis, with application to an AIDS clinical trial study," Under the second round revision in *The Annals of Applied Statistics*.

Jiang, F., Ma, Y., and Wang, Y. (2015a), "Fused kernel-spline smoothing for repeatedly measured outcomes in a generalized partially linear model with functional single index," *The Annals of Statistics*, 0, 0.

Jiang, F., Ma, Y., and Yin, G. (2015b), "Randomization adapted to continuous and discrete covariates in clinical trials," Submitted to *Journal of the American Statistical Association*.

— (2015c), "A semi-parametric transformation frailty model for semi-competing risks survival data." Submitted to *Journal of the American Statistical Association*.

Jiang, F., Tian, L., and Wei, L.-J. (2015d), "Estimating the treatment difference with data from a comparative randomized clinical trial in presence of potential baseline-covariate imbalance," Under preparation.

Owen, S. V. and Froman, R. D. (1998), "Uses and abuses of the analysis of covariance," *Research in Nursing & Health*, 21, 557–562.

Paulsen, J. S., Langbehn, D. R., Stout, J. C., Aylward, E., Ross, C. A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L. J., Duff, K., Kayson, E., Biglan, K., Shoulson, I., Oakes, D., and Hayden, M. (2008), "Detection of Huntingtons disease decades before diagnosis: the Predict-HD study," *Journal of Neurology, Neurosurgery & Psychiatry*, 79, 874–880.

Wang, L. and Leblanc, A. (2008), "Second-order nonlinear least squares estimation," *The Annals of the Institute of Statistical Mathematics*, 60, 883–900.