1 Background

1.1 Literature Review

1.1.1 General Background of Joint Modeling

Longitudinal studies are ubiquitous in biostatistics context. For example, in a randomized clinical trials (RCT) where patients are randomly allocated into different treatment arms and are then followed over time to collect some outcome of interest and risk factors. Repeated measurements will then be produced from this follow-up mechanism. One of the important features of longitudinal data is that the repeated measurements from the same subject are more "similar" to each other compared to those measures from different subjects, i.e. within subject measures tend to be intercorrelated. This feature requires special statistical techniques to handle the correlation thus valid scientific inference can be drawn from the data. As discussed in Diggle et al. (2002), there are mainly three methods we can use to analyze longitudinal data: marginal model, transition model and random effects model. Estimates of the regression coefficients from different models can be interpreted differently and the choice of a model depends on study objectives, the source of correlation as well as the capacity of the model. In this thesis work we will focus on applying random effects model to longitudinal data. A model that contains both random effects and fixed effects is called mixed effects model. The mixed effects model methodology, first introduced by R.A. Fisher (Fisher, 1919), is a statistical tool that is used across a wide variety of disciplines including biostatistical contexts. Mixed effects models are especially popular in researches involving repeated measurements or observations from multilevel (or hierarchical) structure where the correlation between observations is not negligible as discussed above.

In many clinical trails and medical studies, time-to-event (survival) data is commonly generated along with the longitudinal measurements. Often times, the outcome of interest in survival data, for example disease recurrence, possible drop-outs, or death, is correlated with the longitudinal measurements. Simply ignoring and the correlation and fitting two models separately will lead to lost of information and misleading results. Joint model (JM) method for longitudinal and survival data was first proposed by Tsiatis et al. (1995) and Faucett and Thomas (1996) to handle this issue and to obtain unbiased estimators. JM is well studied in recent years, for example, Henderson et al. (2000), Wang and Taylor (2001), and Xu and Zeger (2001). In terms of Bayesian methodology, Guo and Carlin (2004) developed a fully Bayesian method to fit the JM using MCMC methods and implemented them in WinBUGS software. For more detail, see Yu et al. (2004) for a good review of JM topic. Many extensions have also been developed for JM, including considering multiple longitudinal outcomes (Brown et al., 2005; Rizopoulos and Ghosh, 2011), incorporating multiple failure times (Elashoff et al., 2008), and so on.

1.1.2 Longitudinal Quantile Regression in the Setting of Joint Modeling

In most of the JM related works, the longitudinal part is modeled using linear mixed model (LMM), which is a widely used application of mixed effects methods. In brief, a LMM assumes the expected value of the outcome is a linear function of a set of covariates and observations from the same subject share the same unobserved latent variable, i.e. random effect, to account for the correlation among those observations. When conditional on random effects, observations from the same subject are treated as independent. In addition, traditional LMM also assumes the unobserved random error follows normal distribution.

Our concern for the widely used LMM is that in many circumstance the normality assumption of the error term can not always be satisfied, even after trying various ways of transformation. A commonly encountered situation is when there exits outliers or the outcome is skewed, where LMM is not appropriate to use. In other cases, the conditional mean may not be the primary interest and researchers may also be interested the covariate effects on the lower/upper quantiles of the outcome. *For example.* Instead of trying to fix the limitations of LMM, quantile regression is another method that we may consider since it can meet all above needs directly. There are several advantages of quantile regression over the ubiquitous mean regression (a.n.a.

linear regression) model. To list a few, quantile regression provides a much more comprehensive and focused insight into the association between the variables by studying the conditional quantiles functions of the outcome, which may not be observed by looking only at conditional mean of the outcome (Koenker, 2005). In quantile regression, the regression coefficients ($\boldsymbol{\beta}$) are functions of the quantile (τ), and the estimations of $\boldsymbol{\beta}$ vary according to different quantiles. Thus quartile regression provides a way to studying the heterogeneity of the outcome that is associated with the covariates (Koenker, 2005). Moreover, as mentioned above, quantile regression is more robust against outliers in the outcome compared with the mean regression, which is an immediate extension from the property of quantiles.

Quantile regression is becoming more and more popular in the statistical community. Koenker and Bassett Jr (1978) introduced a method in estimating the conditional quantiles. As an introductory material, Koenker and Hallock (2001) briefly covers the fundamentals of quantile regression, parameter estimation techniques, inference, asymptotic theory, etc., his book provides more comprehensive and deeper introduction on quantile regression related topics Koenker (2005). Yu and Moyeed (2001) introduced the idea of Bayesian quantile regression by modeling the error term using asymmetric Laplace distribution (ALD) followed the idea proposed in Koenker and Bassett Jr (1978). Kozumi and Kobayashi (2011) developed a Gibbs sampling algorithm for Bayesian quantile regression models, in which they used a location-scale mixture representation of the ALD. Many works have been done to extend the quantile regression method to accommodate longitudinal data. Jung (1996) developed a quansi-likehood method for median regression model for longitudinal data. Geraci and Bottai (2007) proposed to fit the quantile regression for longitudinal data based on ALD and the estimation is made by using Monte Carlo EM algorithm. Later on, Liu and Bottai (2009) followed the idea of Geraci and Bottai (2007) and extended the model from random intercept to including random slope as well. The study of longitudinal data using quantile regression is becoming popular in recent years. Fu and Wang (2012) proposed a working correlation model for quantile regression for longitudinal data, a induced smoothing method was used to make the inference of the estimators. Fully Bayesian techniques and Gibbs sampling algorithm become possible when the error term is decomposed as the mixture of normal and exponential random variables for the quantile linear mixed model, see Kozumi and Kobayashi (2011) and Luo et al. (2012) for applications. The fully Bayesian method is appealing to us because it is easy to implement and to make inference, the uncertainty of the unknowns is taken into account, and it is flexible in the distribution of random effects. The detailed background about Bayesian quantile linear mixed model will be provided in Section **??**.

1.1.3 Subject Specific Dynamic Predictions Based on Joint Modeling

In recent years, another extension of the JM that attracts increasing attention is to make subject specific predictions for longitudinal or survival outcome based on historical information at hand. In clinical setting, as we monitor the health features of a patient over time, later on those time-varying measurements can be used to derive other useful summary indicators, such as probability of event or recurrence. Thus JM provides a vital tool in predicting future health outcomes for the patients at risk.

Among the literatures, there are several applications of this prediction idea. Yu et al. (2008) used a JM framework to study the longitudinal measures of prostate-specific antigen (PSA) in predicting the probabilities of recurrence of prostate cancer up to four year in the future. In their work, the longitudinal part was modeled using nonlinear hierarchical mixed model and Cox proportional hazards model with time-dependent covariates was used to model the time to clinical recurrences. Both the value of longitudinal outcome and probability of recurrence were predicted. Proust-Lima and Taylor (2009) also worked on the PSA and prostate cancer problem and they used the joint latent class model (JLCM) (Lin et al., 2002) to build a dynamic prognostic tool in predicting the recurrence of prostate cancer, in which they used the maximum likelihood estimate method. As another example, Rizopoulos (2011) illustrated in detail about how to make survival probabilities prediction under the JM framework using Bayesian method. The application is demonstrated using the famous AIDS CD4 cell count data. Another important aspect for all of the prediction works is how to validate the accuracy of the predicted results, which is discussed in all of above

mentioned studies. There are different aspects that we can refer to assess the predictive performance of the model. Rizopoulos (2011) used the receiver operating characteristic (ROC) curve based approach to assess how well the proposed model can differentiate patients who will have events from those who will not. While Proust-Lima and Taylor (2009) computed the absolute error of prediction (EP) curves showing weighted average absolute error of prediction (WAEP) over three years and the EP at one- and three- year horizons. And in Taylor et al. (2013), a simple graphical method was used to assess the predicative accuracy. More technique details about the JM estimation, prediction and results validation will be given in Section **??**.

1.2 Public Health Significance

The significance of our work to the public health area can be summarized into two aspects. First, despite of its popularity, the conditional mean of the outcome is not always the statistics of primary interest in real world studies; in stead, in many cases low or high tail of the outcome variable can be of interest to the researchers. Among others, examples of preference of studying conditional quantiles over the conditional mean of the outcome include cardiovascular studies that focus on the effect of intervention to reduce the high blood pressure for the upper tail patients who are at greater risk; study of low birth weight of infants (Koenker and Hallock, 2001) and HIV studies, where subjects in left tail are more susceptible. Thus our method provides a necessary complement to the traditional linear regression method whenever the conditional quantile if of greater interest in the study.

Second, "accurate" subject specific predictions of the probability of event occurrences is of great importance in clinical setting. For example, people with hypertension is at high risk of having stroke and among those who had experienced stroke and survived, there is also chances of stroke recurrences. Thus, the predicted probability of events in the future can be used as an important reference to assist the physicians to make decisions for additional prevention and intervention choices to current ones in order to achieve better clinical results.

1.3 Research Questions and Specific Aims

The joint model of longitudinal data, using quantile regression, and survival type of data is little studied so far. To our knowledge, Farcomeni and Viviani (2014) is the first work that extended classical JM to incorporate quantile regression model in the longitudinal process. In their paper the parameter estimations is obtained using Monte Carlo expectation and maximization (MCEM) method. Also, there is no work has been done yet to extend the subject specific dynamic predictions method to the the longitudinal quantile regression based JM framework. We aim to extend current research works and fill those gaps.

Specifically, upon finishing the thesis work we aim to achieve the following aims:

- First of all we will develop a fully Bayesian method in estimating the model parameters in the proposed JM, based on which a subject specific dynamic predictions method for survival events will be developed;
- Second, instead of using survival data we will extend our new Bayesian JM method to study the recurrent events data and develop a new method for estimating longitudinal and recurrent event data model that uses quantile regression based JM;
- At last, method for subject specific dynamic predictions of recurrent events will be developed based our results from first and second projects.

To demonstrate the application of our proposed methods, we will use the data from The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) (Davis et al., 1996), which was the largest antihypertensive treatment trial and the second largest lipid-lowering trial ever conducted. And we will mainly focus on antihypertensive component of the trail in studying the effects of risk factors as

well as make dynamic predictions of event occurrences, for subjects at higher tail of the outcome (i.e. blood pressure). The detailed description of the ALLHAT study will be given in Section **??**.

References

Elizabeth R Brown, Joseph G Ibrahim, and Victor DeGruttola. A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73, 2005.

Barry R Davis, Jeffrey A Cutler, David J Gordon, Curt D Furberg, Jackson T Wright, William C Cushman, Richard H Grimm, John LaRosa, Paul K Whelton, H Mitchell Perry, et al. Rationale and design for the antihypertensive and lipid lowering treatment to prevent heart attack trial (allhat). *American Journal of Hypertension*, 9(4):342–360, 1996.

Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.

Robert M Elashoff, Gang Li, and Ning Li. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64(3):762–771, 2008.

Alessio Farcomeni and Sara Viviani. Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint modeling. *Statistics in medicine*, 2014.

Cheryl L Faucett and Duncan C Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in medicine*, 15(15):1663–1685, 1996.

Ronald A Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52(02):399–433, 1919.

Liya Fu and You-Gan Wang. Quantile regression for longitudinal data with a working correlation model. *Computational Statistics & Data Analysis*, 56(8):2526–2538, 2012.

Marco Geraci and Matteo Bottai. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154, 2007.

Xu Guo and Bradley P Carlin. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58(1):16–24, 2004.

Robin Henderson, Peter Diggle, and Angela Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

Sin-Ho Jung. Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, 91(433):251–257, 1996.

Roger Koenker. *Quantile regression*. Cambridge university press, 2005.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Kevin Hallock. Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56, 2001.

Hideo Kozumi and Genya Kobayashi. Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation*, 81(11):1565–1578, 2011.

Haiqun Lin, Bruce W Turnbull, Charles E McCulloch, and Elizabeth H Slate. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457):53–65, 2002.

Yuan Liu and Matteo Bottai. Mixed-effects models for conditional quantiles with longitudinal data. *The International Journal of Biostatistics*, 5(1), 2009.

Youxi Luo, Heng Lian, and Maozai Tian. Bayesian quantile regression for longitudinal data models. *Journal of Statistical Computation and Simulation*, 82(11):1635–1649, 2012.

Cécile Proust-Lima and Jeremy MG Taylor. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*, 10(3):535–549, 2009.

Dimitris Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 2011.

Dimitris Rizopoulos and Pulak Ghosh. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, 30(12):1366–1380, 2011.

Jeremy MG Taylor, Yongseok Park, Donna P Ankerst, Cecile Proust-Lima, Scott Williams, Larry Kestin, Kyoungwha Bae, Tom Pickles, and Howard Sandler. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213, 2013.

AA Tsiatis, Victor Degruttola, and MS Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429):27–37, 1995.

Yan Wang and Jeremy M G Taylor. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96(455):895–905, 2001.

Jane Xu and Scott L Zeger. Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):375–387, 2001.

Keming Yu and Rana A Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4): 437–447, 2001.

Menggang Yu, Ngayee J Law, Jeremy MG Taylor, and Howard M Sandler. Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, 14(3):835–862, 2004.

Menggang Yu, Jeremy M G Taylor, and Howard M Sandler. Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. *Journal of the American Statistical Association*, 103(481): 178–187, 2008.