Computational Statistics and Data Analysis 91 (2015) 40–50

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Joint latent class model of survival and longitudinal data: An application to CPCRA study

Yue Liu [a,*], Lei Liu [c], Jianhui Zhou [b]

[a] Takeda Oncology Co., Cambridge, MA, United States
[b] Department of Statistics, University of Virginia, Charlottesville, VA, United States
[c] Department of Preventive Medicine, Northwestern University, Chicago, IL, United States

ARTICLE INFO

Article history:
Received 21 February 2014
Received in revised form 26 May 2015
Accepted 26 May 2015
Available online 3 June 2015

Keywords:
Counting process
Proportional hazards model
Gaussian quadrature
Mixture model
Dependent censoring
Informative dropout
Frailty model
Mixed effects model
Survival analysis

ABSTRACT

There has been an increasing interest in the joint analysis of repeated measures and time to event data. In many studies, there could also exist heterogeneous subgroups. Thus a new model is proposed for the joint analysis of longitudinal and survival data with underlying subpopulations identified by latent class model. Within each latent class, a joint model of longitudinal and survival data with shared random effects is adopted. The proposed model is applied to Terry Beirn Community Programs for Clinical Research on AIDS study (CPCRA) to characterize the underlying heterogeneity of the cohort and to study the relation between longitudinal CD4 measures and time to death. The proposed model is desirable when the heterogeneity among subjects cannot be ignored and both the longitudinal and survival outcomes are of interest.

1. Introduction

In the clinical trial from the Terry Beirn Community Programs for Clinical Research on AIDS (CPCRA) study (Abrams et al., 1994; Neaton et al., 1994), 467 patients infected with the human immunodeficiency virus (HIV) were randomized into treatment daddanoisene (ddI) and zalcitabines (ddC). In this study, repeated measures of CD4 cell counts were recorded every 2 months from baseline to 20 months. The terminal event death was also recorded. We are interested in the association between CD4 count and death rate, and whether ddI and ddC have different effects on these two outcomes.

Several approaches exist to investigate the relation between longitudinal and time to event data. Some studies consider the repeated measurements as time-dependent covariates in the Cox proportional hazards model, e.g., Robinson et al. (2005) and Bradbury et al. (2008). However, such method requires a complete set of repeated measures in a time-continuous process, while in practice, the biomarker is often measured only at discrete time points. Although imputation techniques (e.g., last observation carried forward) can be adopted to obtain the value of biomarker at event occurrence, such crude practice could lead to inappropriate inferences, especially when the event time is far away from the most recent observation time of the biomarker. Furthermore, there could exist measurements with error which could lead to bias toward the null hypothesis (Prentice, 1982). To deal with this problem, Dafni and Tsiatis (1998) proposed a two-stage model to reduce

---

[*] Corresponding author.
E-mail address: yl7z@virginia.edu (Y. Liu).

http://dx.doi.org/10.1016/j.csda.2015.05.007

the bias of the estimate. However, two-stage model will estimate the longitudinal and survival processes sequentially, instead of simultaneously, which may bring bias especially due to the informative censoring of the longitudinal outcomes. An improvement over the two-stage model is to use data from both longitudinal and survival processes at the same time. Wulfsohn and Tsiatis (1997) proposed a joint model to study the relation between longitudinal CD4 measures and death time, in which a random effects model is adopted for the longitudinal CD4 process, and the survival model shares the same random effects. The parameters are estimated by maximizing the joint likelihood. Methods in Guo and Carlin (2004) and Liu and Huang (2009) were also developed to study the relation between CD4 and death hazard using joint modeling approaches.

However, such joint model only allows subjects to exhibit one pattern. It becomes inappropriate when there are underlying subgroups with different patterns of the response profiles. We thus need to consider the potential subtypes of longitudinal measurements and survival outcomes which may follow different distributions. Latent class models are often used to handle such heterogeneous processes. Lin et al. (2002) proposed a latent class model for joint analysis of longitudinal biomarker and event process data. The uncertainty of the class membership is characterized by a multinomial logistic model. They assumed that longitudinal and event processes are independent within each latent class, and applied their model to prostate-specific antigen readings and the time to prostate cancer data. Their model identified four subgroups representing various levels of prostate antigen trajectories and hazard of prostate cancer. Larsen (2004) used a latent class model to analyze time-to-event and multiple binary indicators simultaneously in Women's Health and Aging Study. Three latent classes were discovered through the joint latent class model, indicating the severity of mobility problem. Roy (2007) discussed how to apply latent class models to identify three missing-data patterns in a depression data with longitudinal Hamilton rating scale measures. Qin et al. (2009) presented a modified two-latent-class model for categorical data with informative dropout in smoking cessation data to represent the extent of smoking stubborn. Proust-Lima and Taylor (2009) also proposed a joint latent class model, investigating the association of prostate-specific antigen biomarker with prostate cancer recurrence. They identified five latent classes, indicating different levels of prostate antigen measures and risk of recurrent prostate cancer occurrence.

Most of the above methods assume conditional independence of longitudinal and survival outcomes within each latent class. Such a conditional independence assumption may be inadequate to capture the correlation between the two outcomes, which might result in underestimation of the association strength and uncertainty of the association structure between these two processes. Of note, a score test was proposed by Jacqmin-Gadda et al. (2010) to assess whether the conditional independence assumption holds in joint latent class models. Garre et al. (2008) proposed a joint latent class change point model to improve the prediction of time to graft failure within the Bayesian framework. However, in their model, the class patterns of trajectories were specified before modeling, which might not be applicable when the mixture patterns are not clear or straightforward. Another limitation is the simple structure in the longitudinal and time to event models which did not consider other covariate effects such as baseline characteristics. Neither did they study the probability of a subject belonging to each latent class in a model based approach. Beunckens et al. (2008) used latent class model to analyze the incomplete longitudinal data. However, they used logistic model for the missing status, without taking account of the time to event. Furthermore, they only evaluated a simplified model in the simulations, with the same time effect across two latent classes, and no covariates were considered in the logistic model.

In this paper, we integrate the joint random effects model into the latent class model framework to better capture the heterogeneity in both longitudinal and survival processes using a frequentist approach. We assume that within each latent class, there is a distinct joint model of longitudinal and survival endpoints with shared random effects. Our model is more comprehensive in that it can identify factors which impact the latent class membership and model the association between repeated measures and time to event processes. Besides, it provides a general setting which can consider class specific effects of covariates on longitudinal and survival models.

The rest of the paper is organized as follows. In Section 2, we present the model and the estimation method. The simulation results to evaluate the performance of our estimation method are shown in Section 3. In Section 4, we apply our model to the motivating CPCRA data. Discussion and concluding remarks are provided in Section 5.

2. Model and estimation

2.1. Latent class model

Suppose that there are N subjects, labeled as $i = 1, 2, \ldots, N$, and K latent classes labeled as $k = 1, 2, \ldots, K$. We define $R_{ik} = 1$ if subject i belongs to class k, and $R_{ik} = 0$ otherwise. The class membership is described in a logistic model with a class-specific parameter vector $\boldsymbol{\alpha}_k$:

$$\pi_{ik} = P(R_{ik} = 1) = \frac{\exp(\mathbf{X}_i^T \boldsymbol{\alpha}_k)}{1 + \sum_{k=1}^{K-1} \exp(\mathbf{X}_i^T \boldsymbol{\alpha}_k)}, \tag{1}$$

where π_{ik} denotes the probability that subject i belongs to latent class k, and $\mathbf{X}_i$ is the covariate vector for subject i. For identification, we take $\boldsymbol{\alpha}_K = 0$, i.e., we treat the last latent class as the reference.