A Bayesian Joint Hierarchical Model for Long-Term Multiple Substance Use and Recovery from Substance Use

Li-Jung Liang[1], Chi-hong Tseng[1], Sitaram Vangala[1], Yih-Ing Hser[2]
[1]Department of Medicine Statistics Core, University of California, Los Angeles, CA,
[2]Integrated Substance Abuse Programs, University of California, Los Angeles, CA

Abstract

Research on how the patterns of addicts' early-period substance use can predict their recovery from long-term substance use, using data from natural history interview studies, is limited. We propose to use a Bayesian joint hierarchical model to investigate the association between patterns of addicts' early-period substance use (longitudinal) and time to recovery from substance use. This approach allows us to properly account for the correlations among multiple drugs within subjects and to provide efficient estimates for the association between time to recovery and long-term use of multiple drugs. A 33-year follow-up study was used to demonstrate our approach.

Key Words: Joint Model, Two-Stage Model, Poly-drug Use

Introduction

Recovery from addiction is a complex process, and often is not homogeneous across individuals. Understanding whether the pattern(s) of addicts' early-period substance use can predict their recovery from substance use has been an interesting research topic for substance abuse researchers. There are several statistical challenges in studying recovery from substance use among long-term addicts. First, there are many different ways to describe or quantify early-period histories of substance use. The early-period history can be described empirically; examples include average monthly usage or frequency of substance use for the first 3 years since the onset of first substance use. The early-period history can also be characterized more sophisticatedly using complex modeling approaches, such as a growth curve model. The second challenge is that poly-drug use (more than one drug, or alcohol and drugs combined) is common among drug users. The pattern of early-period history of substance use is usually estimated separately for each substance. Thus, ignoring correlations among the multiple drugs essentially wastes valuable information, which could possibly result in biased estimates. Third, the association between time to recovery from substance and pattern of early-period substance use cannot be efficiently and properly estimated when these two are modeled separately, leading to inaccurate predictions of recovery.

Joint models for longitudinal outcome and time-to-event data are models that bring these data types together into a single model so that one can infer the dependence and association between the longitudinal outcomes and time to event. There has been much previous work in joint modeling of longitudinal single outcome and survival data; for example, DeGruttola et al (1994), for modeling progression of CD4-lymphocyte count and its relationship to survival time; Tsiatis et al (1995), for modeling the relationship of survival to longitudinal data, with application to survival and CD4 counts in patients with AIDS. The joint modeling approach has not been utilized in substance abuse research because of a few key obstacles. First, there is no available software that can easily fit this type of model. Second, the estimation procedure is usually computationally intensive.

In this study, we present a statistical and conceptual framework that is relatively new in its application to substance abuse research, and the implementation of our framework is computationally feasible. We construct various history functions of the addicts' long-

term use of multiple drugs that can be included in a time to recovery regression model. We propose to use a Bayesian framework that jointly models history functions of multiple substance use and time to recovery from substance use to determine the impact of the history functions on the time to recovery outcome. There are several important features of our approach. First, our joint model considers pattern(s) of multiple drugs simultaneously to properly account for the correlations among multiple drug use within subjects. Second, this model provides more efficient estimates for the associations between time to recovery and history function(s) of addict's early-period substance use, leading to more accurate predictions. Third, a Bayesian approach offers flexibility in implementing a complex hierarchical model that involves different types of outcomes using Markov chain Monte Carlo (MCMC) techniques.

Methods

Motivating Application

The motivating application is a 33-year follow-up study of narcotics addicts (Hser et al, 2001). This study used a natural history interview (NHI) instrument developed and implemented by CALDAR. In this study, a cohort of four hundred seventy-two male heroin addicts (37 Blacks, 273 Hispanics and 162 Whites), admitted in the California Civil Addict Program (CAP) between 1962-1964, were followed-up and interviewed over more than 30 years. During the 1960s, the CAP was the only major publicly-funded drug treatment program available in California. CAP provided a combination of inpatient and outpatient drug treatment program to narcotics-dependent criminal offenders committed under court order. A distinctive feature of this type of study is that researchers collect long-term, multiple substance use history data and behaviors along with other potentially related information (http://www.caldar.org/html/natural-history.html).

Notation

Data in this study are early-period substance use ($y_{ijk}, i = 1, \dots, N, j = 1, \dots, n_i,$ and $k = 2, \dots, p$) for the ith subject at jth month use kth drug, and time to recovery from substance use $(t_i, X_i\ \delta_i)$. Here y_{ij1} is an indicator of incarceration. *Early-period substance use* refers to the first three years of substance usage reported monthly starting from the onset of first drug use, and *recovery from substance use* is defined as when an addict was abstinent from all substance use and not incarcerated for twelve consecutive months after the addict's early-period substance use (Hser, 2007).

Models

The ***two-stage approach*** proposed by Guo and Carlin (2004) involves sequentially fitting a longitudinal model, summarizing patterns of early period substance use in the form of a history function, and a survival model, predicting time to recovery from substance use with the estimated history function as a covariate. The advantage of this approach is ease of implementation in existing software. The stages can be implemented using SAS PROC MCMC and PROC PHREG (Bayesian), respectively (version 9.3, SAS Institute Inc., Cary, NC, USA). We first fit a generalized linear mixed-effects model of primary substance use y_{ij}. The longitudinal model is written as follows:

$$g\left(E\left[y_{ij}\right]\right) = \alpha_i = X'_{ij}\alpha + u_i\ (i = 1, \dots, n;\ j = 1, \dots, n_i) \quad (1)$$

where g is a link function, X_{ij} is a matrix of subject-level covariates, α is a vector of regression coefficients, and u_i is a random intercept that follows a normal distribution with mean zero and variance σ^2. Next, we fit a piecewise-constant baseline hazard model

of time to recovery from substance use. Let $a_0 \equiv 0 < a_1 < \cdots < a_{J-1} < a_J \equiv T$ be a partition of the time axis. The survival model is

$$h(t_i|\beta,\eta) = h_0(t)\exp(X_i'\beta + \eta\psi(\alpha_i|Y_i) \quad (2)$$

$$h_0(t) = \lambda_j;\ a_j \le t < a_{j+1}\ (j = 1,\dots,J) \quad (3)$$

where h is the hazard function and h_0 is the baseline hazard function. Here β is a parameter vector linking baseline covariates to the recovery time and η is a scalar parameter linking the history function to the hazard function. We adapt a Bayesian approach, which assumes prior distributions for all unknown parameters. Normal priors with large variances for regression parameters (α, β, η) and an inverse gamma prior for σ^2 were chosen; a uniform prior was used for the cutoff values a_j's. The means of these prior distributions were set equal to the respective parameters' maximum likelihood estimates.

In this paper, we developed a ***joint hierarchical model*** to link longitudinal poly-drug use $(y_{ij2}, y_{ij3}, \dots, y_{ijp})$ and time to recovery from substance use $(t_i, X_i\ \delta_i)$. Here y_{ij1} represents incarceration. The goal is to investigate whether baseline covariates (e.g., race/ethnicity, age of onset) or history function of poly-drug use (say, *p* different drugs) can predict time to recovery from substance use. We fit a piecewise-constant baseline hazard regression model to the time to recovery data and fit a multivariate logistic model to the longitudinal poly-drug use data. We consider a random intercept model, i.e., $Z_{ij} = 1$ for this study. The time-to-event model is similar to the one described above. This multivariate logistic model allows us to flexibly construct either an average drug effect or q-degree polynomial $\left(t_{ij}^q\right)$ time effects that accommodates possible non-linear pattern of substance use. Another advantage of this model is that different drugs can have their own time effects. The shared random-effects vector U_i follows a p-dim multivariate normal distribution with mean vector zero, and covariance-variance matrix D. The pre-defined posterior history function $\psi(\alpha_i|Y)$ can be estimated and included in the survival regression model as a covariate to represent individual subject's history function of poly-drug use. Our joint modeling framework is described as follows:

$$\left.\begin{array}{c} y_{ij1} \sim Binary\left(\pi_{ij1}\right) \\ \\ \begin{array}{c} y_{ij2} \sim Binary\left(\pi_{ij2}\right) \\ \dots \\ y_{ijp} \sim Binary\left(\pi_{ijp}\right) \end{array} \end{array}\right\} if\ y_{ij1} = 0 \quad (4)$$

$$\text{logit}\left(\pi_{ijk}\right) = X_{ij}'\alpha_k + Z_{ij}U_{ik},\ k = 1,\dots,p \quad (5)$$

$$U_i = \begin{bmatrix} U_{i1} \\ U_{i2} \\ \dots \\ U_{ip} \end{bmatrix} \sim MVN(0, D)$$

Thus, the log-likelihood $l_i^1(\alpha, \mathrm{D}) = \log\left(L_i^1(\alpha, \mathrm{D})\right)$ for the longitudinal poly-drug use is given by

$$l_i^1(\alpha, \mathrm{D}) = \log\left\{\prod_{j=1}^{n_i}\left[Y_{ij}|\alpha, U_i, D\right] \times [U_i|D] \times [\alpha, D]\right\} \propto \sum_{j=1}^{n_i}\left(y_{ij1}\log\left(\pi_{ij1}\right) + \left(1 - y_{ij1}\right)\left[\log\left(1 - \pi_{ij1}\right) + A_{ij}\right]\right) + \log\left(|D|^{\frac{1}{2}}\right) + U_i^T D^{-1}U_i + log([\alpha, D]) \quad (6)$$

$$where\ A_{ij} = y_{ij2}\log(\pi_{ij2}) + (1 - y_{ij2})\log(1 - \pi_{ij2}) + y_{ij3}\log(\pi_{ij3}) + (1 - y_{ij3})\log(1 - \pi_{ij3})$$

For the time-to-recovery data, we used the same hazard function defined in (2) and (3) for subject *i*. The baseline cumulative hazard function is given by

$$H[i.j] = \begin{cases} 0 & t_i < a_j \\ t_i - a_j & a_j \le t_i < a_{j+1} \\ a_{j+1} - a_j & t_i \ge a_{j+1} \end{cases} \quad (7)$$

We define an indicator function

$$N[i,j] = \begin{cases} 1 & a_j \le t_i < a_{j+1} \text{ and } \delta_i = 1 \\ 0 & \text{Otherwise} \end{cases}$$

Thus, the likelihood $L_i^2(\beta, \eta, \lambda)$ for time to recovery is given by

$$L_i^2(\beta,\eta,\lambda) = \prod_{j=1}^{i^*}\left(\exp\{X_i'\beta + \eta\psi(\alpha_i|Y)\}\lambda_j\right)^{N[i,j]} \exp\left(-\exp\{X_i'\beta + \eta\psi(\alpha_i|Y)\}\,\mathrm{H}[i,j]\,\lambda_j\right) \quad (8)$$

where i* is the largest integer with $a_{i*} \le t_{i*}$

In this joint hierarchical model, we have a set of parameters and these parameters have their own prior distributions. They are regression coefficient parameters α_k's for incarceration and all types of substances, β for risk factors, η for an individual addict's history function of poly-drug use, covariance-variance matrix D, and the cutoff values a_j's. We use non-informative priors, i.e., normal priors with large variances for the regression parameters and gamma priors for the cutoff values.

Computation

Parameters were sampled sequentially using the **adaptive *Monte Carlo Markov Chain* (MCMC) algorithm** proposed by Vihola (2011), which is a robust adaptive Metropolis algorithm with coerced acceptance rate (R package: adaptMCMC). The additional feature we added to the algorithm is to create a latent membership $R(\alpha)$, based on the posterior probability vector for the i[th] drug user (in our motivating application, π_{i1}, π_{i2}, π_{i3} for probabilities of incarceration, heroin, and other drug use, respectively, during the early period). This step allows us to identify the high-risk set of drug users based on the pre-defined posterior probability criteria. The algorithm is summarized as follows.

1. Sample $[\alpha \mid \text{rest}, Y, t, \delta]$ from $\log\{ p(\alpha \mid \text{rest}, Y, t, \delta)\} \propto \sum_{i=1}^{K}\sum_{j=1}^{n_i}\{y_{ij1}\log(\pi_{ij1}) + (1 - y_{ij1})[\log(1 - \pi_{ij1}) + A_{ij}\,]\} + \log\{p(\alpha)\}$

2. Sample $[\,D^{-1} \mid \text{rest}, Y, t, \delta] \sim \text{Wishart}\left(\nu_o + K, \left(S_0^{-1} + \sum_{i=1}^{K} U_i U_i^T\right)^{-1}\right)$

3. Sample $[\,U_i \mid U_{(-i)}, \text{rest}, Y, t, \delta]$ from $\log\{p(U_i \mid U_{(-i)}, \text{rest}, Y, t, \delta)\} \propto \sum_{i=1}^{K}\left(\sum_{j=1}^{n_i}(y_{ij1}\log(\pi_{ij1}) + (1 - y_{ij1})[\log(1 - \pi_{ij1}) + A_{ij}\,]) + U_i^T D^{-1} U_i\right)$

4. Sample $[\,\beta \mid \text{rest}, Y, t, \delta]$ from $\log\{ p(\beta \mid \text{rest}, Y, t, \delta)\} \propto \sum_{i=1}^{K}\log(L_i^2(\beta,\eta,\lambda)) + \log\{p(\beta)\}$

5. Sample $[\lambda_i|\ \text{rest}, \text{Y}, \text{t}, \delta] \sim \text{Gamma}\left(a_j + n_j, b_j + \sum_{i=1}^{K} \text{e}^{X_i'\beta+\eta\psi(\alpha_i|Y)}\ \text{H}[i,j]\right), j = 1, \ldots, J$

We demonstrated our proposed approaches in three different models: (1) a simple Cox model without consideration of any history functions, (2) a two-stage model, and (3) a joint hierarchical model.

Results

Sample Characteristics

Hser (2001) found that 66% used heroin before age 20, 85% were arrested before age 18, the mean age at CAP admission was 25, and 67% were incarcerated at least once during their early-phase substance use. During a median of 30 years (range: 7.2 to 30 years) of follow-up, 305 drug users reached the recovery criteria; around 60% of Black and Hispanic drug users vs. 71% of White drug users reached the recovery criteria after the initial three years of substance use (P = 0.044). In the first month of the 3-year early-phase substance use, almost 92% of Black and Hispanic users, compared with less than 80% White users, consumed heroin. Over 65% of non-White users used a single type of substance, and 44% of White users used multiple drugs in the initial month (Table 1).

Estimated History Functions and Time to Recovery among Races

Compared with White drug users, Hispanic drug users had a lower recovery hazard in the model with race only (HR=0.23, 95% creditable interval: 0.01, 0.45). In Table 2, we present a few simple, but interpretable, history functions. For example, the estimated probability of an addict who used heroin and other drugs simultaneously was significantly higher for Whites vs. the other races (P <.05). We included this history function in the time to recovery model (Two-Stage Model B in Table 3) and found the same racial effect on time to recovery from substance use; however, the racial effect was attenuated (HR: 0.72, 95% CI:0.56, 0.91). Moreover, we observed higher frequency of heroin use with other drugs was associated with a longer time to recovery (HR = 0.95 for a 10% increase; 95% CI: 0.92, 0.98).

Our joint model showed similar racial differences in time to recovery as those from the two-stage approach (see Figure 1). Since the joint model took into account the correlation among multiple drugs, we found the Heroin use was negatively associated with use of other drugs (ρ = -0.12, 95% CI: -0.21, -0.03). Furthermore, the Bayesian approach allows us to calculate the posterior probabilities of incarceration, Heroin use, and use of other drugs for each individual. Eighty-eight percent of participants had their posterior probability of Heroin use greater than .80, of which almost 60% were Hispanic users; more than 32% of participants had their posterior probability of use of other drugs greater than .50, of which almost half were White users. The pattern of use of heroin and other drugs by race can be seen in Figure 2. The participants who had high posterior probabilities of use of Heroin and other drugs were considered as having risky drug use behavior. For example, we defined risky drug use behavior as high probability of Heroin use (> .80) and moderate probability of use of other types of substance (> .50). Instead of including the individual probabilities in the survival component of the joint model, we included the indicator of risky drug use behavior as the history function in the survival model for both interpretation and prediction purposes. We found that the participants who were classified as having risky drug use behavior took a longer time to reach the recovery criteria (HR=0.66, 95% CI: 0.47-0.89).

Discussion

In this study, we implemented the commonly used two-stage approach for linking longitudinal and time-to-event data. We developed a joint hierarchical model that links longitudinal and time to event data. History functions can be understood as summaries of early-period substance use, and can be constructed based on simple summary statistics or complex longitudinal multivariate models. They can help researchers understand the behaviors of addicts' early-period substance use. In the motivating application, we observed using both modeling approaches that certain history functions were associated with time to recovery from addiction.

The implementation of two separate models (the two-stage approach) is relatively easy, which makes it useful for more practical applications. However, we only considered the primary substance used in this approach, which may not properly account for the subject-level correlations among multiple drugs. This part can be easily improved upon using the joint modeling approach. There are several advantages of using our proposed joint modeling approach. This approach allows for different sets of parameters for different longitudinal outcomes. For example, it is not surprising that addicts' often use heroin in combination with other substances or alcohol. However, the time effects for different drugs may be varied. Thus, the joint model can accommodate this by allowing each substance to have its own set of time parameters. With this model, we are able to properly account for the correlations among different drugs within subjects. The joint modeling approach also provides efficient estimates of the associations between time to recovery and history function(s) and patterns of drug use behavior, leading to more accurate predictions.

Due to the nature of the study's dependence on findings from previous research, the study itself is presented with some challenges and limitations. The trajectories of poly-drug use can be complex. The high dimensional longitudinal responses, such as multiple substances and alcohol, will make the estimation process challenging. Second, our selection of non-informative priors on fixed-effects may forgo the additional benefits of the joint modeling with informative priors.

References

DeGruttola V., Tu X. M. Modeling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics.* 1994;50(4):1003-1014.

Guo X., Carlin B. P. Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages. *The American Statistician*. 2004;58:16-24.

Hser YI, Hoffman V, Grella CE, et al. A 33-year follow-up of narcotics addicts. *Arch Gen Psychiatry.* 2001;58(5):503–508.

Hser YI. Predicting long-term stable recovery from heroin addiction: findings from a 33-year follow-up study. *Journal of Addictive Diseases.* 2007;26(1):51–60.

Tsiatis A. A., DeGruttola V., Wulfsohn M. S. 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*. 1995;90(429):27-37.

Vihola M. Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and Computing.* 2011;22(5):997–1008.

Acknowledgement: The project described was supported by a Pilot Study Support Program as part of the CALDAR (P30DA016383) from NIDA.

TABLE 1: Number and Type of Drug(s) Used at the initial month of Early-Period Substance Use by Race

Early-Period Substance Use (Initial Month)	**Black (N=37)**	**Hispanic (N=273)**	**White (N=162)**
Type of Drug			
Heroin	34 (91.9%)	251 (91.9%)	85 (79.0%)
Meth	4 (10.8%)	27 (9.90%)	43 (26.5%)
Marijuana	11 (29.7%)	94 (34.4%)	73 (45.1%)
Number of Drugs			
One	26 (70.3%)	178 (65.2%)	85 (52.5%)
Two	10 (27.0%)	91 (33.3%)	72 (44.4%)
All Three	1 (2.70%)	4 (1.47%)	5 (3.09%)

TABLE 2: Selected History Functions Based on Early-Period Substance Use

History Functions	**Black**	**Hispanic**	**White**
Monthly Average of Heroin Use Mean (SE)[1] (P=0.038)	16.3 (1.60)	16.9 (0.59)*	14.4 (0.77)
Prob. of Incarcerated Mean (SE)[2]	0.076 (0.023)	0.124 (0.013)	0.114 (0.016)
Prob. of Heroin Usage Mean (SE)[2]	1.00 (4E-5)	1.00 (2E-5)	0.999 (7E-5)
Prob. of Heroin + Other Drugs Mean (SE)[2]	0.034 (0.023)*	0.056 (0.013)**	0.194 (0.049)
Prob. of Heroin + Alcohol Mean (SE)[2]	0.068 (0.046)	0.153 (0.034)	0.117 (0.035)

[1]Estimated using Mixed-effects model; [2]GLMM; * P <.05, ** P <.001

TABLE 3: Posterior Summary of Recovery from Substance Use from Two-Stage and Joint Models with Selected History Functions

Model Parameter	**Posterior Hazard Ratio (95% Credible Interval)**		
	Two-Stage Models		**Joint Model**
	Model A	Model B	
Race (Ref=White)			
Black	0.73 (0.44, 1.10)	0.64 (0.38, 0.97)	0.66 (0.43, 0.98)
Hispanic	0.78 (0.61, 0.98)	0.72 (0.56, 0.91)	0.73 (0.63, 0.86)
History Function			
Heroin Use (in 10%)	0.92 (0.88, 0.98)		
Heroin + Other Drugs (in 10%)		0.95 (0.92, 0.98)	
Risky Drug Use Behavior (Ref=No)			0.66 (0.47, 0.89)

FIGURE 1: Posterior Recovery from Substance Use for At Least One-Year by Race from Joint Model

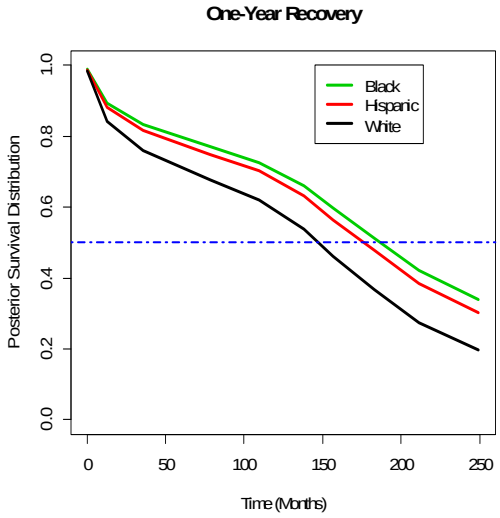

FIGURE 2: Posterior Probabilities of (Incarceration, Use of Heroin and Other Drugs) from Joint Model