Published in final edited form as:

Stat Med. 2019 January 30; 38(2): 175–183. doi:10.1002/sim.7580.

Innovative modeling of naturalistic driving data: Inference and prediction

Paul S. Albert

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20852

Abstract

Naturalistic driving studies provide opportunities for investigating the effects of key driving exposures on risky driving performance and accidents. New technology provides a realistic assessment of risky driving through the intensive monitoring of kinematic behavior while driving. These studies with their complex data structures provide opportunities for statisticians to develop needed modeling techniques for statistical inference. This article discusses new statistical modeling procedures that were developed to specifically answer important analytical questions for naturalistic driving studies. However, these methodologies also have important applications for the analysis of intensively collected longitudinal data, an increasingly common data structure with the advent of wearable devices. In order to examine the sources of variation between- and within-subjects in risky driving behavior, we explore the use of generalized linear mixed models with autoregressive random processes to analyzing long sequences of kinematic count data from a group of teenagers that have measurements at each trip over a 1.5 year observation period starting after receiving their license. These models provide a regression framework for examining the effects of driving conditions and exposures on risky driving behavior. Alternatively, Generalized estimating equations approaches are explored for the situation where we have intensively collected count measurements on a moderate number of participants. In addition to proposing statistical modeling for kinematic events, we explore models for relating kinematic events with crash risk. Specifically, we propose both latent variable and hidden Markov model in models for relating these two processes and for developing dynamic predictors of crash risk from longitudinal kinematic event data. These different statistical modeling techniques are all used to analyze data from the Naturalistic Teenage Driving Study (NTDS), a unique investigation into how teenagers drive after licensure.

Keywords

Hidden Markov model; Intensively collected longitudinal data analysis; Measurement error

*Correspondence to: Paul S. Albert, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20852.

1. Introduction

The high rate of driving accidents among teenagers is a crisis in public health. Understanding the dynamic driving behavior and risks of crashes is important for developing approaches for risk prediction as well as eventual intervention. The typical data structure has been described before in a number of the papers in this issue ([1] and [2]). The focus of this paper is to discuss recent modeling approaches that have been proposed for naturalistic driving studies. In the typical naturalistic driving study instrumentation is installed in a participant's car that assess kinematic measurements (e.g., changes in velocity) from stopping, turning, and related maneuvers. G-force data are collected continuously from the beginning to the end of each trip taken in the outfitted vehicle. Each g-force event over a specified elevated threshold is treated as a count or rate per miles driven. Thus, the typical data structure is a long sequence of counts, each count reflecting the number of g-force events during a trip. Typically, these studies have many events taken on a moderate number of participants. This is an unusual data structure in biostatistics, where most longitudinal data sets have only a few observations taken on a large number of subjects. However, with the advent of new wearable devices that collect extensive amounts of data, there has been an increased amount of research in intensively collected longitudinal data ([3], [5], [4], and [6]). The theoretical properties of Generalized estimating equations (GEE) have been extensively explored for the traditional case where the number of longitudinal measurements per person is small and the number of participants gets increasingly large. In this paper we will review recently methodologies for analyzing intensively collected longitudinal data collected from naturalistic driving studies. These methodologies include GEE, random effects and latent process models, as well as latent class and latent variable modeling approaches. Two important inferential goals are to understand the effects of exposure and driving conditions on risky driving behavior as measured by trip-specific kinematic counts, and to use risky driving behavior as a predictor of crashes. The models we will discuss will focus on one of these two aims. In Section 2, we will discuss random effects and latent processes model for describing the natural history of kinematics across time. These models will be useful for describing both average changes (across individuals and populations) as well as to characterize the variability (both over time and across subjects) in kinematic behavior. In Section 3, we discuss the use of GEE in this setting. We show that the usual application of GEE might be problematic for the naturalistic driving study data structure, and we propose alternative adaptations of usual GEE procedures to alleviate some of the problems. The first two sections focus primarily on modeling kinematics over time and understanding the effects of external factors (covariates) such as night-time driving and teenage passengers on average kinematics behavior. In Section 4, we present different approaches for predicting crashes from longitudinal assessments of kinematic behavior. We propose a standard approach and both a latent variable and latent class modeling approach for addressing this problem. A summary and discussion of future directions is included in the Discussion in Section 5.

2. Model Formulation and Parameter Estimation

As described in the introduction, the data structure for most naturalistic driving studies is a large number of trips on a moderate number of participants. A natural formulation that can

use standard software for analysis is the generalized linear mixed model (GLMM) where the outcome is a count and the number of miles driven is accounted for by using an offset term for the number of miles driven during that trip. Specifically, the inclusion of the offset term $\log m_{ij}$ for the j th trip on the i th participant results in the interpretation of the fixed and random effects as changes on the rate of kinematic measurements per mile ([7]). Specifically, we can model the number of g-force events, Y_{ij} as Poisson with mean characterized by λ_{ij} where

$$\log \lambda_{ij} = \log m_{ij} + X'_{ij}\beta + Z'_{ij}b_i \quad (1)$$

The model (1) is a generalized linear mixed model ([8]) where X_{ij} is a vector of fixed effect covariates such as night versus daytime driving, the presence of passengers, and month of follow-up, while the vector Z_{ij} represents random effects that are often assumed to follow a normal distribution with mean 0 and variance Σ_b . Various techniques have been proposed for parameter estimation ranging from an approximate penalized quasi-likelihood approach ([8]) that performs well for count outcomes when the mean count is relatively large, and full maximum-likelihood using adaptive Gaussian quadrature ([9]) more generally when only a few random effects are included. This model can typically be fit with standard software packages such as R, SAS, and Stata. For applications in driving studies, we usually assume that only a random intercept term is incorporated. Specifically, this involves replacing $Z'_{ij}b_i$ by a simple scalar random effect b_i that has mean 0 and variance σ_b^2 . Such a simplified random effects structure is sensible when the correlation between observations on a given subject is considered exchangeable, that the correlation between consecutive observations does not depend on the difference in time between those observations. Model (1) applied to teenage driving is presented in Simons-Morton et al. ([10]). In this analysis, it was found that driving with an adult and night time driving showed a reduced risk of kinematic events, while having risky friends showed an increased risk of kinematic events.

There are some limitations to this modeling strategy proposed in Simons-Morton et al. ([10]). Specifically, equation (1) assumes that given the individual random effects (the single random effect in our application) then the count data follows a Poisson distribution. This assumption is a rather strong one in that we might expect over-dispersion even after accounting for an individual random effect. Further, we may expect serial correlation in that trips closer together in time may be more highly correlated than trips further apart. There are a number of reasons that could explain both over-dispersion and serial correlation. With regard to over-dispersion, we may have omitted variables (including characteristics of the particular trip) that we do not measure. Similarly for serial correlation, there may be time-dependent factors such as a teen's conflict with his/her's parents which may wax and wane over time and are not measured. Omitting such variables can induce serial correlation.

Kim et al., ([11]) proposed an alternative modeling strategy that accounts for both over-dispersion and serial correlation. This modeling framework extends the generalized linear mixed model to include two additional random components that induces over-dispersion and

serial correlation, respectively. Kim et al. proposed that the longitudinal outcome follows an independent Poisson distribution with conditional mean $\log \lambda_{ij} = \log m_{ij} + X'_{ij}\beta + b_{ij}$, where b_{ij} is a random process that varies across time and across individuals, and can be decomposed as the sum of three sources of variation, $b_{ij} = b_i + O_{ij} + s_{ij}$, where b_i , O_{ij} , and s_{ij} reflect between-subject, over-dispersion, and serial correlation variation, respectively. We assume that each of these random effects follow a normal distribution with mean zero and variances given by σ_b^2 , σ_O^2 , and σ_s^2 , respectively. The serially correlated random effect could be of different forms, but we considered the use of an Ornstein-Uhlenbeck (OU) process that has been used in the biostatistics literature for incorporating serial correlation for longitudinal data ([12]). Specifically, the OU process incorporates a continuous version of an autocorrelation structure, where $\text{var}(s_{ij}) = \sigma_s^2$ and $\text{cov}(s_{ij}, s_{ij'}) = \rho^{d_{ijj'}}$, where $d_{ijj'}$ is the time between observation j and j' on the i th subject. In general then, the variance of b_{ij} is $\sigma_b^2 + \sigma_O^2 + \sigma_s^2$, while the covariance between successive measurements as separated by driving time is $\text{cov}(b_{ij}, b_{ij'}) = \sigma_b^2 + \sigma_s^2 \rho^{d_{ijj'}}$.

Particularly due to the serial correlation, direct maximum-likelihood is difficult since it requires high-dimensional integration to evaluate the target likelihood. However, a Bayesian analysis provides a computationally more efficient way of performing estimation. A Gibbs sampling algorithm was proposed with hierarchical centering to improve the convergence of the Gibbs sampler. A Reversible jump Markov chain Monte Carlo algorithm to accommodate the flexible spline mean structure.

Table 1 shows the model parameters from the latent process model when fit to the NTDS data. Similar to the results reported in Simons-Morton et al. ([10]), we found that passenger presence, time of day, and having risk friends all played a statistically significant role. This is reassuring for the analysis since, in general, ignoring sizable over-dispersion and serial correlation results in liberal inferences (confidence intervals that are too small and statistical tests that have p-values lower than they should be). Aside from making proper inference about the regression coefficients, the latent process model allows us to examine the different sources of variation due to subject and within subject variation. The results in Table 1 demonstrate that the within-subject variation is at least as large as the between-subject variation in kinematic driving behavior. For example, the between-subject variation is the estimated σ_b^2 that is 0.287, while the within-subject variation is the sum of the estimated σ_O^2 and σ_s^2 , that is 0.394. Indeed, this suggests that the degree of tracking in this behavior is rather small and that particular triggers or time-dependent situational factors play a big role in risky kinematic teenager driving behavior. Further, the estimated serial correlation dies diminishes to zero with in approximate three months between trips. For example, the correlation at 0.5, 1, and 2 months is $(1.0175 \times 10^{-16})(1/36) = 0.36$, $(1.0175 \times 10^{-16})(1/18) = 0.36 = 0.12$, and $(1.0175 \times 10^{-16})(1/9) = 0.017$, respectively.

The fact that the within-subject variation is large relative to the between-subject variation has implications for studying teenage driving. First, it suggests that interventions should be targeted for all rather than a subset of teens who drive risky since all teens have the potential

for driving risky. Second, for intervention studies, taking multiple measurements on each subject is recommended to reduce variability.

3. Generalized estimating equations for long sequences of count data

Rather than fitting random effects or latent process models, Zhang et al.([13]) considered using generalized estimating equations (GEE) approaches for fitting repeated kinematic event data in naturalistic driving studies. This methodology where only the first two moments are specified is easier to fit than the latent process models and may be more robust to misspecification of the assumed variance structure. Similar to the latent process model, we specify the marginal mean of Y_{ij} as $E(Y_{ij}) = m_{ij} \exp(\nu + W_i' \alpha + X' \beta)$, where the marginal variance structure can be written as a function of the regression coefficients and variance matrix of the latent processes. Note that we distinguish between the subject-specific and time-dependent covariates (W_i and X_{ij} , respectively) since the statistical properties are different for the two types of covariates. Standard GEE with a robust variance assumption is known to have poor statistical properties when the number of longitudinal measurements is large and the number of subjects is small to moderate (as in this naturalistic driving setting where for the NTDS study, we have 42 subjects who have as many as 3000 trips over the 18 month follow up period). Simulation studies show that for subject-specific covariates, the median robust SEs are severely underestimated compared with the Monte-Carlo variances. Further, the covariate rates are too low (coverage is approximately 90% for 95% targeted rates). The statistical properties are substantially improved when we estimate individual effects using GEE, estimating a fixed effect intercept for each participant, and then regress these effects onto the subject-specific covariates of interest. For this procedure, the robust standard errors are approximately unbiased and the coverage rates are at the nominal level ([13]). However, the statistical properties for estimating the time-dependent coefficients β are still poor (standard errors that are too small and coverage rates that are significant lower than is being targeted). We examined various approaches to improve these properties including within cluster resampling ([14] [15]) where a window of multiple consecutive measurements is sampled from each participant and a GEE estimate of the regression coefficients is obtained. This procedure is repeated multiple times (e.g., $N = 100$) and the average regression coefficient estimate is then obtained. Although estimation of β was improved over GEE, we still had coverage probabilities that were too low relative to targeted rates. As an alternative, we proposed a within cluster resampling procedure that contains separated blocks of measurements. Specifically, we sample block sizes with a certain number of trip measurements (i.e., 100) that are separated by S trips and continue until the end of the sequence. Further, we treated the separate blocks as independent from each other (remember that by including a fixed effect intercept, we remove the exchangeable correlation structure and only have serial correlation remaining). This within-cluster block resampling resulted in nearly nominal level coverage rates for confidence intervals.

Table 2 shows marginal parameter estimates obtained using GEE estimation with within-cluster resampling of separated blocks of consecutive trips. Confidence intervals were estimated using robust variance estimators of the parameter estimates. A bit different than the latent process model, we separately evaluated adult versus teenage passengers and late

night versus early night driving. We saw similar results to the latent process model where driving with passengers (either adult or teenager) reduced risky driving, evening driving reduced risk, and having risky friends increased risky driving. The fact that the kinematic event rate is reduced during evening driving relative to day-time driving is not seen for crash rates, where crash rates are higher in the evening and late night. However, higher kinematics are associated with higher crash rates in the day, early evening, and at night.

There are a few things to note when comparing the latent process versus the GEE approaches for analyzing data with this structure. First, because the link function for both approaches is a log-link, with the exception of intercept term, the coefficients have the same interpretation in both modeling approaches. Second, as compared with the GEE approach, the computational burden is substantially larger for the latent process approach where the convergence of the Gibb's sampling was slow. Third, the latent process model provides additional information about the sources of variation in the intensively collected repeated count data. This additional information was difficult to precisely estimate with the GEE approach.

4. Predicting crashes from intensively collected kinematic data

An important goal in driving research is to use naturalistic driving studies to develop predictors of crash risk from longitudinal kinematic measurements. Simons-Morton et al. ([16]) developed a predictor of whether a crash occurs at a particular month depending on their previous kinematic measurements. Due to the fact that monthly crashes are rare, the occurrence of either a crash or near crash was used as the outcome. Near crashes are defined as an incident for which a crash would have occurred if the other driver had not performed an active maneuver to avoid contact. Simons-Morton et al. used GEE with a logit link function to model the probability of a crash at a particular month as a linear function (on the logit scale) of kinematic measurements taken before the month began and the past history of crashes. An offset to account for the number of miles driven during a particular month was also included in the GEE modeling. Different ways to characterize the past history of kinematic events were explored including summarizing past kinematic exposure as the mean count of elevated events in the past month, 2 weeks, 1 week, 100 miles, and 1,000 miles. A generalized estimating equation with a logit link function demonstrated that an increase in the composite gravitational-force event rate resulted in an increase in the probability of having a crash for all summary measures of past kinematic exposures. Using the past month summary measure did provide the best prediction as measured by a cross-validated AUC of the ROC curve. The GEE estimation used an independence working correlation structure, and was conducted using standard statistical software. The resulting AUC of the ROC curve was 0.76. It is important to recognize for the models that are being considered, unlike in the previous modeling of trip-level kinematic outcomes, we are summarizing both kinematic events and crash occurrence at a monthly level. In part, this was done so that the risk predictor can be easily computed and interpreted thereby making it more useful to the practitioner.

In more recent work, Jackson et al. ([17]) proposed latent variable modeling approaches for studying the relationship between risky driving behavior as measured by longitudinal

kinematic measurements and the occurrence of a crash. Specifically, it was assumed that longitudinal binary (crash occurrence) and count (kinematic events) outcomes are observed and that there is an unobservable underlying process that represents risky driving which influences both outcomes. The model can be viewed as a two-stage model whereby in first stage, the latent variable describes the observed binary sequence (whether or not one or more crashes were observed on a given month over the 18 month observation period), and in the second stage, the latent variable is modeled as a function of the previous observed indicator of a crash and the kinematic count process. Let $Y_{i1}, Y_{i2}, \dots, Y_{in}$ be the sequence of binary outcome variables reflecting the occurrence of a crash or near crash during a particular month j over an n month follow-up period ($n=18$ for the NTDS). Denote X_{ij} as a vector of kinematic counts for individual i during month $j-1$. For a binary latent variable, where b_{ij} is defined as the binary occurrence of being in a risky driving state ($b_{ij}=1$) versus the occurrence of being in a non-risky driving state ($b_{ij}=0$), this model can be expressed as

$$\text{logit}P(Y_{ij} = 1 | \alpha, b_{ij}) = \log(m_{ij}) + \alpha_0 + \alpha_1 b_{ij} \quad (2)$$

and

$$\text{logit}P(b_{ij} = 1 | y_{i,j-1}, x_{i,j-1}, \theta, \beta) = \theta_0 + \theta_1 y_{i,j-1} + \beta' x_{i,j-1} \quad (3)$$

The parameter α_1 characterizes the odds of the latent state on the probability of having a crash during the j th month. The logistic regression model includes an offset term, $\log(m_{ij})$, that accounts for the number of miles driven during that month. Correlation is induced between the crash or near crash occurrences (Y_{ij}) through the correlation in the latent process b_{ij} which in turn is induced through its dependence on lagged responses of the observed process $y_{i,j-1}$.

Model (2 and 3) can be extended in a number of ways. The individual propensity for being in the high or low crash-risk state may vary by individual. In the following model we include a random effect in the latent binary state model,

$$\text{logit}P(b_{ij} = 1 | y_{i,j-1}, x_{i,j-1}, \theta, \beta) = \theta_0 + \theta_1 y_{i,j-1} + \beta' x_{i,j-1} + u_i \quad (4)$$

where u_i is a random effect that is assumed to follow a normal distribution with mean zero and variance λ . Alternatively, The model can be extended to incorporate an ordinal rather than a binary latent variable as in the following formulation:

$$\text{logit}P(Y_{ij} = 1) = \log m_{ij} + \alpha_1 + \alpha_2 b_{ij2} + \dots + \alpha_K b_{ijK} \quad (5)$$

and

$$\text{logit}P(b_{ij} < k) = \alpha_k + \theta_1 y_{i,j-1} + \beta' x_{i,j-1}, \quad k = 1, 2, \dots, K-1. \quad (6)$$

where b_{ijk} is a indicator variable that is equal to 1 when the ordinal state b_{ij} is equal to k , and zero otherwise.

A further extension of (6) can also include the addition of a random effect,

$$\text{logit}P(b_{ij} < k) = \alpha_k + \theta_1 y_{i,j-1} + \beta' x_{i,j-1} + u_i, \quad k = 1, 2, \dots, K-1, \quad (7)$$

where u_i , as in (4) is assumed to be normal with mean zero and variance λ .

Parameter estimation for models (2 and 3) and (5 and 6) can be performed with an E-M algorithm where we iterate between a M- and E-step to obtain maximum-likelihood estimates ([17]). Parameter estimation for models (2 and 4) and (5 and 7) involves the use of Monte-Carlo E-M for which Monte-Carlo evaluation is required to evaluate the high-dimensional integration required in the E-step. Standard errors for the random-effects models were obtained by the use of a non-parametric bootstrap. For the binary and ordinal model that did not include random effects, Louis' method ([18]) for estimating standard errors was used. Table 4 shows parameter estimates for the binary, ordinal, and binary with random effects models. A fit of the ordinal random effects model showed variation close to zero for the random effect and therefore reduced to the ordinal model. Therefore, the ordinal random effect model parameter estimates were not shown. Along with parameter estimates, standard errors, area under the receiver operator curve (AUC), and a measure of the squared

difference between the observations and predictive values ($\sum_{i=1}^N \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$, where N is

the number of participants and n_i is the number of months for the i th subject) were presented.

The binary random effects model fit the data better (based on the AIC), although the estimate of λ , the variance of the subject-specific random effect was relatively small. The table shows the effects of individual kinematic measurements on the latent (binary or ordinal) state. Interpreting the model with a random effect, the estimate of α_1 is large reflecting that when in the high risk state the odds of having a crash is $\exp(2.77)=16.0$. Based on the model, for a typical subject who has driven 100 miles during a particular month, the probability of having a crash or near crash in the high state is 0.29 as compared with 0.02 for the same participant who is in the low risk state. Estimates of the model that governs the probability of being in the high state show sizable effects of the prior month's deceleration, yaw, and lateral accelerations. Interesting, the effect of acceleration was not statistically significant for any of the three models.

An alternative to the random effects model is a hidden Markov model to describe changes in b_{ij} over time. Specifically, b_{ij} follows a two-state Markov chain where the state space is

either 0 reflecting a good driver or 1 reflecting a risky driver. As in the previous models, a Bernoulli random variable describes the binary crash/near crash variable Y_{ij} ,

$$\text{logit}P(Y_{ij} = 1 \mid \alpha, b_{ij}) = \log(m_{ij}) + \alpha_0 + \alpha_1 b_{ij}. \quad (8)$$

The two-state hidden Markov model for b_{ij} can be characterized by an initial distribution for b_{i1} and the transition probabilities of transitioning from $b_{i,j-1} = 0$ to $b_{ij} = 1$ and from $b_{i,j-1} = 1$ to $b_{ij} = 0$. The initial probability can be characterized by one parameter, $r_1 = P(b_{i1} = 1)$, where $r_0 = P(b_{i1} = 0) = 1 - r_1$. The two transition probabilities that can depend on a vector of time-dependent covariates $x_{i,j}$ can be parameterized as

$$\text{logit}P(b_{ij} = 1 \mid b_{i,j-1} = 0) = \gamma_0 + \gamma'_1 x_{i,j-1}. \quad (9)$$

$$\text{logit}P(b_{ij} = 0 \mid b_{i,j-1} = 1) = \gamma_0^* + \gamma_1'^* x_{i,j-1}. \quad (10)$$

The likelihood that is needed to maximize in order to estimate the model parameters is

$$L(\gamma_0, \gamma_1, \gamma_0^*, \gamma_1^*, \alpha_0, \alpha_1) = \prod_{i=1}^N \sum_b \left\{ (r_1 f(y_{i1} \mid b_{i1} = 1) + r_0 f(y_{i1} \mid b_{i1} = 0)) \prod_{j=2}^n f(y_{ij} \mid b_{ij}) \right. \\ \left. P_{b_{i,j-1}, b_{ij}} \right\}, \quad (11)$$

To evaluate the above likelihood, we used an E-M algorithm whereby the forward-backward algorithm ([19]) was used for evaluating the E-step and the M-step was implemented using nlm procedure the R software package. Table 4 shows the estimated model parameters obtained by fitting the hidden Markov model. The estimates show a strong link between crashes and the latent good driving versus poor driving status. Similar to the previously discussed latent variable modeling approaches, the risk of a crash for a hundred miles of driving over a given month is 0.027 in the low risk state and 0.430 in the high risk state. A high kinematic event increased the risk of initiating a transition into the poor driving state ($\hat{\gamma}_1 = 15.5$) and decreased the risk of transitioning from the poor driving state to the good driving state ($\hat{\gamma}_2 = -7.11$). With the hidden Markov model we can compare the performance of predicting crash or near-crash given all the previous composite kinematics events and outcomes. Based on our data, the AUC was estimated as 0.75 which is smaller than our other latent variable modeling approaches.

The hidden Markov model described above incorporates dependence between outcomes through a two state hidden Markov model, where the effect of composite kinematic events is

incorporated by introducing this variable through covariate dependence in the transition probabilities. Jackson et al.[20]) proposed an alternative approach where the longitudinal kinematic measurements are directly linked to the crash/near crash rates through an underlying stochastic process b_{ij} . In this case, the underlying process is a two state hidden Markov chain reflecting two underlying states. The two transition probabilities that govern individual transitions are allowed to dependent on random effects that introduce heterogeneity across subjects. In this formulation, we assume that

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \text{ where } \text{logit}(\pi_{ij}) = \log(m_{ij}) + \alpha_0 + \alpha_1 b_{ij} + \alpha_2 u_i, \quad (12)$$

$$X_{ij} \sim \text{Poisson}(\mu_{ij}), \text{ where } \log(\mu_{ij}) = \log(m_{ij}) + \beta_0 + \beta_1 b_{ij} + \beta_2 t_j + \beta_3 u_i, \quad (13)$$

and

$$p_{01}(u_i): \text{logit}P(b_{ij} = 1 \mid b_{i,j-1} = 0 \mid u_i) = \gamma_{01} + \delta_1 y_{i,j-1} + u_i, \quad (14)$$

$$p_{01}(u_i): \text{logit}P(b_{ij} = 1 \mid b_{i,j-1} = 0 \mid u_i) = \gamma_{10} + \delta_2 y_{i,j-1} + \delta^* u_i. \quad (15)$$

The likelihood for this model is

$$L(\Psi; y, x) = \int_u \sum_b f(y \mid (b, u), \Psi) g(x \mid (b, u), \Psi) h(u; \Psi) du, \quad (16)$$

where Ψ is a vector of all parameters, f and g are the conditional density y 's and x 's, respectively, and h is the random effects distribution, that is assumed normal. Estimation was conducted using both Gaussian quadrature to integrate over the shared random parameter, and the backward-forward algorithm to sum over b . Jackson et al. [20]) fit both a two-state and a three-state underlying Markov chain for b_{ij} and on the basis of a comparison of the AIC, chose a two state model. Parameter estimates for all model parameters are presented in Jackson et al. Most notably, $\delta^* > 0$ indicates a positive correlation between the two transition probabilities, meaning that some subjects are prone to changing more often between states than others. Parameters in the hidden process, γ_1 and γ_2 , shows that transitions between states depends on the previous crash outcomes. A prior crash (in past month) was associated with an increasing probability of making a transition from the good to the poor driving state ($\hat{\delta}_1 = 1.75$), and a decreased probability of making a transition from the poor to the good driving state ($\delta_2 = -2.17$).

5. Discussion

Naturalistic driving studies provide unique methodological challenges. New statistical methodology has been developed to examine the effect of driving conditions on risky driving as measured by complex intensively collected kinematic events. A generalized linear mixed model for the analysis of trip-level kinematic count data is discussed. The model distinguishes between serial, subject-specific, and within-trip over-dispersion. Using this approach, we showed how these complex models can demonstrate that among recently licensure teenage drivers, the within-subject variation is higher than the between-subject variation. This uncovered observation tells researchers that there is room for interventions that reduce this sizable intra-person variability in how kids are driving.

Innovative new GEE approaches for marginal inference of trip-level data were also presented. Our work demonstrated that a straight-forward application of GEE can lead to problematic inference when the number of trips is very large. We propose a two-stage procedure for estimation of subject-specific covariates (using least-squares) and a within-cluster resampling in blocks for trip-level covariates for marginal inference in this setting. Both the generalized linear mixed modeling and the GEE approaches dealt with analyzing a single kinematic measurement at a time (the composite, for example). Analyzing the multivariate kinematic response, that includes multiple correlated counts, would be a subject for future research.

When modeling the time since licensure in the GLMM and GEE modeling frameworks, we considered calendar time as the time-scale of most importance. However, the dynamics of risky behavior over time may be more effected driving time (cumulative sum of the trip durations). Incorporating both types of dynamics into the assessment of risky driving would be an interesting and challenging methodological problem.

New approaches for the dynamic prediction of crashes from kinematic measurements were discussed. We presented both latent variable approaches as well as a hidden Markov model for the joint modeling of crashes and kinematic monthly summary measures. The models all showed that we have a relatively high predictive accuracy for crashing using the past kinematic profiles.

Our prediction approaches all used monthly count data for the kinematic events to predict whether crashes occurred on a monthly basis. For many situations, forming predictive models on a monthly basis makes sense from a practical viewpoint. Events are easy to summarize on a monthly basis and risk is easy to characterize on a monthly basis. However, there are some questions that can best be solved on a trip level basis.

An important, unsolved problem is on distinguishing the more transient versus longer term kinematic behavior profiles on crash risk. Specifically, of interest is on examining whether subject-specific (individual random effect components) versus short term stochastic variation (over-dispersion and serial correlation components) are associated with first crash risk. To address this question, a joint modeling approach can be formulated with the random subject-effect and the within-subject latent processes be separately linked to the time-to-event process with these two shared random processes being in both the longitudinal and survival

models. Specifically, we model the time to the first crash with the proportional hazards model ($t = \lambda_0(t) \exp(\eta_1 b_i + \eta_2(S_{ij} + O_{ij}))$), where the baseline hazard function can be exponential (constant) or piecewise exponential. The proportional hazards model in conjunction with the latent process model for kinematic events will allow investigators to distinguish between the effects of long-term subject-specific kinematic behavior (measured by η_1) and shorter-term fluctuations (measured by η_2) on the risk of a first crash.

Acknowledgments

The research was supported by the Intramural Research Programs of both the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development and the National Cancer Institute.

References

1. Simons-Morton B. Driving in search of analysis. *Statistics in Medicine*.
2. Dawson J. Practical and statistical challenges in driving research. *Statistics in Medicine*.
3. Walls, TA, Schafer, JL, editors *Models for Intensive Longitudinal Data*. Oxford University Press; Oxford:
4. Xiao L, Huang L, Schrack JA, Ferrucci L, Zipunnikov V, Crainiceanu C. 2015; Quantifying the lifetime circadian rhythm of physical activity: a covariate dependent functional approach. *Biostatistics*. 16:352–367. [PubMed: 25361695]
5. Bai J, He B, Shou H, Zipunnikov V, Crainiceanu C. 2013; Normalization and extraction of interpretable metrics from raw accelerometry data. *Biostatistics*. 15:102–116. [PubMed: 23999141]
6. Kim S, Albert PS. Latent variable Poisson models for assessing the regularity of circadian rhythms over time. *Journal of the American Statistical Association*.
7. Zeger SL, Qaqish B. 1988; Markov regression models for time series: a Quasi-likelihood approach. *Biometrics*. 44:1019–1031. [PubMed: 3148334]
8. Breslow NE, Clayton DG. 1994; Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. 88:9–25.
9. Pineiro JC, Chao EC. 2006; Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*. 15:58–81.
10. Simons-Morton BG, Ouimet MC, Zhang Z, Klauer SE, Lee SE, Wang J, Albert PS, Dingus TA. 2011; Crash and risky driving involvement among novice adolescent drivers and their parents. *American Journal of Public Health*. 101:2362–7. [PubMed: 22021319]
11. Kim S, Chen Z, Zhang Z, Simons-Morton B, Albert PS. 2013; Bayesian hierarchical Poisson regression models: an application to a driving study with kinematic events. *Journal of the American Statistical Association* bf. 108:494–503.
12. Taylor JMG, Cumberland WG, Sy JP. 1994; A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*. 89:727–736.
13. Zhang Z, Albert PS, Simons-Morton B. 2012; Marginal analysis of longitudinal count data in long sequences: methods and applications to a driving study. *Annals of Applied Statistics*. 6:27–54. [PubMed: 27087885]
14. Hoffman EB, Sen PK, Weinberg CR. 2001; Within-cluster resampling. *Biometrika*. 88:1121–1134.
15. Follmann D, Proschan M, Leifer E. 2003; Multiple outputation: inference for complex clustered data by averaging analyses from independent data. *Biometrics*. 59:420–429. [PubMed: 12926727]
16. Simons-Morton B, Zhang Z, Jackson JC, Albert PS. 2012; Do elevated gravitational-force events while driving predict crashes and near crashes? *American Journal of Epidemiology*. 175:1075–9. [PubMed: 22271924]

17. Jackson JC, Albert PS, Zhang Z, Simons-Morton B. 2013; Ordinal latent variable models and their application in the study of newly licensed teenage drivers. *Applied Statistics*. 62:435–450. [PubMed: 25284899]
18. Louis TA. 1982; Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B*. 44:226–233.
19. Baum LE, Petrie T, Soules G, Weiss N. 1970; A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*. 41:164–171.
20. Jackson JC, Albert PS, Zhang Z. A two-state mixed hidden Markov model for risky teenage driving behavior. *Annals of Applied Statistics*. 9:849–865.

Table 1

Parameter Estimates from Hierarchial GLMM model

Variable (β)	Posterior Mean	95% HPD
Passenger presence	-0.181	-0.194 to -0.168
Night Driving	-0.192	-0.204 to -0.182
Having Risky Friends	0.406	0.072 to 0.729
σ_b^2	0.287	0.165 to 0.423
σ_O^2	0.269	0.263 to 0.275
σ_S^2	0.125	0.113 to 0.137
	26.824	29.83 to 44.26

Table 2

Parameter Estimates from Hierarchical GEE with within-cluster resampling 100 block of 50 consecutive trips separated by 50 trips. An OU process working correlation structure was used in the GEE estimation.

Variable (β)	Posterior Mean	95% CI
Passenger Adult vs. None	-0.942	-1.079 to -0.777
Passenger Teen vs. None	-0.211	-0.301 to -1.05
Late Night Driving vs Day	-0.274	-0.528 to -0.342
Early Evening vs Day	-0.446	-0.528 to -0.342
Having Risky Friends	0.647	0.104 to 1.191

Table 3

Parameter Estimates (SE) For the Latent Variable Prediction Models

Parameter	Est. for Binary	Est. for Ordinal	Est. for Binary with RE
α_0	-8.30 (0.17)	-10.22 (0.52)	-8.27 (0.18)
$\alpha_1(b_{ij})$	2.68 (0.23)	—	2.77 (0.20)
$\alpha_2(b_{ij2})$	—	1.93 (0.68)	—
$\alpha_3(b_{ij3})$	—	4.41 (0.55)	—
θ_0	-3.27 (0.21)	—	-2.89(0.18)
$\theta_1(y_{ij-1})$	1.11 (0.22)	0.73 (0.27)	1.89 (0.33)
γ_1 (good or average)	—	1.38 (0.16)	—
γ_2 (average or poor)	—	1.87 (0.18)	—
β_1 (acceleration)	1.26 (4.54)	-0.01 (3.77)	7.76 (16.84)
β_2 (deceleration)	86.23 (10.97)	86.28 (12.21)	84.38 (12.62)
β_3 (yaw)	47.39 (5.62)	43.48 (4.99)	49.29 (6.32)
β_4 (lateral acceleration negative)	5.72 (2.16)	7.32 (2.12)	4.11 (2.09)
β_5 (lateral acceleration positive)	19.08 (4.80)	17.52 (4.53)	14.42 (4.11)
λ (RE variance)	—	0.44	0.21
AUC	0.78	0.77	0.78
$\sum_{i=1}^N \sum_{j=1}^{n_i} (y_{ij} - \widehat{y}_{ij})^2$	102.06	104.56	99.87

Table 4Hidden Markov model with x_{ij} being the number of composite events

Parameter	Estimate	Standard error
a_0	-8.24	0.31
a_1	2.76	0.47
γ_0	-3.36	-0.92
γ_1	15.5	8.70
γ_0^*	-0.36	0.71
γ_1^*	-7.11	12.65
λ	0.73	0.20