Received: 30 September 2016 | Revised: 2 November 2017 | Accepted: 5 November 2017

DOI: 10.1002/sim.7574

SPECIAL ISSUE PAPER

WILEY Statistics in Medicine

Semiparametric Bayesian models for evaluating time-variant driving risk factors using naturalistic driving data and case-crossover approach

Feng Guo[1,2] | Inyong Kim[1] | Sheila G. Klauer[2]

[1]Department of Statistics, Virginia Tech, Blacksburg, VA 24060, USA

[2]Virginia Tech Transportation Institute, Blacksburg, VA 24060, USA

Correspondence
Feng Guo, Department of Statistics, Virginia Tech, Blacksburg, VA 24060, USA.
Email: feng.guo@vt.edu

Funding information
Global Automakers, Grant/Award Number: #14-0729-10

Driver behavior is a major contributing factor for traffic crashes, a leading cause of death and injury in the United States. The naturalistic driving study (NDS) revolutionizes driver behavior research by using sophisticated nonintrusive in-vehicle instrumentation to continuously record driving data. This paper uses a case-crossover approach to evaluate driver-behavior risk. To properly model the unbalanced and clustered binary outcomes, we propose a semiparametric hierarchical mixed-effect model to accommodate both among-strata and within-stratum variations. This approach overcomes several major limitations of the standard models, eg, constant stratum effect assumption for conditional logistic model. We develop 2 methods to calculate the marginal conditional probability. We show the consistency of parameter estimation and asymptotic equivalence of alternative estimation methods. A simulation study indicates that the proposed model is more efficient and robust than alternatives. We applied the model to the 100-Car NDS data, a large-scale NDS with 102 participants and 12-month data collection. The results indicate that cell phone dialing increased the crash/near-crash risk by 2.37 times (odds ratio: 2.37, 95% CI, 1.30-4.30) and drowsiness increased the risk 33.56 times (odds ratio: 33.56, 95% CI, 21.82-52.19). This paper provides new insight into driver behavior risk and novel analysis strategies for NDS studies.

KEYWORDS

Bayesian semiparametric, case-crossover, driver behavior, naturalistic driving study, time-variant risk factor

1 | INTRODUCTION

Traffic crashes are one of the leading causes of death in the United States with more than 30 000 fatalities every year.[1] Driver behavior is a major contributing factor for crashes and has been a major area of traffic safety research.[2] To identify and quantitatively evaluate risk associated with driver behaviors is challenging because of the rapid change in driving behavior and its transient effects on crash risk. The absence of objectively collected information at the onset of crashes as well as during normal driving condition is an major obstacle in driving behavior research. The data necessary to conduct such analyses are traditionally drawn from accident databases or post-crash reconstructions. However, both data sources suffer from information bias as the data collected are subject to driver recall errors and the lack of precise time stamps for

the sequence of events. Alternatively, driving data can also be collected in controlled experiments using driving simulators or test tracks, with the caveat that the behavior of participating drivers might differ from real-life driving.

Naturalistic driving study (NDS) is an innovative method for traffic safety research in which participants' vehicles are instrumented with an advanced data collection system that continuously records driving information for substantial periods. Typically, driving variables are collected from multichannel video cameras, a global positioning system (GPS), radar, and multidimensional kinematic sensors. For example, the 100-Car NDS included 102 participants from Northern Virginia and continually collected data[2] for 1 year. The NDS provides large amount of precise and objective driving data that allow crashes or other types of safety critical events to be observed and studied.[2-7] Such precise information on safety events, as well as the normal driving conditions between events, creates unprecedented opportunities for evaluating the safety impacts of driver behavior or other time-variant risk factors.

The novelty and complexity of NDS demand appropriate statistical methodology. The large data collected via NDS bring challenges in data analysis, which often involves labor-intensive video data reduction, especially for time-variant risk factors such texting and cell phone use. Sampling-based epidemiological approaches have been used to extract exposure information at the onset of safety critical events and during normal driving condition, eg, the case-cohort method,[5-9] event rate approach,[10] and the case-crossover study method.[11] Tailored statistical study designs and statistical models are typically required to fit unique characteristics of NDS.

Matched case-crossover studies are special cases of matched case-control studies. The study design uses subjects themselves as controls to control for the confounding effects caused by subject characteristics. The key feature of matched case-crossover studies is that the control information for each stratum is based on the subject's exposure experience and the stratifying variable is the individual subject or case. Because comparisons are made within subjects, time-invariant confounders are inherently controlled.[12] The matched case-crossover is best suited for evaluating the risk of time-variant factors with transient effect.[13,14]

It is generally accepted that matching factors cannot exert a confounding effect on independent covariates included in analyses because the stratum effect is removed through conditioning.[15-18] However, it has been shown that an effect modification by matching factors could exist and parametric and semiparametric varying coefficient models have been developed to assess the effects of matching factors.[19,20] One limitation of these models is the inability to detect stratum effect when the number of controls is unequal across strata, which could lead to distinct within- and between-strata variations. Another issue with standard conditional logistic regression models is that inference results depend only on the difference between covariates of case and controls. Strata with equal covariate values between case and controls do not impact the inference results, which could lead to loss of information. For example, if 80% of the data contain equal covariate values, only 20% of the data actually contribute to the inference. Heagerty and Kurland[21] and McCulloch et al[22] investigated the impact of model assumption violation on the maximum likelihood estimate of a regression coefficient in a generalized linear mixed model. They showed that the informative visit process can bias estimators of parameters of covariates associated with the random effects, while allowing consistent estimation of other parameters.

An alternative approach to overcome the limitations of conditional models is the generalized linear mixed model, which treats stratum effect as a random variable.[23] The multilevel mixed model has been applied to longitudinal data analysis in generalized linear mixed-model setups.[24] The parameters from generalized linear mixed models can be estimated with normal distribution assumptions or a retrospective model approach.[25,26] The retrospective model approach, however, relies on a strong distribution assumption of the stratum random effect, eg, normal distribution, and calculating marginal likelihood can be challenging. Furthermore, estimation of nuisance parameters, ie, stratum effect, may cause unknown effects on the estimation of the parameter of interest.

The object of this paper is to evaluate time-variant driving risk factors with a specific focus on secondary driving tasks engagement and drowsiness using the 100-Car NDS data. The study follows an unbalanced matched case-crossover approach. A hierarchical random effect model with stratum effects is proposed. We use marginal conditional probability to estimate parameters for the full marginal models. Semiparametric Bayesian models are developed to account for variations within and between strata without the need to specify the distribution of random variables. We evaluated the theoretical properties of the proposed models and conducted a simulation study to demonstrate the model's performance with commonly used parametric and semi-parametric models. The model is applied to data extracted from the 100-Car NDS data following a case-crossover study design. The paper is organized as follows: The NDS data and case-crossover approach are introduced in Section 2; the model specification is presented in Section 3; Section 4 introduces the simulation study and model comparison; the application to 100-Car NDS and results are presented in Section 5; Section 6 provides the summary and discussion.

2 | THE 100-CAR NDS AND CASE-CROSSOVER APPROACH

The 100-Car NDS is a large-scale NDS conducted in Northern Virginia that included 102 primary drivers.[2] The vehicles of primary drivers were instrumented with a sophisticated data acquisition system that included 5 video cameras, a GPS, front and rear radar, and multidimensional kinematic sensors. The 5 camera views monitored the driver's face and the driver's side of the vehicle, the forward view, the rear view, the passenger side of the vehicle, and an over-the-shoulder view for the driver's hands and surrounding areas, as illustrated in Figure 1. The videos are in digital format, and specialized software was developed to synchronize and overlay variables for analysis. The driving data were collected continuously from ignition-on to ignition-off for 12 months for each driver. These continuously collected driving data provide objective and precise information on driver's behavior and environmental factors for both the onset of safety-critical events, such as crashes, and for normal driving situations. Such data are critical for quantitatively evaluating the risk of driving behaviors and other time-variant factors.

Through a combination of identifying kinematic triggers (eg, high deceleration rate) and visual confirmation, the study identified 830 crashes and near-crashes (CNCs).[2] Driver behavior, such as distraction, has a short transient effect on crash risk. To extract behavior information prior to a CNC event, trained data reductionists visually examined the short video clips (approximate 6 s) from 5 seconds before the precipitating event, the state of environment and action that start a CNC, of a CNC till the end of event. Dozens of variables related to the driver, traffic, road, and vehicles were recorded. A rigorous data reduction protocol was implemented to ensure the quality of data.[11] These reduced variables represent the exposure to various risk factors at the onset of crashes and near-crashes.

To assess the relative risk of potential risk factors, exposure information under normal driving conditions is also needed. One main challenge in analyzing large NDS data sets is how to extract information from a large amount of video data from normal driving, eg, the 100-Car study collected approximate 43,000 hours of continuous driving data. The reliable approach based on current technology is the visual examination of recorded videos. The manual data reduction process, however, is time-consuming and cost prohibitive to be implemented for all trips. One solution is to use a sampling-based approach, eg, case-cohort and case-crossover methods.[4-8,11] Analogue to classic epidemiological study designs, these approaches are based on a set of safety-critical events and a set of selected controls, which consist of samples of short driving epochs. The control driving segments are typically short to be compatible with the duration of data reduction for crashes, eg, 6 seconds. The sampling-based methods make it feasible to extract the exposure information under normal driving condition by visually examining videos.

FIGURE 1 100-Car naturalistic driving study video camera views (driver face view is blurred for publication purpose)

The advantage of the case-crossover approach lies in its ability to control for potential confounding factors via sampling. In the context of this study, the controls are short 6-second driving windows during non-case, normal driving condition. A predetermined number of baselines were sampled from normal driving conditions for each CNC based on specific matching criteria. These potential baseline epochs were sampled from the time period before the occurrence of the CNC to ensure that the CNC involvement did not alter or affect subsequent driving performance. The sampling schedule is illustrated in Figure 2.

The data used in this study are from a project sponsored by the National Highway Traffic Safety Administration.[11] The data reduction involves both secondary task engagement and eye-glance behavior. An attempt was made to identify up to 15 matched baselines for each CNC. The goal was to locate baselines that matched (as closely as possible) the conditions present during the event. The baseline episodes were required to meet the following matching and data quality criteria:

1. The same participant must be driving that was driving in the event.
2. The baseline must occur during the same general day of the week (weekend versus weekday).
3. The baseline must occur during the same time of day (event time ± 2 h).
4. The baseline must occur at or near the same location (match GPS within 100 m OR match relation to junction).
5. The baseline must occur at a date/time prior to the date/time of the event.
6. The necessary video must be present in order to perform the analysis (forward view and face views were required to be present).
7. All baselines for a given event needed to occur in different trips.
8. If two baselines for two different events occurred in the same file, they could not overlap by more than 10 seconds.

The last criterion is imposed to avoid two baselines being reduced from the same video segment. Software engineers queried the 100-Car study database for potential baseline epochs followed by visual confirmation. Regardless of the substantial efforts to identify controls, a considerable number of CNCs have less than 15 baselines, which led to an unbalanced data set with nonequally sized strata. The distribution of the number of baselines per CNC is shown in Figure 3.

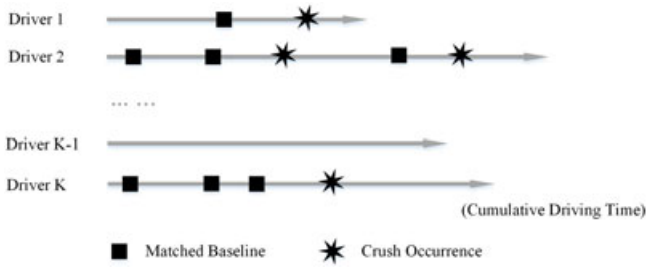

FIGURE 2 Case-crossover baselines [Colour figure can be viewed at wileyonlinelibrary.com]

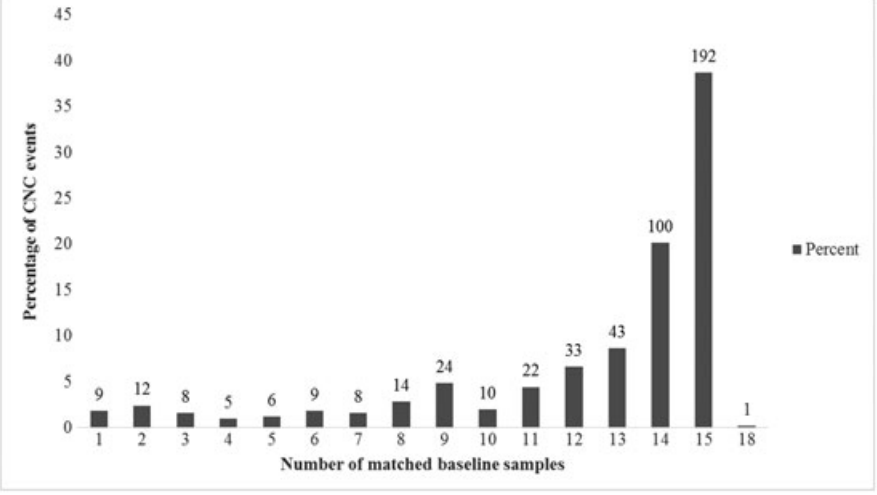

FIGURE 3 Distribution of number of baselines for at-fault crashes and near-crash (CNC) events

TABLE 1 Contingency table

			CNC	Baseline
Distraction	Cell	Dialing	14	60
		Headset	0	62
		Reach	2	61
		Talking	33	421
	Radio		17	598
	Drinking		4	136
	Eating		16	204
	AdjInVeh		4	118
	Normal driving		339	3551
Drowsiness	Drowsy		97	29
	Normal driving		339	3309

To control for the effects of risk factors not controlled by subject drivers, we focused on at-fault or potential at-fault CNCs only. Drowsiness was analyzed separately since it can interact with other types of secondary tasks. The evaluation of a risk factor is with respect to normal driving only, ie, CNC or baseline epochs in which the driver did not engage in any secondary tasks or known risky behavior. The contingency table for the secondary tasks and drowsiness is shown in Table 1.

3 | MODELS

The analysis of a case-crossover study can be viewed as a stratified analysis of conditional, self-matched follow-up studies, each with a sample size of one.[27] The regular condition logistic regression removes the stratum effect by conditioning on the summation of the response variable within the stratum. The standard conditional model, however, cannot catch the unbalanced stratum effects for unbalanced data where the number of controls varies by stratum.

Consider a binary response variable, Y_{ij} for the jth measurement of the ith stratum and p covariate variables $\mathbf{X}_{ij}$, for $j = 1, \ldots, m_i + 1$, and $i = 1, \ldots, N$. For a 1-case-m_i controls design, $Y_{i1} = 1$: case; $Y_{ij} = 0$: control; $j = 2, \ldots, m_i + 1$, and $\sum_{j=1}^{m_i+1} Y_{ij} = 1$. We model the case status Y_{ij} as a function of the covariate X_{ij} in a prospective model form as follows. For simplification, model with a single covariate ($p = 1$) is presented as follows,

$$\begin{aligned} \Pr(Y_{ij} = 1 | X_{ij}, \beta_{0ij}) &= g(\beta_{0ij} + x_{ij}\beta), \\ \beta_{0ij} &= s_i + q_{ij}, \end{aligned} \tag{1}$$

where s_i is a stratum random effect and q_{ij} is a within a stratum random effect. Denote the joint distribution of $(s_i, q_{i1}, q_{i2}, \ldots, q_{i,m_i+1})$ as $F(s, q)$. This two-level mixed model can catch the within- and between-strata effects for unbalanced case-crossover data. We assume that the s_i are independent of the $s_{i'}$ ($s_i \perp\!\!\!\perp s_{i'}$) and q_{ij}s are correlated within the ith stratum ($q_{ij} \not\perp\!\!\!\perp q_{ij'}$). In traditional multilevel mixed model for longitudinal data analysis,[24,28] s_i and q_{ij} are typically assumed to have expectation zero and variances σ_s^2 and σ_i^2, respectively, that is $s_i \sim [0, \sigma_s^2]$, $q_{ij} \sim [0, \sigma_i^2]$. We also assume that that the random effects are independent of covariate X values. Define $\mathbf{q}_i = (q_{i1}, q_{i2}, \ldots, q_{i,m_i+1})$. Note that this two-level mixed model can be found in the literatures related to generalized linear mixed models for longitudinal data analysis.[23-25,28,29] This model can be easily expanded for multiple covariate cases by replacing x_{ij} to $\mathbf{x}_{ij}$ and β to $\boldsymbol{\beta}$. The key for the model fitting is to obtain the following marginal probability

$$\Pr(Y_{i1}, \ldots, Y_{i,m_i+1} | X_{i1}, \ldots, X_{i,m_i+1}, \sum_{j=1}^{m_i+1} Y_{ij} = 1).$$

We do not specify the distribution F but estimate F via a semiparametric Bayesian approach. We considered two approaches for the above marginal probability. The first approach is based on $[Y|X]$, the marginal probability of $[Y_i|X_i]$. The second approach is based on $[[Y|X, s, q]$, the full conditional probability of $[Y_i|X_i, s_i, q_i]$. The second approach has the advantage of being easy to be implement computationally. In the following sessions, we show that under certain conditions these two approaches are approximately equivalent.

3.1 | Marginalized conditional probability of $[[Y|X]]$

Under the mixed model (1), the marginal probability of the response in the ith stratum can be written as

$$\Pr(Y_{i1},\ \dots\ ,Y_{i,m_i+1}|X_{i1},\ \dots\ ,X_{i,m_i+1}) = \int \prod_{j=1}^{m_i+1} \left\{g(x_{ij}\beta + s_i + q_{ij})\right\}^{y_{ij}} \left\{1 - g(x_{ij}\beta + s_i + q_{ij})\right\}^{1-y_{ij}} dF(s_i, \mathbf{q}_i), \tag{2}$$

where $\mathbf{q}_i = (q_{i1}, q_{i2},\ \dots\ , q_{i,m_i+1})$. The marginal conditional probability for stratum i can be calculated as

$$\begin{aligned}
&\Pr(Y_{i1} = 1|X_{i1},\ \dots\ ,X_{im_i},\ \sum_{j=1}^{m_i+1} Y_{ij} = 1)\\
&= \frac{\int g(s_i + q_{i1} + x_{i1}\beta)\prod_{j\neq 1}\{1 - g(s_i + q_{ij} + x_{ij}\beta)\}dF(s_i, \mathbf{q}_i)}{\int \sum_{j=1}^{m_i+1} g(s_i + q_{ij} + x_{ij}\beta)\prod_{j'\neq j}\{1 - g(s_i + q_{ij'} + x_{ij'}\beta)\}dF(s_i, \mathbf{q}_i)}\\
&= \frac{\int g(s_i + q_{i1} + x_{i1}\beta)\prod_{j\neq 1}\{1 - g(s_i + q_{ij} + x_{ij}\beta)\}dF(s_i, \mathbf{q}_i)}{\sum_{j=1}^{m_i+1} \int g(s_i + q_{ij} + x_{ij}\beta)\prod_{j'\neq j}\{1 - g(s_i + q_{ij'} + x_{ij'}\beta)\}dF(s_i, \mathbf{q}_i)}.
\end{aligned}$$

Under logit link function $g(\cdot)$, a conditional probability can be expressed as

$$\Pr(Y_{i1} = 1|X_{i1},\ \dots\ ,X_{im_i},\ \sum_{j=1}^{m_i+1} Y_{ij} = 1) = \frac{1}{\left(1 + \sum_{j=2}^{m_i+1} \exp[-\{(X_{i1} - X_{ij})\beta)\}] \dfrac{\int \dfrac{\exp(q_{ij})}{\prod_{j=1}^{m_i+1} 1+\exp(\beta x_{ij} + s_i + q_{ij})} dF(s_i, \mathbf{q}_i)}{\int \dfrac{\exp(q_{i1})}{\prod_{j=1}^{m_i+1} 1+\exp(\beta x_{ij} + s_i + q_{ij})} dF(s_i, \mathbf{q}_i)}\right)}.$$

In the absence of stratum effects, ie, $\mathbf{q}_i = (0,\ \dots\ , 0)$, the marginal conditional probability reduces to a standard conditional logistic regression,

$$\Pr(Y_{i1} = 1|X_{i1},\ \dots\ ,X_{im_i},\ \sum_{j=1}^{m_i+1} Y_{ij} = 1) = \frac{1}{\left(1 + \sum_{j=2}^{m_i+1} \exp[-\{(X_{i1} - X_{ij})\beta)\}]\right)}.$$

"We note that 'marginal' refers to the integral of random stratum effects and 'conditional' refers to the probability given $\sum_{j=1}^{m_i+1} Y_{ij} = 1$." In the presence of stratum effects, the marginal probability involves multidimensional integration over the space of $(s_i, \mathbf{q}_i)$, which does not have a closed-form solution. In Section 3.2, we develop an alternative approach to obtain the marginal probability.

3.2 | Marginalized conditional probability using $[[Y|X, s, q]]$

The conditional probability based on the logit link function given $(s_i, \mathbf{q}_i)$ can be expressed as

$$\Pr(Y_{i1} = 1|X_{i1},\ \dots\ ,X_{im_i},\ \sum_{j=1}^{m_i+1} Y_{ij} = 1, s_i, \mathbf{q}_i) = \frac{1}{\left(1 + \sum_{j=2}^{m_i+1} \exp[-\{(X_{i1} - X_{ij})\beta + (q_{i1} - q_{ij}))\}]\right)},$$

which does not depend on s_i but q_{ij}.

By letting $\mathbf{r}_i = (q_{i1} - q_{i2}, q_{i1} - q_{i3},\ \dots\ , q_{i1} - q_{im_i+1}) = (r_{i1}, r_{i2},\ \dots\ , r_{i,m_i})$ and $H(\cdot)$ be the probability distribution function of $\mathbf{r}_i$, we have the following expression,

$$\Pr(Y_{i1} = 1|X_{i1},\ \dots\ ,X_{im_i},\ \sum_{j=1}^{m_i+1} Y_{ij} = 1) = \int \frac{1}{\left(1 + \sum_{j=2}^{m_i+1} \exp[-\{(X_{i1} - X_{ij})\beta + (q_{i1} - q_{ij}))\}]\right)} dH(\mathbf{r}_i). \tag{3}$$

It can be shown that the marginal conditional probability based on the second approach is approximately equivalent to the first one under certain conditions. The result is summarized in Lemma 1.

Define $h(s_i, q_{ij}) = \log\{p(Y|s_i, q_{ij})\}/(-N)$ and $\tilde{\Sigma} = [\frac{\partial^2 h(s_i,q_{ij})}{\partial q_{ij}\partial q_{kl}}]^{-1}|_{s_i=\hat{s}_i, q_{ij}=\hat{q}_{ij}}$, where $\hat{s}_i$ and $\hat{q}_{ij}$ are the modes or maximum likelihood estimators of $h(s_i, q_{ij})$. We also define $g(s_i, q_{ij})$ as

$$\int \frac{\exp(q_{ij})}{\prod_{j=1}^{m_i+1} 1 + exp(\beta x_{ij} + s_i + q_{ij})} dF(s_i, \mathbf{q}_i) = E\{g(s_i, q_{ij})\}$$

and $\exp\{-Nh^*(s_i, q_{ij})\} = \exp[\log\{g(s_i, q_{ij}) - Nh(s_i, q_{ij})\}]$.

Lemma 1.

(i) Marginal conditional probabilities using $[Y_i|X_i]$ and $[Y_i|X_i, s_i, q_i]$ are approximately equivalent in the order of $O(1/N)$.

(ii) If $p(y|q_{i1}) \approx p(y|q_{ij}) \approx \ldots \approx p(y|q_{im_{i+1}})$, marginal conditional probabilities using $[Y_i|X_i]$ and $[Y_i|X_i, s_i, q_i]$ are approximately equivalent in the order of $O(1/N^2)$.

(iii) If $\frac{|\Sigma_r^*|^{1/2} exp(-nh(q_{i1}^*-q_{i2}^*,\ldots,q_{i1}^*-q_{im_i}^*))}{|\tilde{\Sigma}_r|^{1/2} p(Y|\hat{q}_{i1}-\hat{q}_{i2},\ldots,\hat{q}_{i1}-\hat{q}_{im_{i+1}})\pi(\hat{q}_{i1}-\hat{q}_{i2},\ldots,\hat{q}_{i1}-\hat{q}_{im_{i+1}})} \approx [1 - \sum_j \alpha_{ij} D_{ij}^q \{\frac{exp(D_{ij}^x + D_{ij}^q)}{1+exp(D_{ij}^x + D_{ij}^q)}\}]$, the two approaches are equivalent in the order of $O(1/N^2)$, where $D_{ij}^x = -(X_{i1} - X_{ij})$ and $D_{ij}^q = -(q_{i1} - q_{ij})$.

Further definitions of α_{ij}, q_{ij}^*, and Σ_r^* are explained in Appendix S1. The proof of Lemma 1 can be shown using the first- and second-order Laplace approximations. An outline of proof is summarized in Appendix S1.

Compared to the standard conditional logistic regression, the marginal conditional approach based $[Y|X, s, q]$ can reduce the bias. The conclusion is summarized in Lemma 2. In addition, we assessed the effects of misspecification of random effect distribution on parameter estimation as summarized in Lemma 3.

Lemma 2.

(i) If $\beta = 0$ in conditional model (3) with $H(r)$, both the proposed two-level mixed effect model and the standard conditional logistic model provide consistent estimates.

(ii) If $|\beta| \neq 0$, the estimator β^* obtained from the standard conditional logistic likelihood is biased toward the null, (ie, $|\beta^*| < |\beta|$). The proposed marginal conditional probability using $[Y|X, s, q]$ can reduce the bias.
We also derived the asymptotic relative bias,which is summarized in Appendix S2. The proof follows Pfeiffer et al.[26]

Lemma 3.

(i) If $\beta = 0$ in conditional model (3), then maximum likelihood estimator $\hat{\beta}^*$ based on either (1) standard logistic regression without random effect $H(r)$, or (2) a misspecified random effect distribution, $H^*(r)$, consistently estimates zero.

(ii) If $\beta \neq 0$, misspecification of the random effects distribution results in inconsistent estimators.

We can show this result by using Neuhaus et al.[23] The outline of proof is presented in Appendix S3.

Lemma 3(ii) highlights the importance of properly specifying random effects distribution. To minimize the potential bias caused by misspecification, we propose a semiparametric Bayesian approach that does not rely on a specific parametric distribution for the random effects. A Dirichlet process prior is used for the unknown distribution of random effect. This model provides robustness and flexibility by allowing multimode distributions for the random effects.

3.3 | Semiparametric Bayesian approach

The proposed semiparametric Bayesian approach does not specify the distribution of r_i, where $r_i = (r_{i1}, \ldots, r_{im_i})$. We assume that r_i follows unknown distribution G from a Dirichlet process (DP) prior,[30,31]

$$r_i|G \sim G,$$
$$G|c_0, G_0 \sim DP(c_0, G_0),$$

where $c_0 > 0$ is a precision or total mass parameter and G_0 is a base measure. We set G_0 as $N(\mu_r, \theta_r)$. The scalar c_0 reflects the prior belief about the similarity between the non-parametric distribution G and the base measure G_0.

To complete the model specification, μ_r is set as 0 and independent hyperpriors are set as follows: $c_0|a_0, b_0 \sim \text{Gamma}(a_0, b_0)$, $\beta|\beta_0, \Sigma_\beta \sim N(\beta_0, \Sigma_\beta)$, and $\theta_r \sim IG(a_r/2, b_r/2)$. The precision or total mass parameter, c_0, of the DP prior can be assumed to be either random or a fixed value. In the random case, c_0 is assumed to follow $\text{Gamma}(a_0, b_0)$ distribution. The inverse gamma prior is parameterized as $E(\theta_r) = b_r/(a_r - 1)$. The complete conditional distribution is proportional to

$$[\beta|X, r_i, \theta_r] \propto \prod_i \frac{1}{1 + \sum_{j=1}^{m_i} \exp[-\{(X_{i0} - X_{ij})\beta + r_i\}]} h(r_i),$$

$$[r_i|X, \beta, \theta_r, \{r_l, l \neq i\}] \propto \sum_{l \neq i} f(r_l|\beta, \theta_r)\delta_{r_l} + c_0 \left\{ \int f(r_i|\beta, \theta_r)h(r_i)dr_i \right\} h(r_i)f(r_i|\beta, \theta_r),$$

where δ_{r_l} is a point mass at r_l.

Markov Chain Monte Carlo (MCMC) method was used to sample from posterior distributions. With probability proportional to $c_0\{\int f(r_i|\beta, \theta_r)h(r_i)dr_i\}$, r_i is sampled from $h(r_i)f(r_i|\beta, \theta_r)$; with probability proportional to $f(r_l|\beta, \theta_r)$, r_i is sampled from δ_{r_l}, which means that $r_i = r_l$ with probability 1. Therefore, r_i has the mixture distribution that can be constructed by the mixture of G_0 and point mass densities.

After sampling the random effects for each stratum, the strata were grouped based on the values of the r_i. After selecting a new r_i for each stratum i, there are k, $0 < k \leq N$ unique r_i values among the r_i's. Denote theses unique values by γ_l, $l = 1, \ldots, k$, where l represents the set of strata with common random effect γ_l. Note that knowing the values of random effects is equivalent knowing k, all of γ_l, all the γ_l's and the stratum memberships l. The γ_ls are k independent observations from $N(0, \theta_r)$. Let $\gamma = (\gamma_1, \ldots, \gamma_k)'$ we have

$$[\theta_r|\beta, \gamma] \propto IG(\frac{k}{2} + a_r, \frac{\sum_l^k \gamma_l}{2} + \frac{1}{b_r}).$$

Since $\{\int f(r_i|\beta, \theta_r)h(r_i)dr_i\}$ does not have a closed form, we calculate this integration by taking a Taylor expansion as follows:

$$\begin{aligned}\left\{ \int f(r_i|\beta, \theta_r)h(r_i)dr_i \right\} &= E_r \left\{ f(\beta, 0) + r_{ij}\frac{\partial f_i(\beta, r_{ij})}{\partial r_{ij}} + \frac{1}{2}r_{ij}^2\frac{\partial^2 f_i(\beta, r_{ij})}{\partial r_{ij}^2} + \frac{1}{3!}r_{ij}^3\frac{\partial^3 f_i(\beta, r_{ij})}{\partial r_{ij}^3}\frac{1}{4!}r_{ij}^4\frac{\partial^4 f_i(\beta, r_{ij})}{\partial r_{ij}^4} + o(r_{ij}^4) \right\} \\ &= l_{ij0} + \frac{1}{2}\theta_i\frac{\partial^2 f_i(\beta, r_{ij})}{\partial r_{ij}^2} + \frac{1}{4!}2\theta_i^2\frac{\partial^4 f_i(\beta, r_{ij})}{\partial r_{ij}^4} + o(\theta_i^2),\end{aligned}$$

where $l_i(r_i)$ is the log density, $l_i(\mathbf{r}_i) = \log\{f_{i\cdot}(\mathbf{r}_i)\}$ and $l_i^{(k)} = \partial^k l_i(\mathbf{r}_i)/\partial \mathbf{r}_i^k$.

A total of 25 000 posterior samples were used for inference after 100 000 burn-in period. The hyperparameter values are set as $\beta_0 = 0$, $\Sigma_\beta = 100$, $\Sigma_{0r} = 100$, $a_r = 4$, $b_r = 2$, $a_0 = 4$, and $b_0 = 1/200$. We also considered $\mu_r|\mu_{0r}, \Sigma_{0r} \sim N(\mu_{0r}, \Sigma_{0r})$ with the same model specification of other hyperpriors. In this case, we set $\mu_{0r} = 0$ and $\Sigma_{0r} = 100$.

Gibbs sampler was used to sample[27] c_0. The c_0 was sampled from conditional distribution $[c_0|X, \beta, \theta_r, r_i, k]$. The prior distribution for the number of clusters, k, can be written as follows[32]:

$$p(k|c_0, N) = c_N(k)N!c_0^k\frac{\Gamma(c_0)}{c_0 + N}, \ k = 1, 2, \ldots, N,$$

where $c_N(k) = P(k|c_0 = 1, N)$ and it does not involve c_0.

Similar to West,[27] we assume $p(c_0|X, \beta, \theta_r, r_i, k) \propto p(c_0)P(k|c_0)$, where $p(c_0)$ is prior density. Suppose that $c_0 \sim \text{Gamma}(a_0, b_0)$, the gamma functions can be written as

$$\frac{\Gamma(c_0)}{\Gamma(c_0 + N)} = \frac{(c_0 + N)\text{Beta}(c_0 + 1, N)}{c_0\Gamma(N)},$$

where Beta is the beta function. The following results can be obtained:

$$p(c_0|k) \propto p(c_0)c_0^{k-1}(c_0 + N)\text{Beta}(c_0 + 1, N) \propto p(c_0)c_0^{k-1}(c_0 + N)\int_0^1 z_0^c(1 - z)^{N-1}.$$

This implies that $p(c_0|k)$ is the marginal distribution from a joint for c_0 and a continuous quantity z $(0 < z < 1)$ such that

$$p(c_0, z|k) \propto p(c_0)c_0^{k-1}(c_0 + N)z^{c_0}(1 - z)^{N-1}.$$

Following above derivation, we get the conditional posterior $p(c_0|z,k)$ and $p(z|c_0,k)$. Under Gamma(a_0,b_0) prior for c_0, the conditional posterior distribution can be expressed as a mixture of two gamma posteriors. The conditional distribution of the mixing parameter given c_0 and k is a simple beta distribution,

$$p(c_0|z,k) \propto \pi_z G(a_0+k, b_0-\log(z)) + (1-\pi_x)G(a_0+k-1, b_0-\log(z)),$$

where weight π_z is defined by $\pi_z/1-\pi_z = (a_0+k-1)/N(b_0-\log(z))$. We also have $p(z|c_0,k) \propto z^{c_0}(1-z)^{N-1}$ $(0<z<1)$ so that $(z|c_0,k) \sim$ Beta(c_0+1,N), a beta distribution with mean $(c_0+1)/(c_0+N+1)$.

Based on the above argument, c_0 can be sampled as follows: at each MCMC iteration, (1) sample z value from beta distribution, conditional on c_0, and k fixed at their most recent values; and (2) sample a new c_0 value from the mixture of gammas based on the same k and the z values just generated.

4 | SIMULATION STUDY AND MODEL COMPARISON

We conducted a simulation study to compare the following 5 alternative methods:

- CMlogit: standard conditional logistic regression model;
- QLlogit: quasi-likelihood approach for fitting the conditional logistic mixed model;

TABLE 2 The average mean square error values of β

Case	Dist. of q_{ij} and m_i	Method	Est$(\hat{\beta})$	MSE	Bias^2	Var
	$q_{ij} \sim N(0,1)$	CMlogit	0.9993	0.0877	0.0403	0.0474
	$m_i \sim Unif\,[5,16]$	QLlogit	1.3646	0.0695	0.0270	0.0425
		MCEMlogit	1.1041	0.0538	0.0092	0.0446
1		pBaylogit	1.1885	0.0522	0.0091	0.0431
		sBaylogit	1.1371	0.0584	0.0149	0.0435
	$q_{ij} \sim N(0,3^2)$	CMlogit	0.8544	0.1299	0.1194	0.0105
	$m_i \sim Unif\,[5,16]$	QLlogit	1.0751	0.0909	0.0387	0.0549
		MCEMlogit	1.0915	0.0815	0.0293	0.0522
2		pBaylogit	1.0977	0.0785	0.0234	0.0551
		sBaylogit	1.1357	0.0802	0.0271	0.0531
	$q_{ij} \sim gamma(1,1)$	CMlogit	1.4650	0.1482	0.1105	0.0323
	$m_i \sim Unif\,[5,16]$	QLlogit	1.4112	0.1099	0.0749	0.0353
		MCEMlogit	1.0298	0.1057	0.0699	0.0381
3		pBaylogit	1.1133	0.1043	0.0674	0.0369
		sBaylogit	1.1883	0.0552	0.0250	0.0302
	$q_{ij} \sim gamma(1,1/2)$	CMlogit	1.5195	0.1443	0.1065	0.0378
	$m_i \sim Unif\,[5,16]$	QLlogit	1.5526	0.1281	0.1043	0.0238
4		MCEMlogit	1.3405	0.1044	0.0796	0.0248
		pBaylogit	1.3337	0.1038	0.0779	0.0201
		sBaylogit	1.1842	0.0472	0.0217	0.0255
		sBayprobit	1.1640	0.1222	0.1009	0.0213
	$q_{ij} \sim gamma(1/2,1)$	CMlogit	1.4217	0.2666	0.2061	0.0605
	$m_i \sim Unif\,[5,16]$	QLlogit	1.4911	0.2189	0.1172	0.0847
5		MCEMlogit	1.4735	0.2041	0.1210	0.0831
		pBaylogit	1.3999	0.2018	0.1283	0.0735
		sBaylogit	1.1715	0.0653	0.0315	0.0338

This simulation generated 1000 matched sets. X_{ij} was generated from Gaussion (0,1) distribution and Y_{ij} was generated from Bernoulli distribution with the following probability $Pr(Y_{ij}=1|X_{ij}) = 1/\{1+\exp(-\beta X_i+\beta_{0ij})\}$, $\beta_{0ij} = s_i + q_{ij}$, where $\beta = 1.2$; CMlogit, conditional model with logit link; QLlogit, quasi-likelihood for conditional mixed model with logit link; MCEMlogit, Monte Carlo expectation maximization method for conditional mixed model with logit link; pBaylogit, parametric Bayesian approach for conditional mixed model with logit link; sBaylogit, semiparametric Bayesian approach for conditional mixed model with logit link.

- MCEMlogit: MCEM approach for fitting the conditional logistic mixed model;
- pBaylogit: parametric Bayesian approach for fitting the conditional logistic mixed model; and
- sBaylogit: semiparametric Bayesian approach for fitting the conditional logistic mixed model.

A Gaussian distribution was used for the random effects of the parametric Bayesian model (PBaylogit). Quasi-likehood approach (QLogit) and Monte Carlo Expectation-Maximization (MCEMlogit) approach estimate the parameters of the conditional likelihood using only the mean and variance of the random effects.

The quasi-likelihood approach (QLogit) approximates the marginal likelihood of (β, σ_q^2) using a Taylor expansion of $L(\beta, \sigma_q^2)$ with respect to the random term $\mathbf{r}$. The marginal likelihood $L(\beta, \theta)$ can be calculated by integrating out $\mathbf{r}$, where $\theta = E(r_i^2) = E(r_l^2)$. The MCEM approach (MCEMlogit) is an iterative procedure that consists of the Monte Carlo (MC) step which samples r_{ij} from the complete conditional distribution $[r_{ij}|X, \beta, \theta]$, the expectation (E) step which calculates the expectation using MC samples, and the maximum (M) step, which estimates the fixed-effect parameters. The sBaylogit model is semiparametric Bayesian models proposed in this study using logit functions.

The simulation study is based on an unbalanced 1-m_i matched case-crossover study. The m_i was generated from uniform [2,16], and 1,000 matched sets were simulated. X_{ij} was generated from Gaussian (0,1) distribution and Y_{ij} was generated from the generalized linear mixed model with logit link

$$Pr(Y_{ij} = 1|X_{ij}, \beta_{0ij}) = \frac{1}{1 + \exp\{-(\beta X_{ij} + \beta_{0ij})\}}$$
$$\beta_{0ij} = s_i + q_{ij},$$

We set $\beta = 1.2$.

The parameter s_i was generated from the Gaussian distribution $N(0, 1)$. The q_{ij} was generated from the following 5 cases—case 1: $q_{ij} \sim N(0, 1)$, case 2: $q_{ij} \sim N(0, 3^2)$, case 3: $q_{ij} \sim$ Gamma$(1, 1)$, case 4: $q_{ij} \sim$ Gamma$(1, 1/2)$, and case 5: $q_{ij} \sim$ Gamma$(1/2, 1)$. These 5 cases were used to assess the implication of the lemma. For CMlogit, we used the normal distribution to estimate CMlogit. Cases 1 and 2 were chosen to investigate accuracy of alternative methods as the variance of the random stratum variable increases as well as the loss of efficiency for the semiparametric Bayesian approach. Cases 3 to 5 were used to evaluate the sensitivity of the models to non-Gaussian random variables with right-skewed distributions and to examine the loss of efficiency for parametric approaches when the Gaussian assumption is violated. For each case and simulated data, parameters were estimated using each of the 6 approaches. We generated 100 data sets for each case.

TABLE 3 The average mean square error values of σ_q

Case	Dist. of q_{ij} and m_i	Method	Est($\hat{\sigma}_q$)	MSE	Bias^2	Var
	$q_{ij} \sim N(0, 1)$	QLlogit	1.0812	0.0219	0.0066	0.0153
1	$m_i \sim Unif\,[5, 16]$	MCEMlogit	1.0700	0.0195	0.0049	0.0146
		pBaylogit	1.0685	0.0190	0.0047	0.0143
		sBaylogit	1.0721	0.0193	0.0052	0.0141
	$q_{ij} \sim N(0, 3^3)$	QLlogit	2.7162	0.1156	0.0805	0.0351
2	$m_i \sim Unif\,[5, 16]$	MCEMlogit	2.7385	0.1015	0.0684	0.0331
		pBaylogit	2.7445	0.1002	0.0653	0.0349
		sBaylogit	2.7362	0.1039	0.0696	0.0343
	$q_{ij} \sim gamma(1, 1)$	QLlogit	1.0911	0.0240	0.0083	0.0157
3	$m_i \sim Unif\,[5, 16]$	MCEMlogit	1.086	0.0186	0.0075	0.0101
		pBaylogit	1.0836	0.0182	0.0070	0.0112
		sBaylogit	1.0714	0.0162	0.0051	0.0111
	$q_{ij} \sim gamma(1, 1/2)$	QLlogit	0.6962	0.0576	0.0385	0.0185
4	$m_i \sim Unif\,[5, 16]$	MCEMlogit	0.6371	0.0489	0.0188	0.0212
		pBaylogit	0.6170	0.0367	0.0137	0.023
		sBaylogit	0.5888	0.0240	0.0079	0.0161
	$q_{ij} \sim gamma(1/2, 1)$	QLlogit	1.2106	0.2396	0.2535	0.0092
5	$m_i \sim Unif\,[5, 16]$	MCEMlogit	1.2096	0.2356	0.2524	0.0039
		pBaylogit	1.2122	0.2384	0.2551	0.0034
		sBaylogit	0.7925	0.0101	0.0073	0.0028

The average mean estimate values and average mean square error (MSE) values were used for model comparison. The average MSE values for β and σ_q^2 are summarized in Tables 2 and 3.

For β, as expected, average MSE values are small except for CMlogit, DLlogit, and sBayprobit. The average MSE values of CMlogit are the largest because this model does not include random stratum effect. The MSE values of QLlogit are the second largest, primarily because the quasi-likelihood uses approximation of marginal likelihood. The MSE values of MCEMlogit, pBaylogit, and sBaylogit are similar to each other. The results imply that a semiparametric method is as efficient as the properly parameterized fully parametric methods.

For estimating σ_q^2, it is clear that the sBaylogit performs universally better than other models in both MSE and bias. The MSE values of all methods in case 2 are larger than those in case 1, which indicates that model performance deteriorates as noise increases. For cases 3 to 5 where the random effects follow non-Gaussian distributions, as expected, the performance of CMlogit, MCEMlogit, and pBaylogit is worse than that of sBaylogit in terms of MSE. It can also be seen that the difference in MSE values of sBaylogit and sBayprobit is relatively small in case 5, which implies that results are less sensitive to the choice of link functions.

In summary, the simulation studies indicate that the quasi-likelihood approach performs worse than others in terms of MSE values. The semiparametric Bayesian approach performs consistently among the most efficient alternative models and the advantages are most pronounced when the distribution of the random effect is non-Gaussian.

5 | APPLICATION

We applied the proposed semiparametric Bayesian model with logit link function to the case-crossover data set from the 100-Car NDS as introduced in Section 2. The 100-Car case-crossover data are 1 to M_i unbalanced matched controls, where the number of matched controls varies $1 \leq M_i \leq 15$. The estimation of σ_q indicates the variation within stratum. As indicated in Lemma 3(ii), the semiparametric Bayesian models provide robust and consistent estimates since they do not rely on the distribution assumption of the random effect. The model-fitting results are summarized in Table 4. The average number of clusters for most models was around 11, which confirms the presence of random stratum effects. This result also suggests that the distribution of random stratum effects was from a mixture distribution. Note that as c_0 becomes larger, the nonparametric mixture distribution G is closer to the parametric distribution G_0. Since the estimated c_0 is small overall, this mixture distribution could not be sufficiently estimated using other parametric alternatives. The results confirm the necessity of using the proposed nonparametric approach for this imbalanced case-crossover data.

The result in Table 4 is based on univariate analysis. Since all of our variables were binary variables, we fit the model by each variable. The model outputs indicate that several driver behavior factors had a significant impact on driving risk. Among these engaging cellphone dialing will driving increases the CNC risk by more than two times (odds ratio [OR]: 2.37, 95% CI, 1.30-4.30). The risk of cell phone use while driving has been a key research area for decades.[6-8,14] Using cell phone and hospital records, Redelmeier and Tibshirani's case-crossover study showed that general cell phone use while driving increase the crash risk four times.[14] Due to lack of detailed subtask information, the risk estimated was with respect for general cell phone use and driving risk was measured by injury crashes. Using the same 100-Car study as this paper but a case-cohort approach, Klauer et al estimated a similar OR of 2.49 for cell phone dialing,[7] which Dingus et al[6] estimated a much large OR of 12.2. The Dingus study, however, was based on a case-cohort study design and a large-scale study, the Second Strategic Highway Research Plan (SHRP2) NDS. Only property damage or more severe crashes were used in risk assessment. All these studies, confirm that engaging in cell phone dialing would undoubtedly contribute to the increasing in driving risk.

Drowsiness increases the CNC risk more than 30 times (OR: 33.56, 95% CI, 21.82-52.19). This number is considerably higher than the 3.4 reported by Dingus et al.[6]

Several factors show inverse effects on driving risk: drinking (OR: 0.27, 95% CI, 0.09-0.77), adjust radio (OR: 0.29, 95% CI, 0.17-0.46), and adjust in-vehicle devices (OR: 0.32, 95% CI, 0.11-0.87). Similar results were obtained by Klauer et al[7] with drinking (OR: 0.44, not significant), adjust ratio and HAV (OR: 0.53), and adjust in-vehicle devices (OR: 0.64, P value $> .05$, not significant). However, Dingus et al[6] showed that all these behaviors significantly increase crash risk (drinking OR: 1.8, adjust ratio: 1.9, adjust climate control: 2.3). While there is no rational to support such behavior would lead to safer driving, their relative mild impact on crash and near-crash risk and driver's self-selection on engaging these behavior on relative safety environmental and traffic conditions might lead to the seemingly protective effects.

No significant results were detected for reaching for cell phone, talking on cell phone, and eating. This is, again, consistent with Klauer et al[7] but differs from Dingus et al. The use of different safety critical events, ie, crashes and near-crashes versus server crashes could be the potential underlying reason for this discrepancy.

TABLE 4 Estimated parameters and 95% Bayesian credible intervals obtained by a semiparametric Bayesian model with logit link

Variable Name	G_0	Parameter	Mean	95%Baysian credible interval
Cell phone: dialing	$N(\mu_r, \theta_r)$	Odd ratio	2.38	[1.30, 4.30]
		μ_r	−2.37	[−3.13, −1.57]
		θ_r	0.26	[0.04, 0.71]
		cluster	10.93	[1, 41]
		c_0	2.74	[0.00, 10.79]
Cell phone: reach	$N(\mu_r, \theta_r)$	Odd ratio	0.27	[0.06, 1.13]
		μ_r	−2.37	[−3.12, −1.60]
		θ_r	0.25	[0.04, 0.68]
		cluster	11.90	[1, 46]
		c_0	3.03	[0.01, 12.61]
Cell phone: talking	$N(\mu_r, \theta_r)$	Odd ratio	0.81	[0.56, 1.18]
		μ_r	−2.37	[−3.19, −1.58]
		θ_r	0.27	[0.03, 0.73]
		cluster	9.33	[1, 34]
		c_0	2.30	[0.01, 8.67]
Drinking	$N(\mu_r, \theta_r)$	Odd ratio	0.27	[0.09, 0.77]
		μ_r	−2.37	[−3.12, −1.56]
		θ_r	0.26	[0.04, 0.70]
		cluster	10.70	[1, 40]
		c_0	2.66	[0.01, 10.57]
Eating	$N(\mu_r, \theta_r)$	odd ratio	0.80	[0.47, 1.36]
		μ_r	−2.38	[−3.14, −1.61]
		θ_r	0.26	[0.03, 0.70]
		cluster	10.40	[1, 39]
		c_0	2.59	[0.01, 10.20]
Adjust radio	$N(\mu_r, \theta_r)$	odd ratio	0.29	[0.17, 0.46]
		μ_r	−2.37	[−3.14, −1.61]
		θ_r	0.25	[0.03, 0.69]
		cluster	10.95	[1, 42]
		c_0	2.75	[0.00, 11.43]
Adjust in-vehicle device	$N(\mu_r, \theta_r)$	Odd ratio	0.32	[0.11, 0.87]
		μ_r	−2.37	[−3.17, −1.63]
		θ_r	0.26	[0.04, 0.71]
		cluster	11.44	[1, 44]
		c_0	2.90	[0.01, 12.35]
Drowsiness	$N(\mu_r, \theta_r)$	Odd ratio	33.56	[21.82, 52.19]
		μ_r	−2.29	[−3.08, −1.54]
		θ_r	0.25	[0.04, 0.68]
		cluster	10.48	[1, 40]
		c_0	2.72	[0.00, 10.86]

This simulation generated 1000 matched sets. X_{ij} was generated from Gaussian (0,1) distribution, and Y_{ij} was generated from Bernoulli distribution with the following probability $Pr(Y_{ij} = 1|X_{ij}) = 1/\{1 + \exp(-\beta X_i + \beta_{0ij})\}$, $\beta_{0ij} = s_i + q_{ij}$, where $\beta = 1.2$; QLlogit, quasi-likelihood for conditional mixed model with logit link; MCEM-logit, Monte Carlo Expectation Maximization method for conditional mixed model with logit link; pBaylogit, parametric Bayesian approach for conditional mixed model with logit link; sBaylogit, semiparametric Bayesian approach for conditional mixed model with logit link.

6 | SUMMARY AND DISCUSSION

To quantitatively evaluate the safety impacts of driver behavior and other time-variant factors requires accurate information and large amounts of data. Large-scale NDS provides comprehensive and objectively collected data for this purpose but also brings challenges in data processing and analysis. This paper used a case-crossover approach to extract event and control information from videos. The semiparametric Bayesian models provide a robust and flexible method for analyzing the reduced data.

The matched case-crossover approach can effectively control for confounding factors through self-matching by subject/case. However, the matching also results in complications associated with the matching mechanism. Conventional conditional logistic regression addresses the stratum effect but has drawbacks, eg, the variation among strata cannot be evaluated and the samples with identical covariate values do not contribute to the inference. As an alternative, generalized mixed effects models usually rely on a strong distribution assumption of the random effect and are more difficulty to calculate the marginal likelihood.

This paper contributes to the modeling of a matched case-crossover study by using a two-level mixed-effect approach and providing an alternative way to calculate a marginal conditional probability derived from the full conditional distribution. The model advances traditional methods by incorporating variation over stratum effects. In the conditional model, the stratum variable is treated as a random effect that depends on the subjects in each stratum. We proved that the estimates of parameters of interest in a conditional mixed model are unbiased and consistent. The proposed model can be applied in any type of matched case-control studies. The simulation study suggests that the semiparametric Bayesian approach is more efficient than other methods with respect to the MSE values and allows more flexibility with regard to lesser known situations.

The application of the proposed model to the 100-Car NDS case-crossover study indicates that the parameter co was relatively small, which implies that F is substantially different from the parametric model. Hence, the data supports of the nonparametric approach than the traditional parametric model. Future research is needed to quantitatively evaluate the difference between the proposed nonparametric approach and traditional parametric approach, which relies on the distribution assumption.

The study did not account for the correlations by multiple crashes and near-crashes from the same driver. We considered that the within stratum correlation has a dominant impact on the inference results. Although this correlation can be incorporated by adding a driver-level random effect, it does increase the computation burden and requires a larger sample size. The model can be relatively easily extended to accommodate within subject correlation with a larger data set.

The safety impacts of a specific secondary task might depend on certain environmental factors or other tasks. This is related to the full covariate conditional mean assumption.[24] To evaluate the effects of multiasking and context of distraction would require larger sample size and would be best be addressed by large scale studies such as the SHRP2 NDS.

The results from application to the 100-Car NDS study indicate that engaging in cell phone dialing while driving would increase the driving risk by 2.38 times, and driving under fatigue and drowsiness would increase risk by 33.56 times. These findings further demonstrate the hazard of distraction and impairment on traffic safety. The results have a direct implication on the impacts of driver behavior on traffic safety and provide critical reference for driving safety regulation, vehicle design, and safety education.

Previous analyses using 100-Car NDS data and the case-cohort approach found similar results.[5,7,8] While the point estimates are not identical, the high-risk secondary tasks that significantly increased risk in the analyses presented here are generally the same secondary tasks that significantly increased risk in the previous analysis. Interestingly, the patterns in tasks that increased risk using a case-cohort analysis also follow similar patterns as the results in case-crossover analyses. This similarity across studies indicates repeatability and robust results in these analyses.

The study is based on crashes and near-crashes, a crash surrogate. As indicated by Guo et al,[4] including crash surrogates in risk analysis tends to underestimation of crash risk. The results from the 100-Car NDS and the large-scale SHRP2 NDS with more than 3400 participants confirm this phenomenon.[6,7] However, for many NDS with relatively small sample size, the crash surrogates can still provide critical information for high risk factors.

While the findings were generally similar across alternative studies using the same 100-Car NDS data, there are some differences worth noting. For example, the odds ratio for fatigue (drowsiness) significantly increased risk in all analyses, it was much higher in the case-crossover results. Our hypothesis for these differences in results is that the matched design of the baseline increased the number of baseline epochs that occurred in and around intersections and/or inclement weather. Not only does risk change based upon roadway geometry and weather but we also hypothesize that driver behavior also changes in terms of willingness to engage in secondary tasks. These differences highlight the importance of accounting

for the variability in risk due to roadway geometries, environment, and drivers' willingness to engage in secondary tasks. The case-crossover approach provides better control on these potential confounding factors.

ACKNOWLEDGEMENTS

This study was supported by grants "Further Analysis of the 100 Car Case Crossover Baseline Dataset" (#14-0729-10) from the Global Automakers and the "Developing Bayesian Models for NDS data" from the National Surface Transportation Safety Center for Excellence.

ORCID

Feng Guo http://orcid.org/0000-0002-2572-481X
Inyong Kim http://orcid.org/0000-0002-5975-4582

REFERENCES

1. National Highway Traffic Safety Administration. Traffic safety facts 2013: a complication of motor vehicle crash data from the fatality analysis reporting system and the general estimates system. report number DOT HS: 812-139, National Highway Traffic Safety Administration: Washington D.C.; 2013.
2. Dingus TA, Klauer SG, Neale VL, et al. The100-Car naturalistic driving study: phase II results of the 100-Car field experiment. report number DOT HS: 810-593, National Highway Traffic Safety Administration: Washington D.C.; 2006.
3. Guo F, Fang Y. Individual driver risk assessment using naturalistic driving data. *Accid Anal Prev*. 2013;61:3-9.
4. Guo F, Klauer SG, Hankey JM, Dingus TA. Near-crashes as crash surrogate for naturalistic driving studies the transportation research record. *J Transp Res Board*. 2147;2010:66-74.
5. Guo F, Hankey JM. Modeling 100-Car safety events: a case-based approach for analyzing naturalistic driving data, the National Surface Transportation Safety Center for Excellence; 2009.
6. Dingus TA, Guo F, Lee S, et al. Driver crash risk factors and prevalence evaluation using naturalistic driving data. In: Proceedings of the National Academy of Sciences 113; 2016:2636-41.
7. Klauer SG, Guo F, Simons-Morton BG, Ouimet MC, Lee SE, Dingus TA. The prevalence and risk of cell phone and other secondary tasks as observed in crashes and near-crashes with novice and experienced drivers. *N Engl J Med*. 2014;370:54-59.
8. Guo F, Klauer SG, Fang Y, et al. The effects of age on crash risk associated with driver distraction. *Int J Epidemiol*. 2017;46(1):258-265.
9. Klauer SG, Dingus TA, Neale VL, Sudweeks JD, Ramsey DJ. The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data. Report No. DOT HS 810 594, National Highway Traffic Safety Administration: Washington D.C.; 2006.
10. Fitch GM, Soccolich SA, Guo F, et al. The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk. Report No. DOT HS 811–757, National Highway Traffic Safety Administration: Washington D.C.; 2013.
11. Klauer SG, Guo F, Sudweeks JD, Dingus TA. An analysis of driver inattention using a case-crossover approach on 100-car Data. report number DOT HS: 811-334, National Highway Traffic Safety Administration: Washington D.C.; 2010.
12. Navidi W. Bidirectional case-crossover designs for exposure with time trends. *Biometrics*. 1998;54:596-605.
13. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*. 1991;133:144-53.
14. Redelmeier DA, Tibshirani RJ. Association between cellular-telephone calls and motor vehicle collisions. *N Engl J Med*. 1997;336:453-8.
15. Brewslow NE, Day NE. Statistical methods in cancer research. *The Analysis of Case-Control Studies*. Vol. 1. International Agency on Cancer: Lyon, France; 1980:248-276.
16. Brewslow NE, Zaho LP. Logistic regression for stratified case-control studies. *Biometrics*. 1988;44:891-899.
17. Brewslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika*. 1988;75:11-20.
18. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York: Wiley; 1989.
19. Kim I, Cohen N, Carroll RJ. Effect heterogeneity by a matching covariates in matched case-control studies: a method for graphs-based representation. *Am J Epidemiol*. 2002;156:463-470.
20. Kim I, Cheong H, Kim H. Semiparametric regression models for detecting effect modification in matched case-crossover studies. *Stat Med*. 2011;96:1458-1468.
21. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*. 2001;88:973-985.
22. McCulloch CE, Neuhaus J, Olin RL. Biased and unbiased estimation in longitudinal studies with informative visit processes. *Biometrics*. 2016;72:1315-1324.
23. Neuhausel JM, Hauck W, Kalbfleisch JD. The effects of mixture distribution misspecification when fitting mixed effects logistic models. *Biometrika*. 1992;79:755-7621.
24. Diggle PJ, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. New York: Oxford University Press; 1994.
25. Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, Elston RC. Ascertainment adjustment: Where does it take us? *Am J Hum Genet*. 2000;76:1505-1514.

26. Pfeiffer R, Gail M. Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika*. 2001;84:933-948.
27. West M. Hyperparameter estimation in Dirichlet process mixture models. Technical report 92-A03, Duke University ISDS: Washington D.C.; 1992.
28. Pinherio JC, Bates DM. *Mixed-effects Models in S and S-PLUS*. Statistics and Computing Series. Springer Verlag: New York; 2000.
29. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press; 2007.
30. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann Stat*. 1973;1:209-230.
31. Ferguson TS. Prior distribution on spaces of probability measures. *Ann Stat*. 1974;2:615-629.
32. Antoniak CE. Mixtures of Dirichlet processes with applications to nonparametric problems. *Ann Stat*. 1974;2:1152-1174.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Guo F, Kim I, Klauer SG. Semiparametric Bayesian models for evaluating time-variant driving risk factors using naturalistic driving data and case-crossover approach. *Statistics in Medicine*. 2017;1–15. https://doi.org/10.1002/sim.7574