Accident Analysis and Prevention 108 (2017) 139–146

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Evaluation of risk change-point for novice teenage drivers

Qing Li[a], Feng Guo[b,c,*], Sheila G. Klauer[c], Bruce G. Simons-Morton[d]

[a] Department of Statistics, University of Wisconsin-Madison, United States
[b] Department of Statistics, Virginia Tech, 250 Drillfield Drive, Blacksburg, VA 24061, United States
[c] Virginia Tech Transportation Institute, United States
[d] Eunice Kennedy Shriver National Institute of Child Health and Human Development, United States

ARTICLE INFO

Keywords:
Change-point
Driving risk
Naturalistic Teenage Driving Study
Non-homogeneous Poisson process
Recurrent events

ABSTRACT

The driving risk of novice teenagers is the highest during the initial period after licensure but decreases rapidly. This paper applies two recurrent-event change-point models to detect the time of change in driving risks. The models are based on a non-homogeneous Poisson process with piecewise constant intensity functions. We show that the maximum likelihood estimators of the change-points can only occur at the event times and they are consistent. A simulation study is conducted to demonstrate the model performance under different scenarios. The proposed models are applied to the Naturalistic Teenage Driving Study, which continuously recorded *in situ* driving behaviour of 42 novice teenage drivers for the first 18 months after licensure using sophisticated in-vehicle instrumentation. The results indicate that approximately half of the drivers have lower risk after 73.0 h of independent driving after licensure while the risk for others increases. On the average the driving risk deceases after the change-point. The results provide critical information for safety education, safety countermeasure development, and Graduated Driver Licensing policy making.

1. Introduction

Traffic crashes are the leading cause of death for teenagers in the United States, and the fatality rate for teenage drivers is substantially higher than that for experienced drivers (National Highway Traffic Safety Administration, 2000). Driving risk is typically measured by the rate of safety critical events. Studies like Williams (2003) have shown that the driving risk among novice teenagers is the highest early in licensure, declines rapidly for a period of time right after independent driving, and changes slowly later. The rapid decline in crash risk after licensure has been documented in a number of recent studies (Guo et al., 2013; Lee et al., 2011; Mayhew et al., 2003; Simons-Morton et al., 2011). Reducing the high risk in the initial driving period is the foundation for the Graduated Driver Licensing (GDL) law and also the target of teenage driver safety education and countermeasure programs.

Various factors contribute to the high driving risk of teenage drivers. Jackson et al. (2013) used high G-force events, i.e., hard braking, to predict the occurrence of crash and near-crash (CNC) events and the findings implied that driving style could contribute to the driving risk. Kim et al. (2013) stated that having friends with poor driving habits and risky behaviour increased the driving risk. Klauer et al. (2014) analyzed the Naturalistic Teenage Driving Study (NTDS) data and concluded that performing secondary tasks while driving would significantly increase the driving risk and the impacts on teenagers were considerably higher than experienced drivers. Ouimet et al. (2014) showed that teenage drivers with higher cortisol level tend to have lower CNC risk and the driving risk decrease faster over time. Simons-Morton et al. (2014) concluded that the driving risk for teenage drivers was directly associated with total time eyes off the forward roadway. The high initial risk and rapid change after licensure are likely to be the results of multiple factors including drivers' experience increase and driving skills.

The duration of the initial high risk period is a key parameter of interest. Many GDL laws impose certain constraints on driving privilege, e.g., teen passengers and night time driving, to decrease the likelihood of CNC events. Using a naturalistic driving study approach, Simons-Morton et al. (2011) showed that the CNC rate declined significantly at six months after licensure. Mayhew et al. (2003) showed similar results with the Nova Scotia Motor Vehicle Accident Database. Guo et al. (2013) examined the variation of the CNC rate change across time and revealed substantial variations among teenage drivers: only about thirty percent of drivers at moderate risk showed a significant decrease in CNC rate after six months, while no significant change was found for high and low risk groups.

A common approach for evaluating the time trend of event rate is to aggregate data into pre-defined time intervals, e.g., three-month

* Corresponding author.
E-mail address: feng.guo@vt.edu (F. Guo).

http://dx.doi.org/10.1016/j.aap.2017.08.007
Received 11 May 2017; Received in revised form 10 July 2017; Accepted 4 August 2017
Available online 04 September 2017
0001-4575/

intervals, for analysis (Guo et al., 2013; Simons-Morton et al., 2011). The low safety-critical event rate observed in real life requires a long time period as base analysis unit for stable estimation. The long time interval, however, reduces the time resolution and can only detect the time of change in terms of months or even quarters. In addition, the calendar time interval used in many studies does not necessarily correspond well with the actual driving time, which is arguable to be a more direct measure of driving experience. It is useful to identify the change-points of crash rates in driving time which may quantify the amount of experience required for relatively safe driving (Simons-Morton et al., 2011). For example, according to the GDL of Virginia (Governor's Highway Safety Association, 2015), the teenagers have to drive at least 45 h under supervision to obtain a license. There is a need for more suitable analytical approaches and new data collection method for estimating the risk change in driving time.

The Naturalistic Driving Study (NDS) method is a novel driving data collection method characterized by continuous data collection from advanced in-vehicle instrumentation (Dingus et al., 2006). There is typically no specific instruction for drivers to ensure that the collected data truly reflect their normal driving behaviour. The NDS provides unrepresented *in situ* driving data under natural, non-experimental driving conditions that could not be obtained from crash database and driving simulator/test track studies. It is especially suitable for evaluating the rapid change in teenage driving risks. The Naturalistic Teenage Driving Study (NTDS) conducted between 2006 and 2009 examined the novice teenage driving. The study included 42 teenage drivers who had just obtained their Virginia driver's licenses; their vehicles were equipped with cameras, accelerometers, Global Positioning Systems, and other vehicle-monitoring devices. The crashes and near-crashes were identified from the collected driving data (Lee et al., 2011). This NTDS provides an opportunity to evaluate the change-points of driving risk with higher precision for novice teenage drivers.

Change-point models are a natural fit for detecting the risk change-point. Differing from terminal diseases, majority of crashes or safety critical events are not fatal and most likely will occur again in the future. In this paper, we focus on developing change-point models in the context of recurrent event. There have been a number of change-point detection models based on constant hazard functions but in a non-recurrent-event context (Loader, 1991; Matthews and Farewell, 1982; Nguyen et al., 1984; Yao, 1986). Ghosh et al. (1993) and Karasoy and Kadilar (2007) presented Bayesian estimators of the change-point assuming a piecewise constant hazard function. Mü and Wang (1994) reviewed different change-point estimators in hazard functions. Pons (2002) estimated the change-point in the time-varying covariates in a Cox model by maximizing the likelihood. Wu et al. (2003) proposed a non-parametric change-point estimator of the hazard function in the counting process context allowing for random censoring. Dupuy (2006) provided consistent estimators of the change-point for both hazard function and regression coefficients in a parametric survival regression model.

Piecewise constant intensity function is robust (Lawless and Zhan, 1998), and is the assumption of a large part of literatures on change-point detection in the recurrent-event context. Lawless and Zhan (1998) analyzed the interval-grouped recurrent-event data under such assumption, where only the event counts in given intervals were recorded. Achcar et al. (2007) and West and Odgen (1997) implemented likelihood approaches to estimate the change-point for only one individual with recurrent events. Frobish and Ebrahimi (2009) proposed a maximum likelihood estimator (MLE) and a Nelson-Aalen estimator for estimating the unknown change-point of multiple individuals with recurrent events. Li et al. (2017) proposed a Bayesian finite mixture model to detect the change-points in the driving risk of teenage drivers and cluster the drivers by risk patterns.

Recurrent-event models are commonly based on counting processes (Cook and Lawless, 2007, Chap. 1). A counting process is a non-decreasing stochastic process $\{N(t); t \geqslant 0\}$ with positive integer values, in which $N(t)$ is the cumulative number of events before or at time t. According to the Doob–Meyer decomposition theorem, $N(t) = \Lambda(t) + M(t)$, where $\Lambda(t)$ is the cumulative intensity process and $M(t)$ is a martingale (Andersen et al., 1993). If $\Lambda(t)$ exists, the intensity is $\lambda(t|\pounds_t) = \lim_{\nabla \to 0} P(N[t, t+\nabla) \geqslant 1|\pounds_t)/\nabla$, where $\pounds_t$ is the process history before time t, ∇ is a small time interval, and $N[t, t+\nabla)$ is the number of events between time t and $t+\nabla$.

The Poisson process is canonical for event count analysis, of which the intensity function $\lambda(t|\pounds_t)$ has the form $\lambda(t)$. Given the marginal probability of an incident at time t, $\lambda(t)$ is the intensity function of a Poisson process. The non-homogeneous Poisson process (NHPP) is a Poisson process when the intensity parameter $\lambda(t)$ is not a constant (Ross, 2006, p. 32). Considering the NHPP $\{N(t); t \geqslant 0\}$ with cumulative intensity $\Lambda(t)$, the number of events that occurred between time zero and t is a random variable that is Poisson distributed with parameter $\Lambda(t)$. The intensity function $\lambda(t)$ of a NHPP can increase or decrease abruptly and the shift point in time is the change-point.

This paper advances the likelihood-based approach of Frobish and Ebrahimi (2009) in several ways. First, we relax the constraint that the censoring times be larger than the change-points for all subjects. Second, the intensity functions for individuals could vary, which fits the driving risk change better because of high variability among subjects. Third, we infer the standard errors (SEs) of the intensity estimators by the information matrix in identical-intensity models. The confidence intervals (CIs) and SEs of change-point estimators are obtained by block bootstrapping, where only the drivers are re-sampled with the event times left unchanged (Field and Welsh, 2007). The change-points of the bootstrapped samples are detected by the method in this paper and are used to calculate the CIs and SEs. Lastly, multiple change-points are allowed in our models and we propose a method to determine the number of change-points. All the change-points are assumed unknown, which is different from the approach in Frobish and Ebrahimi (2009) which considered two change-points and assumed the first change-point to be known.

The rest of the paper is as follows. Section 2 introduces the NTDS and presents an exploratory data analysis. Section 3 describes the change-points detection in the context of recurrent-event models. Section 4 reports the simulation findings, and Section 5 shows the NTDS data application results. Section 6 presents concluding remarks.

2. Naturalistic teenage driving methods

The NTDS is a naturalistic driving study focusing on novice teenage drivers. The participants consist of 22 females and 20 males with an average age of 16.4 years. Recruitment of newly-licensed teenagers was conducted via newspaper advertisements, flyers, and driving schools from the New River Valley and Roanoke Valley areas of the Commonwealth of Virginia. The participants were still subject to the GDL when they were in this study. To be eligible for the study, the drivers have to be younger than 17 years old and hold a provisional driver's license with the maximum independent driving time to be three weeks. (Lee et al., 2011). The data collection started within two weeks after licensure. The average driving time over the 18-month study period per participant was 250.6 h and the stand deviation (SD) was 107.5 h.

The driving risk is quantitatively measured by crashes and near-crashes (CNC) rate. A crash involves energy transfer between the participant vehicle and another vehicle or object; a near-crash is similar to a crash except that the driver makes an evasive maneuver to avoid a crash (Lee et al., 2011). It is a state-of-practice to combine crashes and near-crashes in driving risk evaluation mainly because of the relatively small sample sizes in naturalistic driving studies (Guo et al., 2010; Guo and Fang, 2013; Klauer et al., 2010, 2014; Ouimet et al., 2014; Simons-Morton et al., 2011). The validity of using crash and near-crash has been examined in previous studies (Guo et al., 2010).

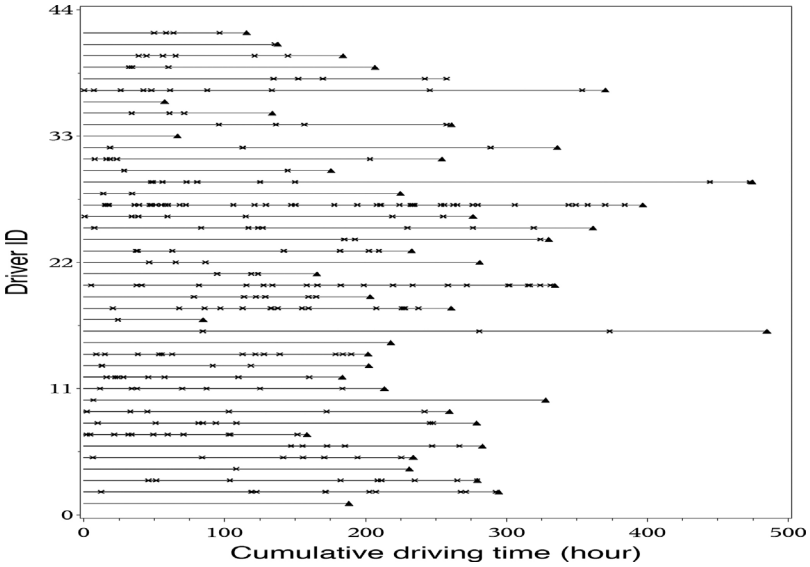

Fig. 1. Event plot of NTDS data. Cross is the time to event, and triangle is the end of study.

A driver might encounter multiple CNC events during the first 18 months after licensure and the average number of CNC events is 6.6 per participant for the NTDS. Fig. 1 shows the event plot by driver. The distribution of events is highly unbalanced among drivers: four participants experienced no safety events while 38 involved in a total of 279 CNC events, including 37 crashes and 242 near-crashes. Fig. 1 also shows that the censoring time varies. The final analysis includes 271 CNC events due to eight missing event times.

The event intervals are based on driving hours. The driving hours, instead of mileage, was used to be consistent with the Graduated Driver Licensing laws, which are typically based on hours of driving. Furthermore, the driving time and distance are highly correlated. Fig. 2 shows the histogram and the boxplot of the inter-event times for all drivers that are heavily skewed to the right. The inter-failure time ranges from 0.002 to 294.4 h with a mean of 28.9 and SD of 38.7 h.

3. Change-point models for recurrent events

This section presents the identical-intensity model and the subject-specific intensity model to detect when crash and near-crash risks decrease significantly for novice teenage drivers.

The models proposed by Frobish and Ebrahimi (2009) assumed that the event occurrence followed a NHPP with a piecewise constant intensity function having two change-points. The number of change-points are usually unknown in practice. To increase the flexibility of the methodology, we propose models to allow multiple unknown change-points and use information based criteria (Akaike, 1974) to determine the number of change-points.

Besides, the models in Frobish and Ebrahimi (2009) required that the censoring times to be larger than the second change-point. However, the censoring time varies substantially from 57.4 h to 562.5 h for the NTDS. The constraint that the change-point should be smaller than the minimum censoring time would result in an unacceptable upper bound for the change-points. We relax this assumption and the censoring times can be smaller than the change-points.

Furthermore, Dingus et al. (2006), Guo et al. (2013), Klauer et al. (2009), Musselwhite (2006), Rolls et al. (1991), and Ulleberg (2001) have shown that clusters existed among the teenagers with different patterns of risk change. The identical-intensity model assumes that all subjects share the same intensity function and it cannot accommodate the heterogeneity among drivers. To address this issue, we develop a subject-specific intensity model as well.

To develop the notation, denote n_j to be the total number of CNC events and C_j be the cumulative driving time for subject j, $j = 1, \ldots, m$. The events occurred at the ordered times $t_{j1}, \ldots, t_{jn_j}$. The inter-event durations between two consecutive events are $x_{j1}, \ldots, x_{jn_j}$, where $x_{ji} = t_{ji} - t_{j,(i-1)}$, $i = 1, \ldots, n_j$ and $t_{j0} = 0$.

We assume that the CNC event counts follow a NHPP with piecewise constant intensity functions and all subjects share the same d unknown change-points, $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_d)$, $1 \leqslant d \leqslant N$ and d is fixed. N is the total number of events for all the drivers during the study. $\tau_L \leqslant \tau_1 < \tau_2 < \cdots < \tau_d \leqslant \tau_U$, τ_L and τ_U are the known lower and upper bound of change-points respectively. τ_U is required to guarantee the consistency of the proposed estimator. For the ease of notation in the later formulas, define $\tau_0 = \tau_{d+1} = 0$. The d change-points partition the time axis into $d + 1$ intervals: $Y_1, Y_2, \ldots, Y_{d+1}$, $Y_p = (\tau_{p-1}, \tau_p]$ for $p = 1, 2, \ldots, d$, $Y_{d+1} = (\tau_d, +\infty)$. S_p is the index set where $C_j \in Y_p$. The number of elements in S_p is k_p. $N_j^{(p)}$ is the number of events for the jth driver in Y_p, and $\sum_{p=1}^{d+1} N_j^{(p)} = n_j$. I_p is the indicator function on Y_p, $p = 1, 2, \ldots, d + 1$. Each driver joined the study at time zero. Then the

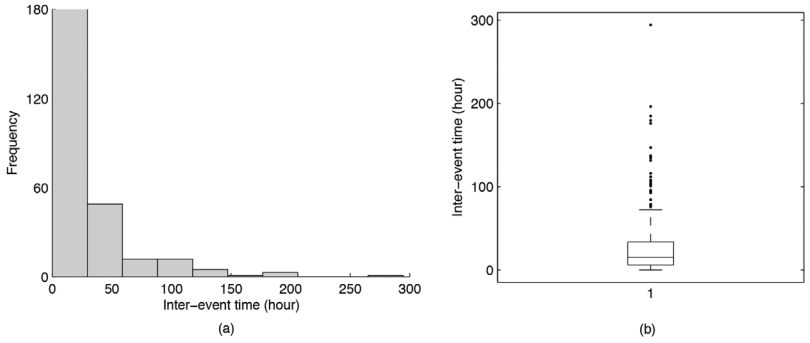

Fig. 2. Distribution of inter-event time from NTDS data: (a) histogram; (b) boxplot.

total number of events for all the drivers is $N = \sum_{j=1}^{m} n_j$.

Depending on whether the intensity functions are treated as equal among subjects, two models are proposed, namely the identical-intensity model and the subject-specific intensity model as discussed below. For a data set, both types of model with different number of change-points can be fit. The best model is then chosen based on the Akaike information criterion (AIC), which combines the goodness of fit for the data and a penalty for model complexity (Akaike, 1974). A smaller AIC indicates a better model and AIC difference greater than four suggests a significant difference between two models (Burnham and Anderson, 2002).

3.1. Identical-intensity model

This model generalizes the likelihood-based model by Frobish and Ebrahimi (2009) to allow for d unknown change-points. We also relax the restriction that the change-points should be smaller than the censoring times, as some participants might drive for a very short period of time.

We assume that m drivers share the identical piecewise constant intensity function $\lambda(t) = \sum_{p=1}^{d+1} \lambda_p I_p(t)$. Integrating $\lambda(t)$ yields

$$\Lambda(t) = \sum_{p=1}^{d+1} \left[\sum_{q=1}^{p-1} \lambda_q(\tau_q - \tau_{q-1}) + \lambda_p(t - \tau_{p-1}) \right] I_p(t).$$

The likelihood for the jth driver (Thompson, 1988) is

$$L_j(\boldsymbol{\lambda}, \boldsymbol{\tau}) = \exp[-\Lambda(C_j)] \prod_{i=1}^{n_j} \lambda(t_{ji}) = \exp[-\Lambda(C_j)] \prod_{p=1}^{d+1} \lambda_p^{N_j^{(p)}}, \quad \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d+1}).$$

The log-likelihood for all the drivers combined is

$$\mathrm{LogL}(\boldsymbol{\tau}, \boldsymbol{\lambda}) = \sum_{p=1}^{d+1} \log \lambda_p \sum_{j=1}^{m} N_j^{(p)} - \sum_{j=1}^{m} \Lambda(C_j), \tag{1}$$

where $\sum_{j=1}^{m} \Lambda(C_j) = \sum_{p=1}^{d+1} \lambda_p \sum_{j \in S_p} C_j + \sum_{p=1}^{d} (m - \sum_{q=1}^{p} k_q)\tau_p(\lambda_p - \lambda_{p+1})$.

Setting the derivatives of $logL(\cdot)$ over $\boldsymbol{\lambda}$ to $\mathbf{0}$ and solving equations yields the MLE:

$$\widehat{\lambda}_p = \frac{\sum_{j=1}^{m} N_j^{(p)}}{\sum_{j \in S_p} C_j - \tau_{p-1}(m - \sum_{q=1}^{p-1} k_q) + \tau_p(m - \sum_{q=1}^{p} k_q)}, \quad p = 1, 2, \ldots, d+1. \tag{2}$$

$\partial^2 \mathrm{logL}(\boldsymbol{\tau}, \boldsymbol{\lambda})/\partial\lambda_p^2 = -\sum_{j=1}^{m} N_j^{(p)}/\lambda_p^2$, which are always non-positive, indicating that $\widehat{\lambda}_p$'s in Eq. (2) are the MLE. $\boldsymbol{\tau}$ has to be estimated first to obtain the numerical value of $\widehat{\boldsymbol{\lambda}}$.

The standard deviation of the estimation can be derived from the information matrix (Lehmann and Casella, 1998),

$$\mathrm{SE}(\widehat{\lambda}_p) = \frac{\widehat{\lambda}_p}{\sqrt{\sum_{j=1}^{m} N_j^{(p)}}}, \quad p = 1, \ldots, d+1. \tag{3}$$

Following Frobish and Ebrahimi (2009), we derive the following theorems.

Theorem 1. *The values of* $\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_d$ *that maximize* $\mathrm{logL}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\tau})$ *in the identical-intensity model can only locate at the observed event times.*

Proof. Plug $\widehat{\boldsymbol{\lambda}}$ into Eq. (1), the profile log-likelihood is:

$$\mathrm{logL}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\tau}) = \sum_{p=1}^{d+1} \log \frac{\sum_{j=1}^{m} N_j^{(p)}}{\sum_{j \in S_p} C_j - \tau_{p-1}(m - \sum_{q=1}^{p-1} k_q) + \tau_p(m - \sum_{q=1}^{p} k_q)} \sum_{j=1}^{m} N_j^{(p)} - N,$$

where N is the total number of events for all the drivers during the study. $\mathrm{logL}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\tau})$ is not a continuous function of $\boldsymbol{\tau}$ but leaps at every event time. Order $\{t_{ji} | 1 \le j \le m, 1 \le i \le n_j\}$ to be $t_{(1)}, t_{(2)}, \ldots, t_{(N)}$. For $p = 1, 2, \ldots, d$, when $t_{(c_p)} \leqslant \tau_p \leqslant t_{(c_p+1)}$ for any fixed index c_p, $N_j^{(p)}$, $\sum_{j \in S_p} C_j$ and $\sum_{q=1}^{p-1} k_q$ are also fixed. Consequently, $\mathrm{logL}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\tau})$ is continuous and differentiable in any fixed set $\{t_{(c_p)} \leqslant \tau_p \leqslant t_{(c_p+1)}, \ p = 1, 2, \ldots, d\}$. It follows that

$$\frac{\partial^2 \mathrm{logL}(\cdot)}{\partial \tau_p^2} = \frac{(m - \sum_{q=1}^{p} k_q)^2 \sum_{j=1}^{m} N_j^{(p)}}{\{\sum_{j \in S_p} C_j - \tau_{p-1}(m - \sum_{q=1}^{p-1} k_q) + \tau_p(m - \sum_{q=1}^{p} k_q)\}^2} + \frac{(m - \sum_{q=1}^{p} k_q)^2 \sum_{j=1}^{m} N_j^{(p+1)}}{\{\sum_{j \in S_{p+1}} C_j - \tau_p(m - \sum_{q=1}^{p} k_q) + \tau_{p+1}(m - \sum_{q=1}^{p+1} k_q)\}^2},$$

which is non-negative. Thus, the values that maximize $logL(\cdot)$ on each interval $[t_{(c_p)}, t_{(c_p+1)}]$ can only be one of the endpoints which are the event times. $\widehat{\boldsymbol{\tau}}$ is the set of d event times between τ_L and τ_U that maximizes $logL(\cdot)$:

$$\widehat{\boldsymbol{\tau}} = \mathrm{argmax}_{\tau_L \leqslant t_p \leqslant \tau_U, t_p = t_{ji} | 1 \le p \le d, 1 \le j \le m, 1 \le i \le n_j} \mathrm{logL}(\boldsymbol{t}), \quad \boldsymbol{t} = (t_1, t_2, \ldots, t_d). \tag{4}$$

□

Note: for all the models in this paper, it can be proved that $\sum_{j=1}^{m} \widehat{\Lambda}(C_j) = N$ by linear algebra. This is consistent with the fact that the other component of the counting process $N(t)$ is a martingale, which has a mean of zero.

Theorem 2. *The MLE of* $\boldsymbol{\tau}$ *on* $[\tau_L, \tau_U]$ *for the identical-intensity model is consistent.*

Proof. Following Frobish and Ebrahimi (2009), $logL(\cdot)$ is finite on $[\tau_L, \tau_U]$, which satisfies the conditions listed in Van der Vaart (1998, p.48, Theorem 5.14). The MLE of $\boldsymbol{\tau}$ on $[\tau_L, \tau_U]$ is consistent as a result. □

3.2. Subject-specific intensity model

Considering a change-point model where intensity function varies by subject, the intensity for subject j is $\lambda_j(t) = \sum_{p=1}^{d+1} \lambda_{jp} I_p(t)$, $j = 1, \ldots, m$. Following the similar steps as in Section 3.1, $\mathrm{logL}(\cdot) = \sum_{p=1}^{d+1} \sum_{j=1}^{m} N_j^{(p)} \log \lambda_{jp} - \sum_{j=1}^{m} \Lambda(C_j)$, where $\sum_{j=1}^{m} \Lambda(C_j) = \sum_{p=1}^{d+1} \sum_{j \in S_p} C_j \lambda_{jp} + \sum_{p=1}^{d} \tau_p \sum_{j \in (\cup_{q=1}^{p} S_q)^c} (\lambda_{jp} - \lambda_{j,p+1})$. The MLE of λ_p's are:

$$\widehat{\lambda}_{jp} = \begin{cases} 0, j \in \cup_{q=1}^{p-1} S_q, \\ \frac{N_j^{(p)}}{C_j - \tau_{p-1}}, j \in S_p, \quad p = 1, 2, \ldots, d+1. \\ \frac{N_j^{(p)}}{\tau_p - \tau_{p-1}}, j \in (\cup_{q=1}^{p} S_q)^c, \end{cases} \tag{5}$$

Following the similar argument as in Section 3.1, we have the following theorems.

Theorem 3. *The values of* $\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_d$ *that maximize* $logL(\cdot)$ *in the subject-specific intensity model locate at the event times.*

Proof. $\mathrm{LogL}(\cdot) = \sum_{p=1}^{d} \sum_{j \in (\cup_{q=1}^{p} S_q)^c} N_j^{(p)} \log \frac{N_j^{(p)}}{\tau_p - \tau_{p-1}} + \sum_{p=1}^{d+1} \sum_{j \in S_p} N_j^{(p)} \log \frac{N_j^{(p)}}{C_j - \tau_{p-1}} - N.$ On each interval $[t_{(c_p)}, t_{(c_p+1)}]$,

$$\frac{\partial^2 \log L(\cdot)}{\partial \tau_p^2} = \frac{\sum_{j\in(\cup_{q=1}^{p} S_q)^c} N_j^{(p)}}{(\tau_p - \tau_{p-1})^2} + \sum_{j\in S_{p+1}} \frac{N_j^{(p+1)}}{(C_j - \tau_p)^2} + \frac{\sum_{j\in(\cup_{q=1}^{p+1} S_q)^c} N_j^{(p+1)}}{(\tau_{p+1} - \tau_p)^2}, \quad p = 1, \ ..., d,$$

which are non-negative for all values of τ_p. □

Theorem 4. *The MLE of* $\boldsymbol{\tau}$ *on* $[\tau_L, \tau_U]$ *for the subject-specific intensity model is consistent.*

Proof. Similarly to Frobish and Ebrahimi (2009), $logL(\cdot)$ is finite on $[\tau_L, \tau_U]$, which guarantees the consistency of the MLE of $\boldsymbol{\tau}$ on $[\tau_L, \tau_U]$. □

4. Simulation studies

We conducted a simulation study to investigate the performance of the proposed models with different sample sizes, intensity rates, and change-point values.

Data from a NHPP with piecewise constant intensity functions is generated based on the distribution of the inter-event times (Klein and Roberts, 1984). For each subject, the cumulative density function (CDF) of the ith inter-event time $X_i = T_i - T_{(i-1)}$ conditional on the previous event times $T_0 = t_0 = 0$, $T_1 = t_1$, …, $T_{i-1} = t_{i-1}$ is

$$F_{t_{i-1}}(x) = \Pr[X_i \le x | T_s = t_s, s = 1, 2, ..., i-1] = 1 - \exp[\Lambda(t_{i-1}) - \Lambda(t_{i-1} + x)],$$

where $\Lambda(\cdot)$ is defined in Section 3. Given the first i events, the next event time t_{i+1} is the ith event time plus the $(i+1)$th inter-event time with CDF F_{t_i}. Accordingly, the generation of the data follows the procedure below:

Step 1: $i = 0$, initialize $t_{(i)} = 0$;
Step 2: Generate $x_{(i+1)}$ with CDF F_{t_i} randomly;
Step 3: $t_{(i+1)} = t_{(i)} + x_{(i+1)}$;
Step 4: $i = i + 1$, go to Step 2.

For the jth individual, the above procedure is repeated until t_{i+1} is larger than the censoring time C_j. $t_1, t_2, \ldots, t_i$, $i = n_j$ are the event times. The process can be run for m times to generate the event times for m drivers. For each combination of parameters, we generate $B = 5000$ data sets with censoring times uniformly from 400 to 500. For the subject-specific intensity model with one change-point, assume $\lambda_{j1} \sim$ *Exponential* (40) and $\lambda_{j2} \sim$ *Exponential* (20), where 40 and 20 are the means of the exponential distributions. For the subject-specific intensity model with two change-points, assume $\lambda_{j1} \sim$ *Exponential* (10), $\lambda_{j2} \sim$ *Exponential* (40), and $\lambda_{j3} \sim$ *Exponential* (10).

We apply models in Section 3 to the simulated data. τ_L is 0 and τ_U is 300. The root mean square error (RMSE) of $\hat{\tau}$ is $\sqrt{(1/B)\sum_{k=1}^{B}(\hat{\tau} - \tau)^2}$ and the absolute percentage bias (%) is $(1/B)|\sum_{k=1}^{B}(\hat{\tau} - \tau)|/\tau \times 100\%$. The coverage probabilities of the 95% confidence interval (CI) are obtained by bootstrapping 1000 times.

Table 1 shows the average, RMSE, absolute percentage bias (%) and coverage probability for the identical-intensity model. The RMSEs are reasonably small and the absolute percentage biases (%) of the parameters are within 3.7% of the true values. The coverage probabilities are above 86%. When the sample size m increases, the RMSE, absolute percentage bias (%) and coverage probability tend to decrease. The estimators behave well overall in the identical-intensity model.

Table 2 shows the results of the subject-specific intensity model with one or two change-points. When the sample size m increases, the RMSE, absolute percentage bias (%) and the coverage probability become smaller. $\hat{\tau}$ behaves fine for the subject-specific intensity model.

In summary, the RMSE, absolute percentage bias (%) and coverage probability decrease when the number of drivers increases. For the subject-specific models, RMSE and |bias (%)| of the parameters are greater than the identical-intensity model, potentially due to increased model complexity and the limited information from the data. Overall, these models give satisfactory results under different scenarios.

Table 1
Simulation results for the identical-intensity model. m is the number of drivers, τ is the change-point, and λ is the intensity rate.

No. of change-points	m	Parameter	True value	Average of estimates	RMSE	|Bias (%)|	Coverage probability (%)
1	40	τ	60	59.1	5.2	1.5	86.8
		λ_1	30	30.8	3.8	2.7	94.1
		λ_2	10	10.0	0.8	0.0	93.8
	80	τ	60	59.6	3.1	0.7	88.2
		λ_1	30	30.4	2.6	1.3	95.3
		λ_2	10	10.0	0.6	0.0	94.2
	40	τ	90	89.0	5.1	1.1	86.6
		λ_1	30	30.5	3.0	1.7	94.5
		λ_2	10	9.9	0.8	1.0	93.5
	80	τ	90	89.5	3.0	0.6	87.7
		λ_1	30	30.3	2.1	1.0	95.0
		λ_2	10	10.0	0.6	0.0	94.2
2	40	τ_1	50	51.0	3.8	2.0	86.6
		τ_2	120	119.3	9.9	0.6	94.0
		λ_1	10	10.2	2.3	2.0	96.0
		λ_2	40	41.5	4.7	3.8	93.2
		λ_3	20	19.9	1.3	0.5	93.3
	80	τ_1	50	50.4	1.6	0.8	87.0
		τ_2	120	119.6	5.1	0.3	93.7
		λ_1	10	10.1	1.6	1.0	94.9
		λ_2	40	40.6	2.9	1.5	95.5
		λ_3	20	20.0	0.9	0.0	94.0
3	40	τ_1	80	79.6	4.8	0.5	94.5
		τ_2	150	150.5	4.9	0.3	97.5
		τ_3	220	219.3	4.8	0.3	91.5
		λ_1	60	60.7	4.4	1.2	98.3
		λ_2	30	29.1	3.5	3.0	93.2
		λ_3	60	61.5	5.2	2.5	96.6
		λ_4	30	29.8	2.1	0.7	94.9

Table 2
Estimation of τ in simulation for the subject-specific intensity model. m is the number of drivers, and τ is the change-point.

No. of change-points	m	τ	Average of estimates	RMSE	|Bias (%)|	Coverage probability (%)
1	40	60	68.0	19.8	13.3	89.7
	80	60	65.2	14.9	8.7	93.1
	120	60	65.1	14.7	8.5	94.6
	40	75	79.1	19.6	5.5	87.0
	80	75	76.4	13.6	1.9	90.4
	120	75	76.4	13.6	1.9	93.0
	40	90	90.8	19.9	0.9	85.6
	80	90	87.6	13.2	2.7	86.2
	120	90	88.8	12.9	1.3	88.5
2	40	$\tau_1 = 50$	53.4	7.2	6.8	83.7
		$\tau_2 = 120$	114.2	10.4	4.8	87.7
	80	$\tau_1 = 50$	50.5	5.6	7.0	89.6
		$\tau_2 = 120$	115.3	8.2	3.9	89.4

5. Application of the intensity models to the NTDS

We applied both types of models with no more than five change-points to the NTDS data. Because of the relatively small sample size, fitting models with more change-points would result in instability of the estimators. Zero and 300 h are chosen as the lower and upper bound of change-points respectively, as the change is likely during the initial period of driving (Simons-Morton et al., 2011). The SEs of rates are

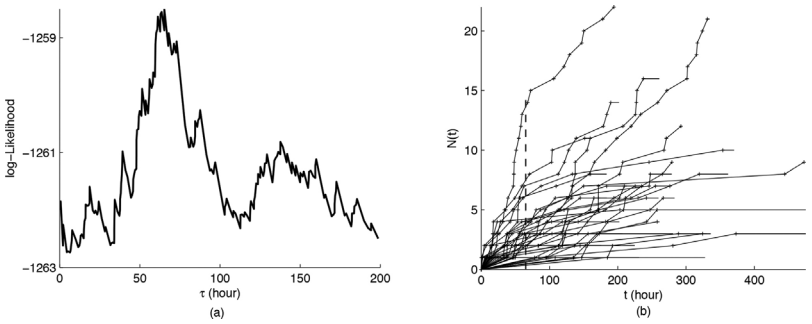

Fig. 3. The identical-intensity model with one change-point: (a) log-likelihood vs. change-point τ, the highest log-likelihood corresponds to the MLE of change-point; (b) the cumulative event counts of each driver over driving time t. The counts are marked by crosses and connected by drivers. The censoring times are connected with the last event times by horizontal lines for each driver. The vertical dashed-line shows the estimated τ.

Table 3
The NTDS application results.

Model	No. of change-points	Parameter	Estimate	95% CI	SE	AIC
Identical-intensity	1	τ	65.3	[42.3, 96.0]	10.9	2523.0
		λ_1	33.7	[21.7, 47.7]	3.5	
		λ_2	23.0	[15.0, 31.1]	1.7	
	2	τ_1	34.0	[24.3, 73.0]	16.4	2524.0
		τ_2	61.2	[58.4, 128.9]	24.3	
		λ_1	28.0	[18.0, 45.9]	4.4	
		λ_2	41.2	[9.2, 60.2]	6.0	
		λ_3	23.1	[14.2, 34.7]	1.7	
Subject-specific intensity	1	τ	73.0	[45.8, 93.8]	11.5	2415.2
	2	τ_1	31.7	[31.7, 72.6]	20.1	2446.6
		τ_2	73.0	[62.7, 108.4]	22.2	

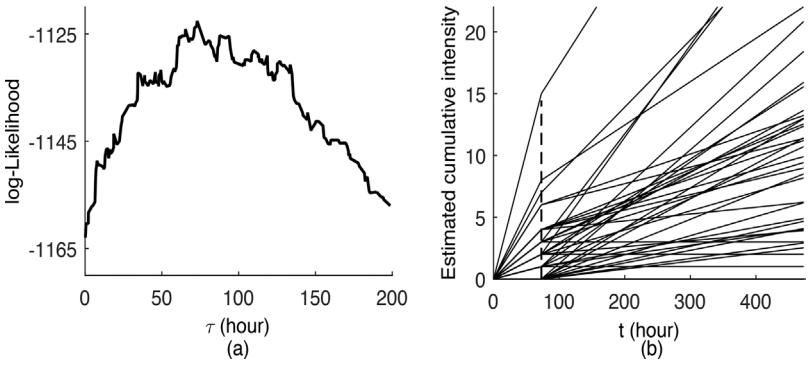

Fig. 4. The subject-specific intensity model with one change-point: (a) log-likelihood vs. change-point τ; (b) the estimated cumulative intensity function $\widehat{\Lambda}_j(t)$ for each driver with the estimated change-point in the vertical line.

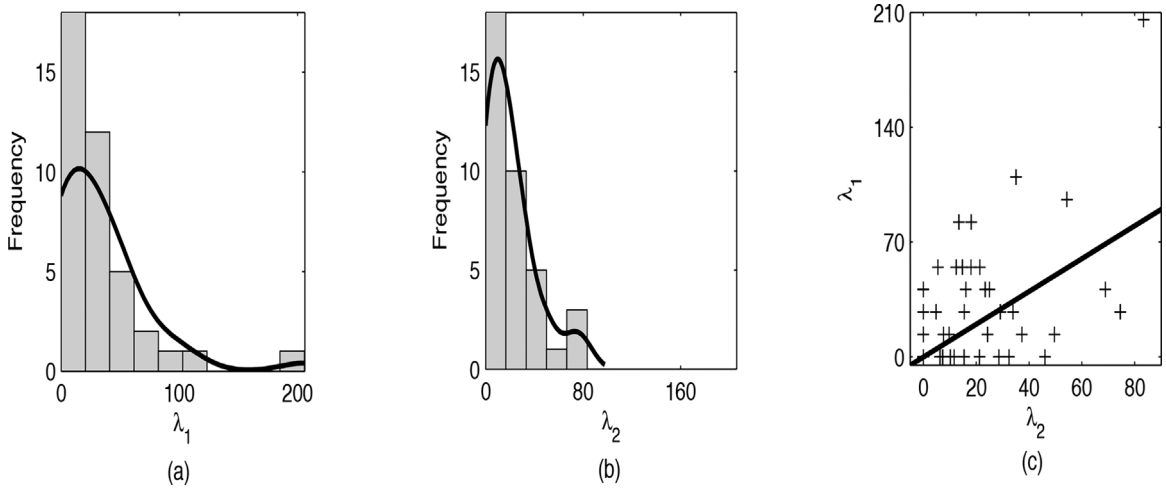

Fig. 5. The intensity rates for all drivers of the subject-specific intensity model with one change-point: (a) histogram of $\widehat{\lambda}_1$ with kernel fitting; (b) histogram of $\widehat{\lambda}_2$ with kernel fitting; (c) scatter plot of $\widehat{\lambda}_1$ vs. $\widehat{\lambda}_2$.

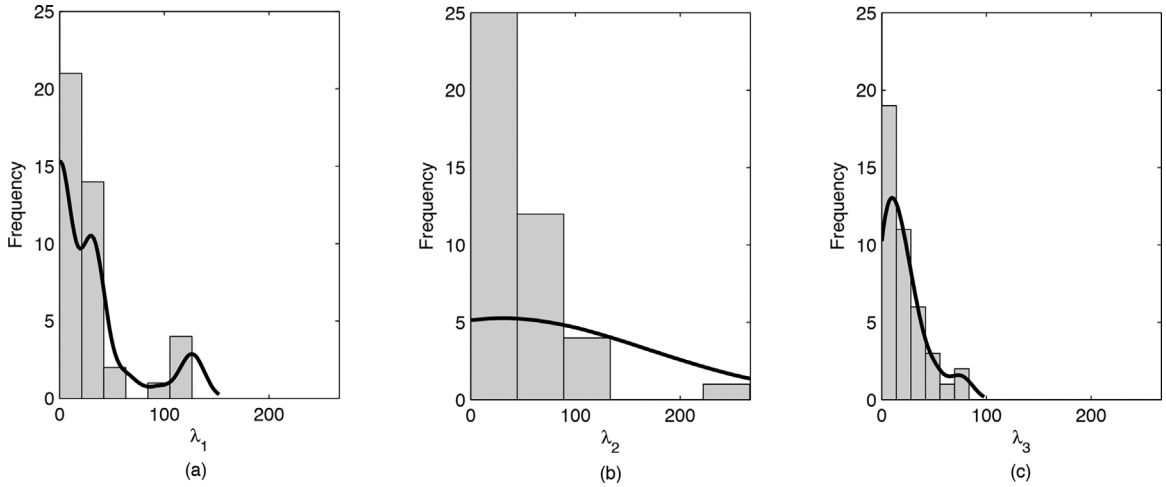

Fig. 6. The histogram and kernel fitting of intensity rates for all drivers in the subject-specific intensity model with two change-points: (a) $\widehat{\lambda}_1$; (b) $\widehat{\lambda}_2$; (c) $\widehat{\lambda}_3$.

Table 4
Intensity rates of the subject-specific intensity model.

No. of change-points	Parameter	Average of λ	95% CI	SD	NTDS average
1	λ_1	32.3	[21.2, 46.4]	39.6	32.6
	λ_2	20.8	[15.2, 27.2]	21.1	23.0
2	λ_1	27.8	[19.7, 46.5]	6.9	27.8
	λ_2	35.7	[14.1, 53.4]	11.1	36.2
	λ_3	20.8	[14.4, 36.0]	5.1	23.0

estimated by Eq. (3) for the identical-intensity model and the SEs of the change-points for both types of models are estimated by block bootstrapping. CIs are estimated by block bootstrapping, where only the drivers are re-sampled 10,000 times and the event-times remain unchanged. The best model is chosen based on the Akaike information criterion (AIC). For both types of models in NTDS application, the AIC is increasing when decreasing the number of change-points. The models with one or two change-points are shown in this section.

For the identical-intensity model with one change-point, Fig. 3 illustrates the change of log-likelihood over change-point and the estimated cumulative function for all drivers. The slopes of the cumulative functions are the corresponding intensity rates, which are the average number of CNC events per teenager per 1000 h. Table 3 shows that the change-point is at 65.3 h of driving after first licensure. The intensity before the change-point is 33.7 CNC events per teenager per 1000 h and becomes 23.0 after the change-point, which is equivalent to the rates computed from the NTDS data directly by the following equation:

$$\text{Average incident rate} = \frac{\text{Total number of events}}{\text{Total driving time} \times \text{Number of drivers}}.$$

For the identical-intensity model with two change-points, τ_2 is close to the change-point estimated by the identical-intensity model with one change-point. Table 3 shows that the intensity increases from 28.0 CNC events per teenager per 1000 h to 41.2 and later drops to 23.1.

For the subject-specific intensity model with one change-point, Fig. 4 shows the log-likelihood over change-point and the estimated cumulative intensity function for each driver. Table 3 shows that the change-point is at 73.0 h of driving after first licensure. Fig. 5 shows two different patterns: approximately half of the drivers have lower risk after the change-point while the risk for others increases. $\widehat{\lambda}_1$ is larger than $\widehat{\lambda}_2$ on the average, which indicate a risk decrease after the change-point on the average.

For the subject-specific model with two change-points, Table 3 shows that the second change-point is equal to the change-point estimated by the subject-specific model with one change-point. The first change-point is close to the first change-point for the identical-intensity model with two change-points. Fig. 6 displays the histogram of the rates for all drivers with the kernel fitting.

For the subject-specific intensity model, the sample averages of the rates are close to the NTDS averages as shown in Table 4, which indicates a good model fitting.

In brief, for the two types of models with no more than five change-points, the subject-specific intensity model with one change-point has the smallest AIC, hence the best model. According to this model, approximately half of the drivers have lower risk after 73.0 h of driving while the risk for others increases. On the average, the intensity rate decreases from 32.3 CNC events per teenager per 10,000 h to 20.8. The AIC is increasing when the number of change-points is increasing for both types of models.

6. Discussion and conclusions

Crash rates among novice teenage drivers are high initially, yet decline rapidly for a period of time, and then decline relatively slower thereafter. The observation is consistent with the development of expertise in complex psycho-motor tasks of all sorts (Keating, 1978). It is useful to identify the change-points in crash rates which may quantify the amount of experience required for relatively safe driving.

It is critical to identify the period during which novices are at greater risk so that parents and policy makers can provide more guidance to protect them and other drivers from their mistakes. Previous research suggested that the change-point occurred at about six months of driving. However, the actual amount of driving time varies considerably during the first six months after licensure. Therefore, identifying the change-point of risk in driving hours provides critical information to improve young drivers' education, safety counter measures, and Graduated Driver Licensing regulations.

This paper develops two alternative recurrent-event models for detecting the change-point of driving risk for novice teenage drivers. Our analysis shows that the subject-specific intensity model with one change-point gives the best fit among the models for NTDS: the change-point is at 73.0 h of driving after licensure with a 95% CI [45.8, 93.8]; approximately half of the drivers have lower risk after the change-point while the risk for others increases. The average driving time per teenager in the NTDS was 68.1 h for five months and 82.3 h for six months. Therefore, the change-point obtained by the subject-specific intensity

model with one change-point is between five and six months after licensure for most novice teens, which is consistent with the findings of Mayhew et al. (2003). In practice, the subject-specific intensity model is more flexible in accommodating different risk patterns among subjects because of variability in CNC rates and change-points in the sample data. More attention should be paid to those teenagers whose driving risk might increase after the change-point. The vehicle technology and driving culture change rapidly. It is important to understand that the estimation from this study could be affected by these changes. Therefore, the results are more likely to change dynamically instead of static.

There are several future extensions for the current work. The subject-specific intensity model can better incorporate individual driver effect, although this type of model is likely leading to over-parameterization. An alternative is using random effects in intensity rates. Another possible extension would be relaxing the condition on identical change-points for all drivers, because clusters may exist among the teenagers considering the different patterns of change (Rolls et al., 1991). Specifically, we are interested in determining if there are particular subgroups with different underlying change patterns, which is partially addressed in Li et al. (2017). Other possible extensions would be considering covariates like gender to predict the change-points and changing the form of the piecewise constant intensity function. The resulting findings would help in developing provisions on graduated licensing policies for teenage drivers to encourage low-risk driving (Williams, 2003).

Acknowledgements

We would like to thank the authors who provided support for our work. This study is partially funded by the National Institute of Child Health and Human Development (funding number: HHSN27520053405C).

References

Achcar, J.A., Loibel, S., Andrade, M.G., 2007. Interfailure data with constant hazard function in the presence of change-points. REVSTAT 5, 209–226.
Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control, AC, vol. 19 716–723 System identification and time-series analysis.
Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer-Verlag, New York.
Burnham, K.P., Anderson, D., 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York.
Cook, R.J., Lawless, J.F., 2007. The Statistical Analysis of Recurrent Events. Statistics for Biology and Health. New York, Springer-Verlag.
Dingus, T., Klauer, S., Neale, V., Petersen, A., Lee, S., Sudweeks, J., et al., 2006. The 100-car naturalistic driving study: phase II – results of the 100-car field experiment. Technical Report DOT-HS-810-593. National Highway Traffic Safety Administration, Washington, DC.
Dupuy, J.-F., 2006. Estimation in a change-point hazard regression model. Stat. Probab. Lett. 76, 182–190.
Field, C.A., Welsh, A.H., 2007. Bootstrapping clustered data? J. R. Stat. Soc. Ser. B. Stat. Methodol. 69 (3), 369–390.
Frobish, D., Ebrahimi, N., 2009. Parametric estimation of change-points for actual event data in recurrent events models? Comput. Stat. Data Anal. 53 (3), 671–682.
Ghosh, J.K., Joshi, S.N., Mukhopadhyay, C., 1993. A Bayesian approach to the estimation of the change-point in a hazard rate. In: Basu, A.P. (Ed.), Advances in Reliability. Elsevier Sciences Publishers, North-Holland, pp. 141–170.
Governor's Highway Safety Association, 2015. State Laws & Funding: Virginia.
Guo, F., Fang, Y., 2013. Individual driver risk assessment using naturalistic driving data. Accid. Anal. Prev. 61, 3–9.
Guo, F., Klauer, S., Hankey, J., Dingus, T., 2010. Using near-crashes as a crash surrogate for naturalistic driving studies. J. Transp. Res. Board 2147, 66–74.
Guo, F., Simons-Morton, B.G., Klauer, S.E., Ouimet, M.C., Dingus, T.A., Lee, S.E., 2013. Variability in crash and near-crash risk among novice teenage drivers: a naturalistic study? J. Pediatrics 163 (6), 1670–1676.
Jackson, J.C., Albert, P.S., Zhang, Z., Simons-Morton, B., 2013. Ordinal latent variable models and their application in the study of newly licensed teenage drivers? J. R. Stat. Soc. Ser. C: Appl. Stat. 62 (3), 435–450.
Karasoy, D., Kadilar, C., 2007. A new Bayes estimate of change-point in the hazard function. Comput. Stat. Data Anal. 51, 2993–3001.
Keating, D., 1978. A search for social intelligence. J. Educ. Psychol. 70, 218–223.
Kim, S., Chen, Z., Zhang, Z., Simons-Morton, B.G., Albert, P.S., 2013. Bayesian hierarchical Poisson regression models: an application to a driving study with kinematic events? J. Am. Stat. Assoc. 108 (502), 494–503.
Klauer, S., Dingus, T.A., Neale, V.L., Sudweeks, J., Ramsey, D.J., 2009. Comparing real-world behaviors of drivers with high vs. low rates of crashes and near-crashes. Technical Report DOT-HS-811-091. National Highway Traffic Safety Administration, Washington, DC.
Klauer, S., Guo, F., Sudweeks, J., Dingus, T., 2010. An analysis of driver inattention using a case-crossover approach on 100-car data. Technical Report DOT-HS-811-334. National Highway Traffic Safety Administration, Washington, DC.
Klauer, S.G., Guo, F., Simons-Morton, B.G., Ouimet, M.C., Lee, S.E., Dingus, T.A., 2014. Distracted driving and risk of road crashes among novice and experienced drivers. N. Engl. J. Med. 370, 54–59.
Klein, R.W., Roberts, S.D., 1984. A time-varying Poisson arrival process generator? Simulation 43 (4), 193–195.
Lawless, J., Zhan, M., 1998. Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. Can. J. Stat. 26, 549–565.
Lee, S., Simons-Morton, B., Klauer, S., Ouimet, M., Dingus, T., 2011. Naturalistic assessment of novice teenage crash experience. Accid. Anal. Prev. 43, 1472–1479.
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, 2nd ed. New York, Springer-Verlag.
Li, Q., Guo, F., Kim, I., Klauer, S., Simons-Monton, B., 2017. A Bayesian finite mixture change-point model for assessing the risk of novice teenage drivers. J. Appl. Stat (in press).
Loader, C.R., 1991. Inference for a hazard rate change point? Biometrika 78 (4), 749–757.
Matthews, D., Farewell, V., 1982. On testing for a constant hazard against a change-point alternative. Biometrics 38, 463–468.
Mayhew, D., Simpson, H., Pak, A., 2003. Changes in collision rates among novice drivers during the first months of driving. Accid. Anal. Prev 35, 683–691.
Müller, H.G., Wang, J.-L., 1994. Change-point models for hazard functions. Change-Point Problems, Volume 23 of IMS Lecture Notes Monogr. Ser. Institute of Mathematical Statistics, Hayward, CA, pp. 224–241.
Musselwhite, C., 2006. Attitudes towards vehicle driving behavior: categorizing and contextualizing risk. Accid. Anal. Prev. 38, 324–334.
National Highway Traffic Safety Administration, 2000. Traffic safety facts 2000: young drivers. Technical Report DOT-HS-809-336. National Center for Statistics & Analysis, Washington, DC.
Nguyen, H.T., Rogers, G.S., Walker, E.A., 1984. Estimation in change-point hazard rate models? Biometrika 71 (2), 299–304.
Ouimet, M.C., Brown, T.G., Guo, F., Klauer, S.G., Simons-Morton, B.G., Fang, Y., Lee, S.E., Gianoulakis, C., Dingus, T.A., 2014. Higher crash and near-crash rates in teenaged drivers with lower cortisol response: an 18-month longitudinal, naturalistic study. JAMA Pediatrics 168 (6), 517–522.
Pons, O., 2002. Estimation in a Cox regression model with a change-point at an unknown time? Statistics 36 (2), 101–124.
Rolls, G., Hall, R.D., Ingham, R., McDonald, M., 1991. Accident Risk and Behavioral Patterns of Younger Drivers. AA Foundation for Road Safety Research, Hampshire, UK.
Ross, S.M., 2006. Simulation, 4th ed. Academic Press.
Simons-Morton, B.G., Guo, F., Klauer, S.G., Ehsani, J.P., Pradhan, A.K., 2014. Keep your eyes on the road: young driver crash risk increases according to duration of distraction? J. Adolesc. Heal. 54 (5), S61–S67.
Simons-Morton, B.G., Ouimet, M.C., Zhang, Z., Lee, S.E., Klauer, S.E., Wang, J., Albert, P.S., Dingus, T.A., 2011a. Crash and risky driving involvement among novice adolescent drivers and their parents. Am. J. Public Health 101, 2362–2367.
Simons-Morton, B.G., Ouimet, M.C., Zhang, Z., Lee, S.E., Klauer, S.G., Wang, J., Chen, R., Albert, P.S., Dingus, T.A., 2011b. The effect of passengers and risk-taking friends on risky driving and crashes/near crashes among novice teenagers. J. Adolesc. Health 49, 587–593.
Thompson, W., 1988. Point Process Models with Applications to Safety and Reliability. UK, Chapman and Hall.
Ulleberg, P., 2001. Personality subtypes of young drivers. Relationship to risk-taking preferences, accident involvement, and response to a traffic safety campaign. Transp. Res. F: Traffic Psychol. Behav. 4, 279–297.
Van der Vaart, A.W., 1998. Asymptotic Statistics, Volume 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, Cambridge University Press.
West, R., Odgen, R., 1997. Continuous-time estimation of a change-point in a Poisson process. J. Stat. Comput. Simul. 56, 293–302.
Williams, A., 2003. Teenage drivers: patterns of risk. J. Saf. Res. 34, 5–15.
Wu, C.Q., Zhao, L.C., Wu, Y.H., 2003. Estimation in change-point hazard function models? Stat. Probab. Lett. 63 (1), 41–48.
Yao, Y.-C., 1986. Maximum likelihood estimation in hazard rate models with a change-point? Commun. Stat. A: Theory Methods 15 (8), 2455–2466.