Lifetime Data Analysis manuscript No.
(will be inserted by the editor)

Recurrent Event Modeling and Analysis of Occurrence of Mass Shootings in the United States

Edsel A. Peña · Lu Wang

Received: date / Accepted: date

Abstract In this paper we analyze the data set gathered by *Mother Jones* magazine concerning mass shootings in the United States during the period from August 20,1982 to January 31, 2017. We limit to those mass shootings with at least four fatalities, excluding the shooter or shooters. We utilize dynamic recurrent event models to model the occurrences of mass shootings, with the models taking into consideration dynamic or internal covariates, such as the accumulated number of mass shootings up to the time of interest. Of particular interest is the detection of a contagion effect, which is the phenomenon in which the rate of occurrence of a mass shooting increases relative to an ambient rate a certain period after a mass shooting. Goodness-of-fit tests of the fitted dynamic models are performed using Pearson-type statistics and forecasting of mass shootings using the fitted models are also discussed.

Keywords Contagion Effect · Cox Regression Model · Dynamic Event-Time Models · Exponential Regression Model · External Covariates · Internal Covariates · Mass Shootings · Pearson-Type Goodness-of-Fit Tests · Weibull Regression Model

1 Threat and Menace of Mass Shootings

The occurrence of a mass shooting is one of the most unnerving and depressing events that happens in our society. Despite the fact that the proportion of deaths from mass shootings is very minuscule relative to all deaths from gun violence, drug-related crimes, accidents, etc. (see, for instance, [5,6]), deaths from mass shootings send tremors to the very fabric of our society because of its senselessness, its irrationality, its randomness, its unexpectedness, and its being so devoid of

Edsel A. Peña
Department of Statistics, University of South Carolina, Columbia, SC 29208 USA
E-mail: pena@stat.sc.edu

Lu Wang
Department of Statistics, University of South Carolina, Columbia, SC 29208 USA
E-mail: lw3@email.sc.edu

explanations. It leads to introspection and re-examination of many aspects of our society, such as gun control and freedom to possess arms, basic rights of citizens, violence and race relations, gender issues, diversity and immigration, political and socio-economic issues, mental health issues, education, religious and moral values, the press and media, the Internet, etc. It has brought deep sadness to many people including our political leaders such as when President Obama was brought to tears while giving a speech related to the mass shooting at Sandy Hook Elementary School in Newtown, Connecticut in December 2012, as well as to spontaneous healing and forgiveness as when this same President started singing *Amazing Grace* during his eulogy in June 26, 2015 for the Honorable Reverend Clementa Pinckney and the eight other victims of the Charleston, South Carolina AME Church massacre. See, for instance, the Washington Post article [3] about aspects of mass shootings that have occurred in the United States over the years.

The probabilistic modeling of mass shooting occurrences is complicated by the possible phenomenon of a 'contagion effect' - the tendency of a higher rate of incidence of mass shootings a short period after an occurrence. There are many potential explanations of such a phenomenon, if indeed it exists. One of them is that with the heightened 24/7 media coverage of such events, potential mass shooters consider the opportunity to commit a mass shooting as a way for recognition because of the intense media coverage. However, this explanation remains a hypothesis since it is difficult to establish this unequivocally with the available observational data. On the other hand, it maybe possible to detect such an increase in incidence of mass shootings a certain period after a mass shooting has occurred, since under ordinary circumstances it is theoretically plausible to assume that mass shootings are occurring on a purely random manner at some ambient rate, for example, according to a non-homogeneous Poisson process.

A major goal of this paper is to demonstrate that general dynamic models for recurrent events could be utilized to model real-world phenomena, in particular the occurrence of mass shootings in our society. It will be demonstrated that dynamic models are better able to model intrinsic features inherent in this mass shooting phenomenon, such as the contagion effect, relative to static-type models.

2 *Mother Jones* Mass Shootings Data Set

The definition of a mass shooting varies in the literature, hence leading to different data sets pertaining to the occurrences of mass shootings. In this paper we follow the definition of a mass shooting in the magazine *Mother Jones*, which defines a mass shooting as having at least four fatalities, excluding the shooter or shooters. *Mother Jones* has kept track of the occurrences of such events in the United States since 1982 [6] and we will utilize their data set. In the later stages of their recording the occurrences of mass shootings, *Mother Jones* started including those events with at least three fatalities. However, since we are interested in the modeling of the successive occurrences of these events and since in the beginning they simply kept track of those with at least four fatalities, we exclude those with only three fatalities in the database. This data set, with some of the variables, is provided in appendix section A. This is for the period from August 20, 1982 to January 31, 2017. The number of days during this time period was $\tau = 12583$ days. This data set includes the following variables:

– `Date`: date of the occurrence of the mass shooting.
– `DayOfWeek`: the day of the week when mass shooting occurred.
– `Location`: this is the place where the mass shooting occurred.
– `Fatalities`: this is the count of the number of deaths, excluding those of the shooter or shooters, in the mass shooting.
– `NumDaysBetw`: the number of days between successive mass shootings.
– `NumDaysFromFirst`: the number of days starting from August 20, 1982, the date of the first recorded mass shooting, which we shall consider as the time origin.

We provide some descriptive summaries of this *Mother Jones* data set. Figure 1 plots the number of days between mass shootings at each of the occurrences of a mass shooting together with a distributional histogram of the inter-event times. One may observe from this plot that the inter-event times are decreasing as time increases. Figure 2 depicts the number of fatalities at each of the mass shooting events together with its distributional histogram. It is not evident that the number of event fatalities increases or decreases as time increases. Another interesting summary is the days of the week in which mass shootings occur. Table 1 provides a frequency/percentage table for the number of mass shootings for each of the seven days of the week. A chi-square goodness-of-fit test of the null hypothesis that the probabilities of mass shootings for each of the days are equal leads to a p-value of 0.1154, hence based on the observed *Mother Jones* data set, it could not be concluded that some days are more prone to mass shootings at the 5% level of significance.

Fig. 1 Plot of the number of days from the time origin of mass shooting occurrences and the inter-event times and a histogram of the inter-event times.

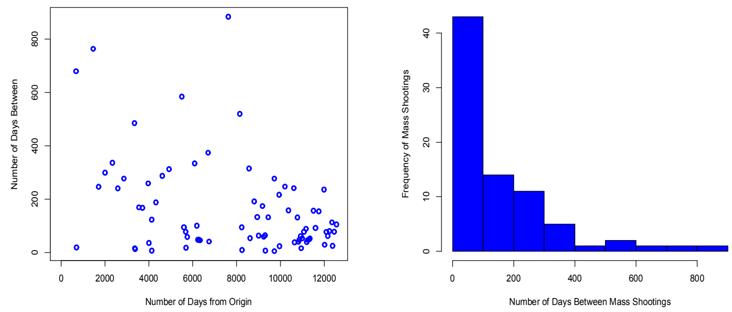
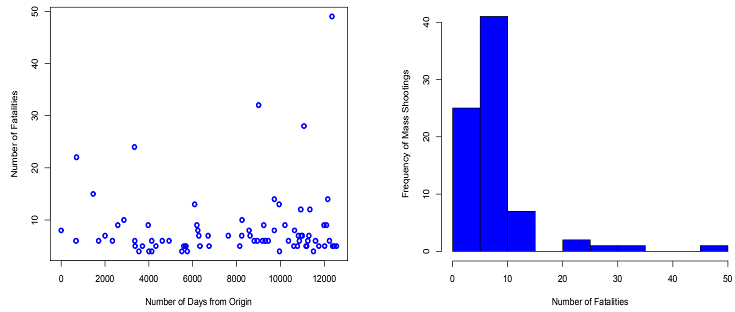


Table 1 Frequency and percentages of occurrences of mass shootings for each day of the week.

Day of Week	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
Frequency	11	15	8	16	15	10	4	79
Percentage	13.9	19.0	10.1	20.3	19.0	12.7	5.1	100

Fig. 2 Plot of the number of days from origin of mass shooting occurrences and the number of fatalities.



3 Stochastic Models

Before proceeding, we first introduce our notation to facilitate describing the models to be considered. We let $\{N(s), 0 \le s \le \tau\}$ be the stochastic process in which $N(s)$ is the number of mass shootings that have occurred over $(0, s]$ with $N(0) = 0$. Note that we do not count the mass shooting at the time origin. τ represents the time of the end of the monitoring period. We let $\mathfrak{F}_s$ denote the history or all the information up to time s. Associated with this stochastic process are the sequences of random variables $\{S_0 = 0, S_1, S_2, \ldots, S_K, S_{K+1} = \tau\}$ with $K = N(\tau-)$, which are the successive times in which the mass shootings occurred, and $\{T_1, T_2, \ldots, T_K, \tau - S_K\}$ which denotes the successive inter-event times. In the *Mother Jones* data set, the S_k's are given by the variable `NumDaysFromFirst`, while the $\{T_k\}$'s are given by the variable `NumDaysBetw`. We also introduce the process $\{F(s), 0 \le s \le \tau\}$ with $F(s)$ denoting the total number of fatalities up to time s, including the number of fatalities at the mass shooting that occurred at the time origin. Thus, at time S_k the number of fatalities is $\Delta F(S_k) = F(S_k) - F(S_k-)$, which are the values contained in the variable `Fatalities` in the *Mother Jones* data set.

At this point we describe the general specification of the model for the counting process $\{N(s), 0 \le s \le \tau\}$. We first introduce the backward recurrence time process $\{\mathcal{E}(s), 0 \le s \le \tau\}$, where $\mathcal{E}(s) = s - S_{N(s-)}$, which is the elapsed time up to s since the last mass shooting. The general stochastic model that we consider for the process $\{N(s)\}$ is of form

$$\begin{aligned} &\Pr\{dN(s) \ge k | \mathcal{F}_{s-}\} = \\ &\lambda_0[\mathcal{E}(s)] \exp\left\{\mathbf{I}(s)\kappa + \mathbf{X}(s)\beta\right\}(ds) I\{k = 1\} + o_p(ds) \end{aligned} \tag{1}$$

where $I\{\cdot\}$ is the indicator function, $\mathbf{I}(s) = (I_1(s), I_2(s), \ldots, I_p(s))$ is a vector of internal covariates, and $\mathbf{X}(s) = (X_1(s), X_2(s), \ldots, X_q(s))$ is a vector of external covariates, both of which are measurable with respect to $\mathcal{F}_{s-}$. We will allow the internal covariate vector to be dependent on a parameter. See [10] for discussions of internal and external covariates. The regression coefficients are $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_p)^{\mathrm{T}}$

and $\beta = (\beta_1, \beta_2, \ldots, \beta_q)^{\intercal}$. The function $\lambda_0(\cdot)$ is a baseline hazard rate function, which could either be parametrically specified or non-parametrically specified. Observe that the effective age used in $\lambda_0(\cdot)$ is the time elapsed since the last mass shooting, the backward recurrence time. Our reason for doing so is our thinking that upon occurrence of a mass shooting, a re-start or a renewal transpires. At the same time, we include in the model the potential impact of internal covariates and external covariates which could increase or decrease the intensity of mass shootings relative to the rate $\lambda_0(\cdot)$. This model belongs to the general class of dynamic recurrent event models in [13].

If we define the process $\{A(s), s \le s \le \tau\}$ via

$$A(s) = \int_0^s \lambda_0[\mathcal{E}(v)] \exp\{\mathbf{I}(v)\kappa + \mathbf{X}(v)\beta\}\, dv, \tag{2}$$

then the process $\{M(s), 0 \le s \le \tau\}$ with $M(s) = N(s) - A(s)$ is a square-integrable zero-mean martingale with predictable quadratic variation process $\{\langle M\rangle(s), 0 \le s \le \tau\}$ given by $\langle M\rangle(s) = A(s)$. For theoretical background, see [2]. The model parameters are $\lambda_0(\cdot)$, κ, β, and any other parameter in the internal covariate vector. Under this model, the likelihood function based on the data $\{N(s), 0 \le s \le \tau\}$ is

$$L_\tau = \left[\prod_{k=1}^{K} \lambda_0(T_k) \exp\{\mathbf{I}(S_k)\kappa + \mathbf{X}(S_k)\beta\}\right] \times \\ \exp\left\{-\int_0^\tau \lambda_0[\mathcal{E}(v)] \exp\{\mathbf{I}(v)\kappa + \mathbf{X}(v)\beta\}\, dv\right\}.$$

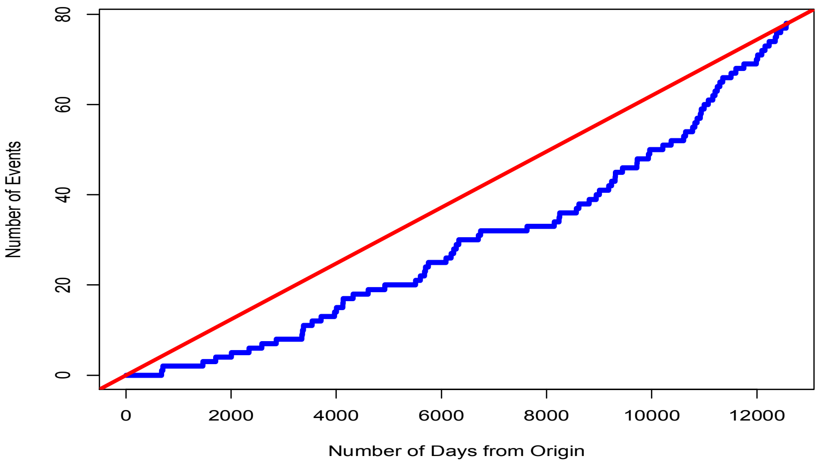
By taking specific forms of $\lambda_0(\cdot)$ and the internal covariates $\mathbf{I}$ and external covariates $\mathbf{X}$, we obtain special models. In the succeeding sections, we consider fitting simple models belonging to this general class of models. We defer consideration of models that have external covariates to future papers, but focus instead in this paper on those with an internal covariate representing the number of previous mass shootings. Likelihood-based inference for these models have been discussed in several papers. When the model is parametric in the sense that the baseline hazard $\lambda_0(\cdot)$ belongs to a parametric family, then the vector of ML estimators is the maximizer of L_τ. If the baseline hazard is non-parametrically specified, the approach using profile and/or partial likelihoods are as discussed in [13].

4 HPP Model

The homogeneous Poisson process (HPP) is typically the first model to consider when fitting recurrent event data. As such, we first consider an HPP as a possible model for the occurrences of mass shootings. The HPP model arises from the general model by taking $\lambda_0(t) = \lambda$ and excluding the internal and external covariates from the model. Thus, there is only one model parameter, λ, which is the rate of mass shooting occurrences. See the recent pedagogical paper [12] concerning the HPP model. We fitted this HPP model using the *Mother Jones* data set. The maximum likelihood estimate (MLE) of λ is

$$\hat{\lambda} = \frac{K}{\tau} = \frac{78}{12583} = 0.0062,$$

Fig. 3 Plot of the number of days from origin of mass shooting occurrence and the cumulative number of mass shootings. Time origin corresponds to August 20, 1982. The red line passing through zero has slope equal to $\hat{\lambda} = 0.0062$, which is the maximum likelihood estimate of the rate of the fitted HPP model.



where $K = 78$ is the total number of observed mass shootings over the monitoring period $(0, \tau]$. In our setting, the monitoring period is from 0 days (time origin) to $\tau = 12583$ days. The times of mass shootings are provided by the variable `NumDaysFromOrigin` in the Mother Jones data set. A plot of this data is provided in Figure 3. The last point in this plot corresponds to the pair of value $(\tau, 78)$, where 78 is the value of K. The straight line passing through zero is the line whose slope is $\hat{\lambda}$.

In [12] a procedure for testing the adequacy of the HPP model, given event occurrence times over a monitoring interval, was presented. This procedure was called the V-test. Applying this V-test, we find the value of the statistic to be $V = 224.26$ with an associated p-value of 0.0003 for testing the null hypothesis that the HPP model holds. Thus, based on the *Mother Jones* data set, the HPP model is an inadequate model for the occurrences of mass shootings in the United States. The inadequate fit could also be noted from the fact that the line $\hat{\lambda}t$ is always above the graph of $(S_k, k), k = 0, 1, 2, \ldots, K$, where S_k is the time of occurrence of the kth mass shooting. If an HPP model is adequate, we would see that the straight line and the graph of $\{(S_k, k)\}$ will be close to each other. In fact, the observed plot appears to indicate that the inter-event times of the mass shootings are ominously getting stochastically shorter as time increases.

5 Dynamic Recurrent Event Models

Noting that the HPP model does not fit well the observed data, we now consider a more general dynamic model for the occurrences of mass shootings. The simplest parametric dynamic model that utilizes $N(s-)$ as the sole dynamic covariate is the Weibull dynamic regression model, which includes as a special case the exponential dynamic regression model (cf., [10]). This specifies that

$$\lambda_0(t; \theta = (\alpha, \eta)) = (\alpha\eta)(\eta t)^{\alpha-1} \quad \text{and} \quad I(s) = N(s-).$$

When $\alpha = 1$, then this is the dynamic exponential regression model. When $\lambda_0(\cdot)$ is simply assumed to be some hazard rate function, then we obtain a dynamic Cox proportional hazards model [4]. These models could be fitted easily using the `survreg` and `coxph` object functions in the `survival` library in the `R` statistical platform [14]. The results of these model fittings are provided below.

For the exponential dynamic regression model the fitted model has

$$\hat{\eta} = 0.003319 \quad \text{and} \quad \hat{\kappa} = 0.019541.$$

It is found that κ is significantly different from zero. In our initial fittings, we also included the number of fatalities of the preceding mass shooting, but this did not turn out to be a significant predictor, hence we did not include this dynamic covariate in the Weibull and Cox PH model fittings. For the Weibull dynamic regression model, the estimates of the parameters are

$$\hat{\alpha} = 1.1285, \hat{\eta} = 0.0031, \hat{\kappa} = 0.0215.$$

It is found that α is significantly different from 1.0 (the exponential baseline hypothesis), and κ is significantly different from 0.

Fitting the semi-parametric Cox proportional hazards model with $I(s) = N(s-)$ as the internal dynamic covariate, we find the estimate of the associated regression coefficient to be

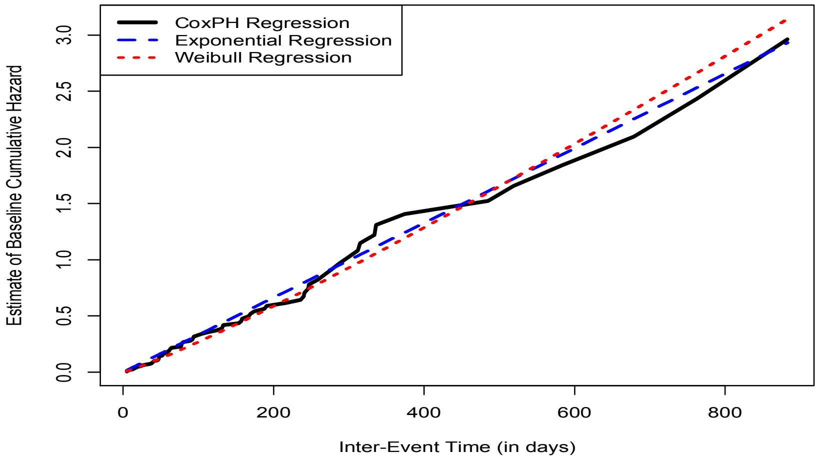
$$\hat{\kappa} = 0.02012.$$

A test of the hypothesis $\kappa = 0$ leads to the conclusion that the dynamic covariate is an important predictor.

In all of these fits, we note that the estimates of the regression coefficient of $N(s-)$ are all positive, indicating that there is an increase in the rate of occurrence of mass shootings with an increase in the number of occurrences of previous mass shootings. Of course, this could simply be that $N(s-)$ is a surrogate of calendar time and as time increases there is an increase in the rate of mass shootings, possibly due to an increase in the population of people or the higher availability of guns. See, for instance, [5,15,6].

For these three models, we also estimated the baseline cumulative hazard function, which is a functional parameter encoding the rate of mass shootings after correcting for the effect of time or prior mass shootings. Figure 4 provides the three estimates together. Observe that all three estimates are all close to being linear. However, note that the nonparametric (or semi-parametric) estimate, the so-called Aalen-Breslow-Nelson (ABN) estimate, obtained via the Cox PH model shows a bump over the linear and Weibull fits in the region from about 300 days to 500 days. This bump *may* be a manifestation of the so-called contagion effect with the fitted Weibull baseline serving as an ambient rate of the occurrence of mass shootings. However, we again point out that this is a difficult hypothesis to prove with the available data. At the same time this demonstrates a potential advantage of the semi-parametric model over the parametric Weibull model since

Fig. 4 Super-imposed plots of the estimated baseline cumulative hazard functions under the dynamic exponential, Weibull, and Cox proportional hazards models, with internal covariate being the number of prior mass shootings.



it may have the ability to potentially tease out such contagion effect. To further explore this issue, we present a plot of the estimates of the baseline hazard probabilities in Figure 5 and obtain associated kernel estimates of the baseline hazard rate function $\lambda_0(\cdot)$ based on this ABN estimate. The kernel estimator used is given by

$$\hat{\lambda}_0(t) = \int_0^\infty \frac{1}{h} K\left(\frac{v-t}{h}\right) d\hat{\Lambda}_0(v) = \sum_{l=1}^{L} \frac{1}{h} K\left(\frac{t_l - t}{h}\right) \hat{\lambda}_{0l}$$

where $t_l, l = 1, 2, \ldots, L$, are the jump times of the ABN estimate $\hat{\Lambda}_0$, which are $\hat{\lambda}_{0l} = \hat{\Lambda}_0(t_l) - \hat{\Lambda}_0(t_l-), l = 1, 2, \ldots, L$. These are the estimated baseline hazard probabilities depicted in Figure 5. $K(\cdot)$ is a kernel function which we set to be the Epanechnikov kernel

$$K(v) = (1 - v^2)^3 I\{|v| \le 1\}.$$

In our implementation, we specified four values for the bandwidth h: 100, 200, 300, and 400. Figure 6 shows these kernel estimates of the baseline hazard rate $\lambda_0(\cdot)$. In all four estimates, we notice the increasing trend and the bump(s) in the hazard rate estimates from time zero until about 300 days, which indicate an increase in the chances of another mass shooting over this period just after a mass shooting. In fact, looking at the first estimate corresponding to the bandwidth $h = 100$, leading to the wiggliest estimate, we notice the first bump in the baseline hazard rate estimate at 71 days (hazard rate estimate of 0.002898) and the second bump at 296 days (hazard rate estimate of 0.004923). These first two bumps in the baseline hazard estimate may be an indication of the presence of a contagion effect in mass shootings.

Fig. 5 Estimates of the discrete hazard probabilities based on the ABN estimate of $\Lambda_0(\cdot)$.

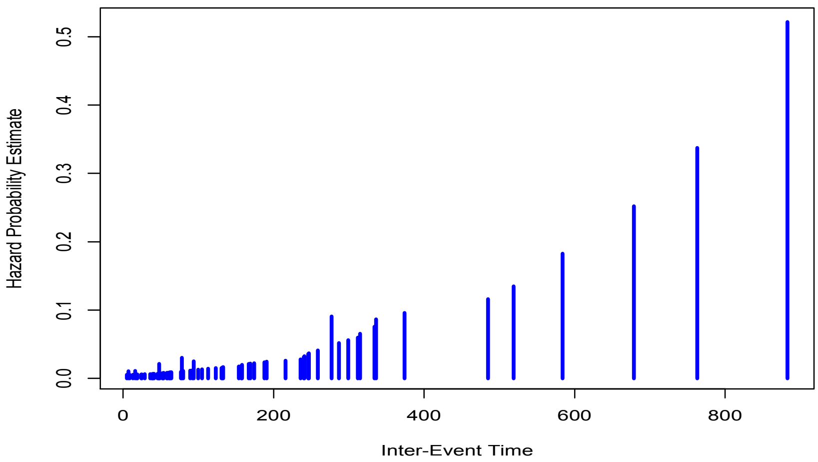
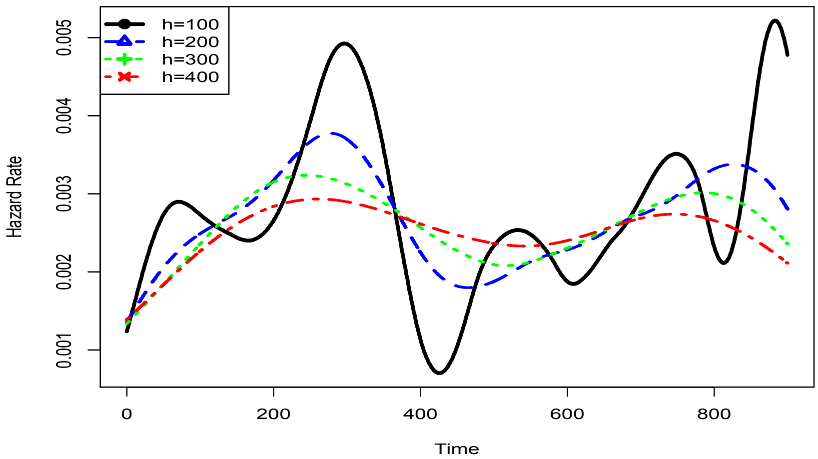


Fig. 6 Super-imposed plots of the kernel estimates of the baseline hazard rate function corresponding to four bandwidths. The bandwidths chosen were $h \in \{100, 200, 300, 400\}$ days.



There has been papers discussing the possibility of such a contagion effect. The paper [15] discussed the potential impact of media coverage after mass shootings and suicides in the context of increasing the incidence of subsequent mass shootings or suicides. The authors presented mathematical contagion models that tried to tease out the contagion effect arising from the enhanced media coverage. They also examined the impact of mental health illness and firearm availability in states. Based on the data sets that they analyzed about mass killings and school shootings they found a significant contagion effect. The *Pacific Standard* article [11] also examined the impact of intense media coverage of mass shootings in the context of an increasing incidence of mass shootings over time.

5.1 Bootstrapping Dynamic Recurrent Event Models

In this subsection we discuss the process of bootstrapping from these dynamic recurrent event models. This will enable us to estimate the standard errors of estimates, and also enable the assessment of significance in goodness-of-fit procedures discussed in the next subsection.

5.1.1 Dynamic Weibull Model

How do we generate the bootstrap samples? Such bootstrap sample generation should take into account the dynamic aspects of the event generation. Consider first the situation where the baseline hazard is Weibull. Based on the observed data, we are able to estimate the Weibull parameters (α, η) and the regression coefficient κ. Let the estimates be denoted by $(\hat{\alpha}, \hat{\eta}, \hat{\kappa})$. To generate one (parametric) bootstrap sample over the monitoring period $[0, \tau]$, we could implement the algorithm presented in appendix section B, which is coded in the `R` syntax. This algorithm incorporates the dynamicity of the event occurrences. Through this bootstrapping procedure we are able to obtain nonparametric estimates of the standard errors of estimates of the model parameters. For the fitted Weibull regression, using Mboot = 10000 bootstrap replications, we found the following bootstrap standard error (bSE's) estimates of the parameter estimates:

$$bSE(\hat{\alpha}) = 0.10888;\ bSE(\hat{\eta}) = 0.00061;\ \text{and}\ bSE(\hat{\kappa}) = 0.00618.$$

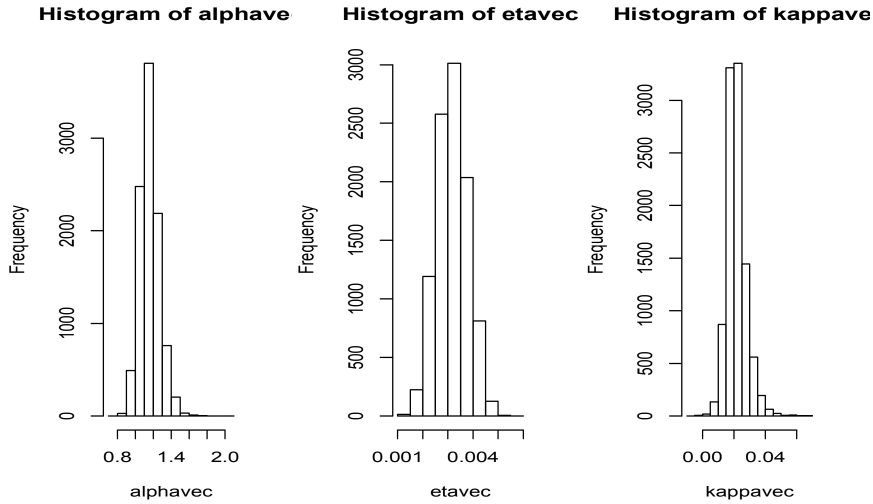
The histograms of the 10000 bootstrap estimates of α, η, and κ are provided in Figure 7. We note that in our bootstrap implementation we put an upper limit to the observed number of events in a bootstrap sample to 1000 (the `maxEvents` in the input to the algorithm). Theoretically, there is the possibility of explosion, that is, the number of events observed in a finite interval increases without bound (see, for instance, [7]) with positive probability. In practical situations, such an eventuality will not be observed, hence putting a large upper limit to the number of observed events is an acceptable solution to this 'blowing up' phenomenon, though this solution may induce some bias. However, we assess that the induced bias is negligible since the proportions of bootstrap samples reaching the upper limit of 1000 was very low out of the 10000 bootstrap samples.

5.1.2 Dynamic Cox Model

When a non-parametric baseline hazard rate function $\lambda_0(\cdot)$ is specified, the portion in the algorithm for the Weibull model containing the two code lines

```
U = runif(1)
T = (-exp(-kappa*Xcur)*log(U))^(1/alpha)/eta
```

Fig. 7 Histograms of the estimates of α, η, and κ based on the 10000 bootstrap samples under the dynamic Weibull regression model.



need to be replaced by generating a value from the observed inter-event times with probabilities induced by the ABN estimate of $\Lambda_0(\cdot)$. The replacement code line is presented after Lemma 1.

To amplify, denote the estimate of $\Lambda_0(\cdot)$ by $\hat{\Lambda}_0(\cdot)$ with jump points $v_1 < v_2 < \ldots < v_L$. For a covariate value of x, then the hazard probabilities under the dynamic model are

$$\hat{\lambda}_j(x) = \hat{\lambda}_{0j} \exp\{\hat{\kappa} x\}, j = 1, 2, \ldots, L, \tag{3}$$

where $\hat{\lambda}_{0j} = \hat{\Lambda}_0(v_j) - \hat{\Lambda}_0(v_{j-1})$ for $j = 1, 2, \ldots, L$ are the estimates of the baseline hazard probabilities at the observed complete inter-event times. It is possible that $\hat{\lambda}_j(x)$ as computed exceeds 1, so if this occurs we set the value to 1. However, these hazard probabilities need not induce a proper set of probabilities on the set $\{v_1, v_2, \ldots, v_L\}$ if $\hat{\lambda}_L(x) < 1$. To induce a proper set of probabilities, we always set $\hat{\lambda}_L(x) = 1$, equivalent to considering the largest observed gap-time as complete. That imposing this condition leads to a proper set of probabilities on $\{v_1, v_2, \ldots, v_L\}$ follows from the following lemma.

Lemma 1 *Let $\lambda_1, \lambda_2, \ldots, \lambda_L$, be such that $\lambda_j \in [0, 1], j = 1, 2, \ldots, L-1$, and $\lambda_L = 1$. Then, letting*

$$p_j = \left[\prod_{i=1}^{j-1}(1 - \lambda_i)\right] \lambda_j, j = 1, 2, \ldots, L,$$

we have $\sum_{j=1}^{L} p_j = 1$, that is, $p_1, p_2, \ldots, p_L$, determines a probability mass function.

Proof A simple inductive proof establishes the result, so we leave the proof as an exercise.

For these $\hat{\lambda}_j(x), j = 1, 2, \ldots, L$, with $\hat{\lambda}_L(x) = 1$, we then define the probabilities

$$\hat{p}_j(x) = \left\{ \prod_{l=1}^{j-1} [1 - \hat{\lambda}_j(x)] \right\} \hat{\lambda}_j(x), j = 1, 2, \ldots, L.$$

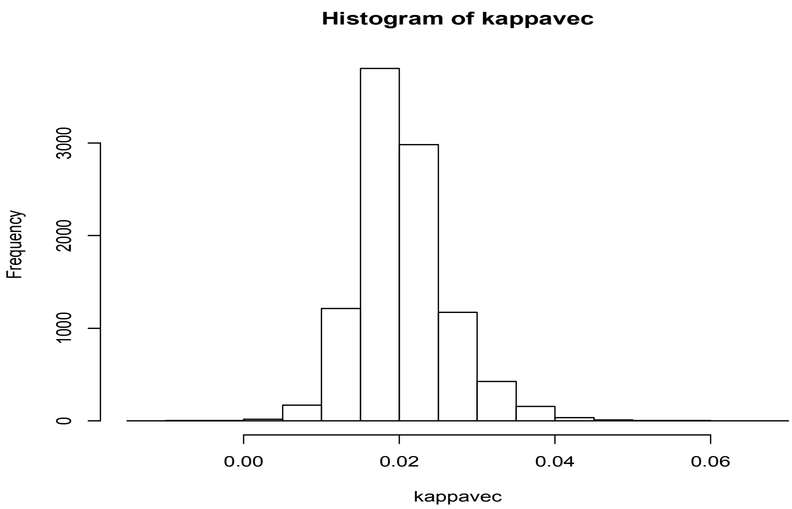
An observation is then generated from the set $\mathbf{v} = \{v_1, v_2, \ldots, v_L\}$ according to the probabilities $\mathbf{p} = (\hat{p}_1(x), \hat{p}_2(x), \ldots, \hat{p}_L(x))$ by using the `R` code

```
T = sample(v,1,prob=p)
```

with `x = Xcur` in the computation in (3). This is the command that replaces the two code lines mentioned above to generate an observation from the semi-parametric dynamic Cox model. The algorithm in `R` syntax is provided in appendix section C. It takes as input the arguments `timein` and `hazin`, which are the distinct complete inter-event times and the baseline hazard probability estimates, respectively, from the ABN estimate.

In this nonparametric bootstrap, the issue of explosion does not arise since there are just a finite number of possible positive values of the generated inter-event time. Also, as in the parametric model, we could use this bootstrap procedure to obtain an estimate of the standard error of the estimator of κ, whose estimate is $\hat{\kappa} = 0.02012$. The bootstrap estimate of its standard error, based on 10000 bootstrap replications, is $bSE(\hat{\kappa}) = 0.00582$. The estimate of the bootstrap distribution of the estimator of κ based on the 10000 bootstrap replications is displayed in Figure 8.

Fig. 8 Histogram of the estimates of κ based on the 10000 bootstrap samples under the dynamic Cox regression model.



5.2 Goodness-of-Fit Testing of Fitted Models

In the simple dynamic models that were fitted, if we partition the monitoring period $(0,\tau]$ into $L+1$ intervals given by $0 \equiv t_0 < t_1 < t_2 < \ldots < t_L < t_{L+1} \equiv \tau$, then the number of events observed in the interval $I_l = (t_{l-1}, t_l]$ is $O_l = N(t_l) - N(t_{l-1})$. If the assumed model holds, then the estimated expected number of events in the interval I_l is given by

$$\begin{aligned}
\hat{E}_l &= \int_{t_{l-1}}^{t_l} \hat{\lambda}_0[\mathcal{E}(v)] e^{\hat{\kappa}N(v-)} dv \\
&= \sum_{j=1}^{K+1} \int_{\max(t_{l-1},S_{j-1})}^{\min(t_l,S_j)} I\{\max(t_{l-1}, S_{j-1}) < \min(t_l, S_j)\} \times \\
&\quad \hat{\lambda}_0(v - S_{N(v-)}) e^{\hat{\kappa}N(v-)} dv \\
&= \sum_{j=1}^{K+1} e^{\hat{\kappa}(j-1)} I\{\max(t_{l-1}, S_{j-1}) < \min(t_l, S_j)\} \times \\
&\quad \int_{\max(t_{l-1},S_{j-1})-S_{j-1}}^{\min(t_l,S_j)-S_{j-1}} \hat{\lambda}_0(w) dw \\
&= \sum_{j=1}^{K+1} e^{\hat{\kappa}(j-1)} I\{\max(t_{l-1}, S_{j-1}) < \min(t_l, S_j)\} \times \\
&\quad \left[\hat{\Lambda}_0(\min(t_l, S_j) - S_{j-1}) - \hat{\Lambda}_0(\max(t_{l-1}, S_{j-1}) - S_{j-1})\right].
\end{aligned}$$

For the Weibull baseline, we will have

$$\hat{\Lambda}_0(t) = (\hat{\eta}t)^{\hat{\alpha}} I\{t > 0\};$$

whereas, for the non-parametrically specified baseline, we will use the ABN estimate of $\Lambda_0(\cdot)$ from the `coxph` fit to evaluate $\hat{\Lambda}_0(t)$.

Analogously to the goodness-of-fit test of Akritas [1] (see also the goodness-of-fit procedure in [8, 9]), we may use the Pearson-type test statistic

$$Q^2 = \sum_{l=1}^{L+1} \frac{(O_l - \hat{E}_l)^2}{\hat{E}_l}$$

for assessing the goodness-of-fit of the fitted model. Significance of the observed value of Q^2 could be assessed by comparing to a chi-square distribution with degrees-of-freedom L minus the number of estimated parameters, or by generating bootstrap samples and obtaining an estimate of the null (that is, that the assumed model is adequate) sampling distribution of the Q^2 statistic.

A bootstrap assessment of the significance of the Q^2 appears to be more reliable in this case since if we utilize the ML estimates of the parameters based on the ungrouped data, then the chi-square distribution may not be the appropriate approximate distribution to use. There is also the un-examined impact of the randomness of K, the number of events observed over the monitoring period. We believe that the bootstrap approximation of the P-value automatically incorporates these issues. There is also the subjective decision of how many intervals

Table 2 The interval upper boundaries together with the observed frequencies and estimated expected frequencies under the dynamic Weibull model and the dynamic Cox model in the implementation of the Pearson-type goodness-of-fit for the fitted models. The lower boundary of the first interval is zero.

Interval Upper Boundary	Observed Freq Frequency	Expected Freq (Weibull)	Expected Freq (Cox)
1398.11	2	4.899	4.328
2796.22	5	9.647	9.418
4194.33	10	8.387	8.240
5592.44	4	9.270	9.028
6990.55	11	10.191	10.735
8388.67	4	13.479	12.491
9786.78	12	12.989	13.537
11184.89	14	15.065	14.625
12583.00	16	19.553	19.169
		$Q^2 = 14.78$ (Bootstrap-$P = .07$)	$Q^2 = 13.00$ (Bootstrap-$P = .16$)

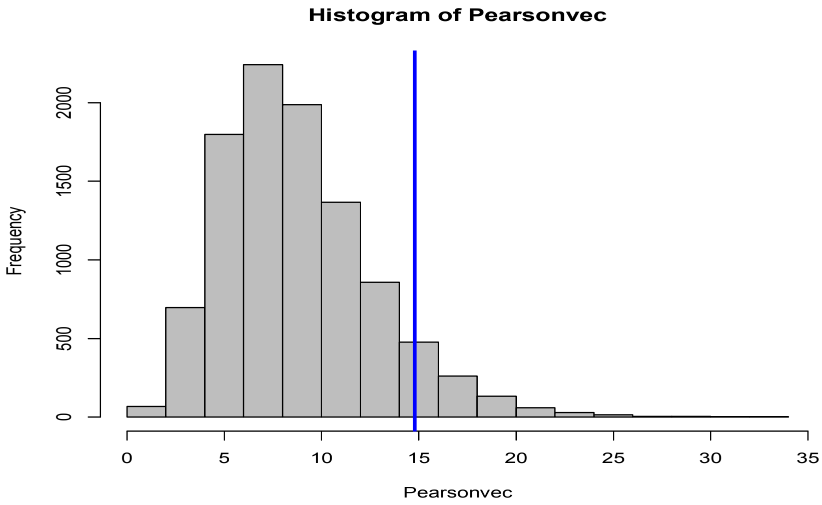
in the partition and what the boundaries should be. The simplest, but which may not be best, is to use equal length intervals, and to have $L \approx \sqrt{K}$ intervals in the partition. Clearly, these issues require more in-depth theoretical studies. Our goal in this paper is to simply utilize a simple *ad hoc* goodness-of-fit procedure to assess the adequacy of the fitted models.

We implemented these ideas by developing appropriate object functions in R [14]. When we applied to the mass shooting data set, we used 10000 bootstrap replications, equal-length partition, and with $L = 8$. Table 2 presents the intervals together with the observed frequency and the estimated expected frequencies under the dynamic Weibull model and dynamic Cox model. Indicated in the last two rows of this table are the values of Q^2 together with their bootstrap P-values. For the dynamic Weibull model, we find $Q^2 = 14.7870$ and the associated bootstrap P-value is 0.0769. This P-value is close to 0.05, possibly indicating that the dynamic Weibull model is not the most appropriate model. Figure 9 presents the histogram of the Q^2 goodness-of-fit statistic for the 10000 bootstrap samples under the Weibull model. For the dynamic Cox model, we find $Q^2 = 13.0051$ with an associated P-value of 0.1659. This appears to indicate that the dynamic Cox model is a better fit to the data than the dynamic Weibull model. We hypothesize that this could be due to the fact that the Weibull model will not be able to model a contagion effect or a bump in the baseline hazard, whereas the dynamic Cox model with a nonparametrically-specified baseline hazard will be able to do so. But this remains a hypothesis and this will be difficult to validate with the existing data. We also mention that among the 10000 bootstrap replications, there were some outlying values of the Q^2-statistic under the dynamic Cox model. This could be seen from the histograms of the Q^2 and the logarithm of this Q^2 statistic which are both depicted in Figure 10.

6 Forecasting Mass Shootings

An oft-quoted saying, attributed to different people (Nostradamus, Niels Bohr, Mark Twain, Yogi Berra, others), is that:

Fig. 9 Histograms of the Q^2 goodness-of-fit statistic based on the 10000 bootstrap samples under the dynamic Weibull regression model. The vertical blue line corresponds to the observed Q^2-statistic.



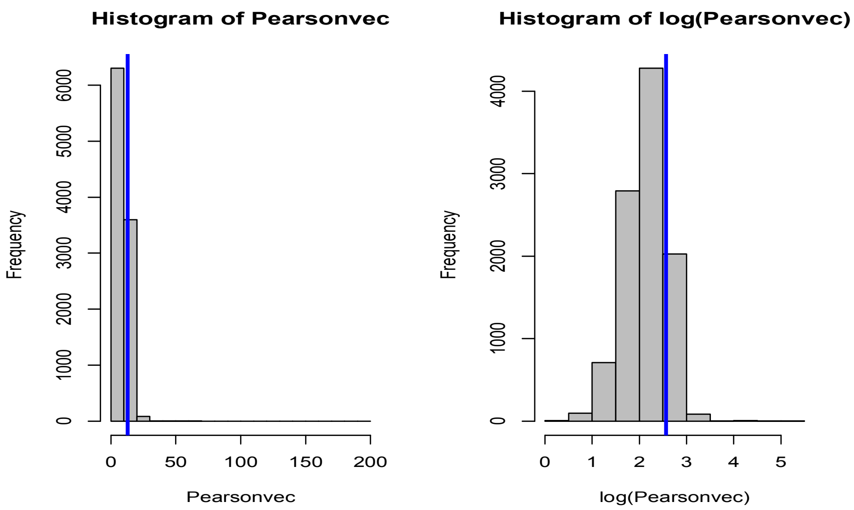
It's Difficult to Make Predictions, Especially About the Future.

And so it is the case with predicting or forecasting when the next mass shooting in the United States is going to occur, even with the aid of the stochastic models fitted to the observed data. Mass shootings are as unpredictable as earthquakes. However, these fitted models provide some guidance on future mass shooting occurrences. For instance, starting from February 1, 2017, one may inquire about the probability that at least one mass shooting will occur in the US during the next four months, that is, until May 31, 2017, which covers 120 days, given the information until January 31, 2017. Using the fitted dynamic Cox model, an estimate of this probability is

$$\begin{aligned}
&\widehat{\Pr}\{25 < T^* \le 146|\text{data until } 1/31/2017\} \\
&= 1 - \widehat{\Pr}\{T^* > 146|\text{data until } 1/31/2017\} \\
&= 1 - \left[\hat{\bar{F}}_0(146)/\hat{\bar{F}}_0(25)\right]^{\exp\{(\hat{\kappa})(78)\}} \\
&= 1 - \left[\prod_{\{l:\ 25<t_l\le 146\}} [1-\hat{\lambda}_{0l}]\right]^{\exp\{(.02012)(78)\}} \\
&= 1 - 0.1780 \\
&= 0.8220,
\end{aligned}$$

where T^* represents the time-to-occurrence of the next mass shooting starting from January 6, 2017, the time of last mass shooting prior to February 1, 2017. The value of 25 is the number of days from 1/6/2017 until 2/1/2017, while the

Fig. 10 Histograms of the Q^2 and $\log(Q^2)$ goodness-of-fit statistics based on the 10000 bootstrap samples under the dynamic Cox regression model. The vertical blue line corresponds to the observed Q^2 and $\log(Q^2)$ statistics.



value of 146 is the number of days from 1/6/2017 until 5/31/2017. Note that the division by $\hat{\bar{F}}_0(25)$ in the second line is because information up to 1/31/2017 indicated that $T^* > 25$, which becomes a conditioning event.

Thus, there is a not-so-insignificant probability of about 82% of at least one mass shooting occurring during the period from February 1, 2017 to May 31, 2017, given the information up to January 31, 2017. Of course, from the perspective of helping to prevent the occurrence of a mass shooting, the probability estimate above will not be of direct help since it does not pinpoint when and where a mass shooting will occur nor could it help in identifying potential mass shooter(s). [**Note:** As of May 1, 2017, the date of initial draft of this manuscript, there has indeed been at least one event since February 1, 2017 that qualifies as a mass shooting. The first occurred last February 6, 2017 in Yazoo City, Mississippi claiming four victims, and another one at Toomsuba, Mississippi last February 21, 2017, also with four victims.]

7 Concluding Remarks

In this paper we analyzed data pertaining to the commission of a mass shooting in the United States spanning the period from August 20, 1982 to January 31, 2017. The data was compiled by the magazine *Mother Jones*, and it includes mass shootings with at least four fatalities, excluding the shooter or shooters. The advent of a mass shooting is one of the most unnerving events in our society. The analysis performed in this paper utilized dynamic event-time models, dynamic in the sense that the occurrence of a mass shooting depends to some extent on what has transpired before. We fitted four models: the homogeneous Poisson process, a dynamic exponential regression model, a dynamic Weibull regression model, and a dynamic Cox model, the latter being a semi-parametric model. We found that the first two models did not fit the data well, while the Weibull-based and the Cox-based models offer better fits to the observed data, with the Cox-based model having the advantage in that it potentially detects a contagion effect through the nonparametric baseline hazard. A contagion effect is one when the rate of occurrence of a mass shooting bumps up a certain period after the last mass shooting, and this has been a topic of interest in several papers dealing with mass shootings. However, we emphasize that it is difficult to validate the presence of this contagion effect on the basis of available data. Both the dynamic Weibull regression and dynamic Cox regression fitted models indicate that the number of prior mass shootings has an effect in terms of the waiting-time for the next occurrence of a mass shooting, with this effect being to stochastically shorten such waiting time. It is certainly conceivable that the number of prior mass shootings serves as a surrogate to calendar time and that as time progresses, the rate of incidence of mass shootings also increases owing to an increase in availability of guns or due to an increasing population size.

We also proposed methods for validating the fitted models through Pearson-type goodness-of-fit tests. However, the determination of the significance of the observed values of these Pearson-type test statistics is done via bootstrapping methods. This led to an examination of the proper way in which to generate bootstrap samples that incorporates the dynamic mechanisms in which events occur. Using the proposed goodness-of-fit methods, we found that the Cox-based model best fits the observed data, though the procedure did not lead to rejection (at level of significance 5%) of the dynamic Weibull model. We also discussed the issue of forecasting the advent of a mass shooting using the fitted Cox dynamic model, with the caveat that the fitted models will not truly be of practical value in terms of preventing when and where a mass shooting will occur nor in pinpointing potential mass shooter(s).

Further studies are warranted regarding the modeling and analysis of mass shootings. More elaborate dynamic models incorporating information from external covariates and possibly with additional information other than those provided by the *Mother Jones* data set will clearly be of interest. It would also be interesting to change the internal covariate from the number of mass shootings that have occurred *since* the time origin to the number of mass shootings that have occurred *within* a certain period, say two years, prior to the time under consideration.

Acknowledgements E. Peña acknowledges grant NIH P30GM103336-01A1 awarded to the University of South Carolina. The authors thank Taeho Kim, Beidi Qiang, Shiwen Shen, and Jeff Thompson for their ideas during research meetings.

References

1. Akritas, M.: Pearson-type goodness-of-fit tests: the univariate case. Journal of the American Statistical Association **83**, 222–230 (1988)
2. Andersen, P., Borgan, O., Gill, R., Keiding, N.: Statistical Models Based on Counting Processes. Springer-Verlag, New York (1993)
3. Berkowitz, B., Gamio, L., Lu, D., Uhrmacher, K., LIndeman, T.: The math of mass shootings. The Washington Post (January 8, 2017)

4. Cox, D.: Regression models and life tables (with discussion). Journal of the Royal Statistical Society **34**, 187–220 (1972)
5. Duwe, G.: Mass Murder in the United States: A History. McFarland & Company, Inc., North Carolina (2007)
6. Follman, M., Aronsen, G., Pan, D.: A guide to mass shootings in America. Mother Jones (January 8, 2017)
7. Gjessing, H.K., Røysland, K., Peña, E.A., Aalen, O.O.: Recurrent events and the exploding Cox model. Lifetime Data Anal. **16**(4), 525–546 (2010)
8. Hjort, N.: Goodness-of-fit tests in models for life history data based on cumulative hazard rates. Annals of Statistics **18**, 1221–1258 (1990)
9. Hollander, M., Peña, E.A.: A chi-squared goodness-of-fit test for randomly censored data. J. Amer. Statist. Assoc. **87**(418), 458–463 (1992)
10. Kalbfleisch, J.D., Prentice, R.L.: The statistical analysis of failure time data, second edn. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ (2002)
11. Keller, J.: Does the media cause mass shootings? Pacific Standard (October 3, 2016)
12. Liu, P., Peña, E.A.: Sojourning With the Homogeneous Poisson Process. Amer. Statist. **70**(4), 413–423 (2016)
13. Peña, E.A., Slate, E.H., González, J.R.: Semiparametric inference for a general class of models for recurrent events. J. Statist. Plann. Inference **137**(6), 1727–1747 (2007)
14. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013). ISBN 3-900051-07-0
15. Towers, S., Gomez-Lievano, A., Khan, M., Mubayi, A., Castillo-Chavez, C.: Contagion in mass killings and school shootings. PLOS One **10**(7), 1–12 (2015)

A *Mother Jones* Mass Shootings Data Set

In the data set below, the variable name abbreviations are: F = Fatalities; T = NumDaysBetw; and S = NumDaysFromFirst. The time origin coincides with 8/20/82, so the value of S corresponds to the number of days elapsed since this date.

	Date	DayOfWeek	Location	F	T	S
1	6/29/84	Thursday	Dallas, Texas	6	679	679
2	7/18/84	Tuesday	San Ysidro, California	22	19	698
3	8/20/86	Tuesday	Edmond, Oklahoma	15	763	1461
4	4/23/87	Wednesday	Palm Bay, Florida	6	246	1707
5	2/16/88	Monday	Sunnyvale, California	7	299	2006
6	1/17/89	Monday	Stockton, California	6	336	2342
7	9/14/89	Wednesday	Louisville, Kentucky	9	240	2582
8	6/18/90	Sunday	Jacksonville, Florida	10	277	2859
9	10/16/91	Tuesday	Killeen, Texas	24	485	3344
10	11/1/91	Thursday	Iowa City, Iowa	6	16	3360
11	11/14/91	Wednesday	Royal Oak, Michigan	5	13	3373
12	5/1/92	Thursday	Olivehurst, California	4	169	3542
13	10/15/92	Wednesday	Watkins Glen, New York	5	167	3709
14	7/1/93	Wednesday	San Francisco, California	9	259	3968
15	8/6/93	Thursday	Fayetteville, North Carolina	4	36	4004
16	12/7/93	Monday	Garden City, New York	6	123	4127
17	12/14/93	Monday	Aurora, Colorado	4	7	4134
18	6/20/94	Sunday	Fairchild Air Force Base, Washington	5	188	4322
19	4/3/95	Sunday	Corpus Christi, Texas	6	287	4609
20	2/9/96	Thursday	Fort Lauderdale, Florida	6	312	4921
21	9/15/97	Sunday	Aiken, South Carolina	4	584	5505
22	12/18/97	Wednesday	Orange, California	5	94	5599
23	3/6/98	Thursday	Newington, Connecticut	5	78	5677
24	3/24/98	Monday	Jonesboro, Arkansas	5	18	5695
25	5/21/98	Wednesday	Springfield, Oregon	4	58	5753
26	4/20/99	Monday	Littleton, Colorado	13	334	6087

```
27  7/29/99 Wednesday                     Atlanta, Georgia  9 100  6187
28  9/15/99   Tuesday                   Fort Worth, Texas  8  48  6235
29  11/2/99    Monday                    Honolulu, Hawaii  7  48  6283
30 12/30/99 Wednesday                      Tampa, Florida  5  46  6329
31 12/26/00   Tuesday            Wakefield, Massachusetts  7 374  6703
32   2/5/01    Monday              Melrose Park, Illinois  5  41  6744
33   7/8/03   Tuesday               Meridian, Mississippi  7 883  7627
34  12/8/04 Wednesday                      Columbus, Ohio  5 519  8146
35  3/12/05  Saturday               Brookfield, Wisconsin  7  94  8240
36  3/21/05    Monday                 Red Lake, Minnesota 10   9  8249
37  1/30/06    Monday                  Goleta, California  8 315  8564
38  3/25/06  Saturday                   Seattle, Washington  7  54  8618
39  10/2/06    Monday      Lancaster County, Pennsylvania  6 191  8809
40  2/12/07    Monday                Salt Lake City, Utah  6 133  8942
41  4/16/07    Monday               Blacksburg, Virginia 32  63  9005
42  10/7/07    Sunday                  Crandon, Wisconsin  6 174  9179
43  12/5/07 Wednesday                     Omaha, Nebraska  9  59  9238
44   2/7/08  Thursday                  Kirkwood, Missouri  6  64  9302
45  2/14/08  Thursday                    DeKalb, Illinois  6   7  9309
46  6/25/08 Wednesday                 Henderson, Kentucky  6 132  9441
47  3/29/09    Sunday            Carthage, North Carolina  8 277  9718
48   4/3/09    Friday               Binghamton, New York 14   5  9723
49  11/5/09  Thursday                   Fort Hood, Texas 13 216  9939
50 11/29/09    Sunday                Parkland, Washington  4  24  9963
51   8/3/10   Tuesday             Manchester, Connecticut  9 247 10210
52   1/8/11  Saturday                      Tucson, Arizona  6 158 10368
53   9/6/11   Tuesday                  Carson City, Nevada  5 241 10609
54 10/14/11    Friday               Seal Beach, California  8  38 10647
55  2/22/12 Wednesday                   Norcross, Georgia  5 131 10778
56   4/2/12    Monday                  Oakland, California  7  40 10818
57  5/20/12    Sunday                  Seattle, Washington  6  48 10866
58  7/20/12    Friday                    Aurora, Colorado 12  61 10927
59   8/5/12    Sunday                 Oak Creek, Wisconsin  7  16 10943
60  9/27/12  Thursday              Minneapolis, Minnesota  7  53 10996
61 12/14/12    Friday               Newtown, Connecticut 28  78 11074
62  3/13/13 Wednesday             Herkimer County, New York  5  89 11163
63  4/21/13    Sunday              Federal Way, Washington  5  39 11202
64   6/7/13    Friday             Santa Monica, California  6  47 11249
65  7/26/13    Friday                     Hialeah, Florida  7  49 11298
66  9/16/13    Monday                    Washington, D.C. 12  52 11350
67  2/20/14  Thursday                 Alturas, California  4 157 11507
68  5/23/14    Friday            Santa Barbara, California  6  92 11599
69 10/24/14    Friday               Marysville, Washington  5 154 11753
70  6/17/15 Wednesday           Charleston, South Carolina  9 236 11989
71  7/16/15  Thursday              Chattanooga, Tennessee  5  29 12018
72  10/1/15  Thursday                     Roseburg, Oregon  9  77 12095
73  12/2/15 Wednesday           San Bernardino, California 14  62 12157
74  2/20/16  Saturday           Kalamazoo County, Michigan  6  80 12237
75  6/12/16    Sunday                     Orlando, Florida 49 113 12350
76   7/7/16  Thursday                        Dallas, Texas  5  25 12375
77  9/23/16    Friday                       Burlington, WA  5  78 12453
78   1/6/17    Friday             Fort Lauderdale, Florida  5 105 12558
79     <NA>      <NA>                                 <NA> NA  25 12583
```

B Bootstrap Sampling Algorithm: Dynamic Weibull Model

```
BootSampWeibull <-
 function(alpha=1.1287,eta=0.0031,kappa=0.02146,tau=12583,maxEvents=1000)
{
```

```
S = 0
Svec = 0
Tvec = NULL
Deltavec = NULL
Xvec = 0
Xcur = 0
K = 0

ok = TRUE

while(ok) {

 U = runif(1)
 T = (-exp(-kappa*Xcur)*log(U))^(1/alpha)/eta

 if((S+T) < tau) {
  K = K + 1
  if(K > maxEvents) {ok=FALSE; print("Explosion Occurring!")} #cut-off explosion
  S = S + T
  Xcur = Xcur + 1
  Svec = c(Svec,S)
  Xvec = c(Xvec,Xcur)
  Tvec = c(Tvec,T)
  Deltavec = c(Deltavec,1)
 }
 else {
  ok = FALSE
  Svec = c(Svec,tau)
  Xvec = c(Xvec,Xcur)
  Tvec = c(Tvec,tau-S)
  Deltavec = c(Deltavec,0)
 }
}

return(list(alpha=alpha,eta=eta,kappa=kappa,tau=tau,
 K=K,Svec=Svec,Tvec=Tvec,Deltavec=Deltavec,Xvec=Xvec))
}
```

C Bootstrap Sampling Algorithm: Dynamic Cox Model

```
BootSampCoxPH <-
 function(timein=time,hazin=haz,kappa=0.02012,tau=12583)
{

L = length(timein)
probs = rep(0,L)

S = 0
Svec = 0
Tvec = NULL
Deltavec = NULL
Xvec = NULL
Xcur = 0
K = 0

ok = TRUE

while(ok) {
```

```
curhaz = hazin*exp(kappa*Xcur)
for(l in 1:L) {curhaz[l] = min(c(1,curhaz[l]))}
curhaz[L] = 1

probs[1] = curhaz[1]
for(l in 2:L) {probs[l] = probs[l-1]*(1/curhaz[l-1] - 1)*curhaz[l]}

T = sample(timein,1,prob=probs)

 if((S+T) < tau) {
  K = K + 1
  S = S + T
  Xcur = Xcur + 1
  Svec = c(Svec,S)
  Xvec = c(Xvec,Xcur)
  Tvec = c(Tvec,T)
  Deltavec = c(Deltavec,1)
 }
 else {
  ok = FALSE
  Svec = c(Svec,tau)
  Xvec = c(Xvec,Xcur)
  Tvec = c(Tvec,tau-S)
  Deltavec = c(Deltavec,0)
 }
}
return(list(kappa=kappa,tau=tau,K=K,Svec=Svec,Tvec=Tvec,
 Deltavec=Deltavec,Xvec=Xvec,tau=tau))
}
```

Lifetime Data Analysis manuscript No.
(will be inserted by the editor)

Recurrent Event Modeling and Analysis of Occurrence of Mass Shootings in the United States

BLINDED

Received: date / Accepted: date

Abstract In this paper we analyze the data set gathered by *Mother Jones* magazine concerning mass shootings in the United States during the period from August 20,1982 to January 31, 2017. We limit to those mass shootings with at least four fatalities, excluding the shooter or shooters. We utilize dynamic recurrent event models to model the occurrences of mass shootings, with the models taking into consideration dynamic or internal covariates, such as the accumulated number of mass shootings up to the time of interest. Of particular interest is the detection of a contagion effect, which is the phenomenon in which the rate of occurrence of a mass shooting increases relative to an ambient rate a certain period after a mass shooting. Goodness-of-fit tests of the fitted dynamic models are performed using Pearson-type statistics and forecasting of mass shootings using the fitted models are also discussed.

Keywords Contagion Effect · Cox Regression Model · Dynamic Event-Time Models · Exponential Regression Model · External Covariates · Internal Covariates · Mass Shootings · Pearson-Type Goodness-of-Fit Tests · Weibull Regression Model

1 Threat and Menace of Mass Shootings

The occurrence of a mass shooting is one of the most unnerving and depressing events that happens in our society. Despite the fact that the proportion of deaths from mass shootings is very minuscule relative to all deaths from gun violence, drug-related crimes, accidents, etc. (see, for instance, [5,6]), deaths from mass shootings send tremors to the very fabric of our society because of its senselessness, its irrationality, its randomness, its unexpectedness, and its being so devoid of explanations. It leads to introspection and re-examination of many aspects of our society, such as gun control and freedom to possess arms, basic rights of citizens, violence and race relations, gender issues, diversity and immigration, political and socio-economic issues, mental health issues, education, religious and moral values,

BLINDED

the press and media, the Internet, etc. It has brought deep sadness to many people including our political leaders such as when President Obama was brought to tears while giving a speech related to the mass shooting at Sandy Hook Elementary School in Newtown, Connecticut in December 2012, as well as to spontaneous healing and forgiveness as when this same President started singing *Amazing Grace* during his eulogy in June 26, 2015 for the Honorable Reverend Clementa Pinckney and the eight other victims of the Charleston, South Carolina AME Church massacre. See, for instance, the Washington Post article [3] about aspects of mass shootings that have occurred in the United States over the years.

The probabilistic modeling of mass shooting occurrences is complicated by the possible phenomenon of a 'contagion effect' - the tendency of a higher rate of incidence of mass shootings a short period after an occurrence. There are many potential explanations of such a phenomenon, if indeed it exists. One of them is that with the heightened 24/7 media coverage of such events, potential mass shooters consider the opportunity to commit a mass shooting as a way for recognition because of the intense media coverage. However, this explanation remains a hypothesis since it is difficult to establish this unequivocally with the available observational data. On the other hand, it maybe possible to detect such an increase in incidence of mass shootings a certain period after a mass shooting has occurred, since under ordinary circumstances it is theoretically plausible to assume that mass shootings are occurring on a purely random manner at some ambient rate, for example, according to a non-homogeneous Poisson process.

A major goal of this paper is to demonstrate that general dynamic models for recurrent events could be utilized to model real-world phenomena, in particular the occurrence of mass shootings in our society. It will be demonstrated that dynamic models are better able to model intrinsic features inherent in this mass shooting phenomenon, such as the contagion effect, relative to static-type models.

2 *Mother Jones* Mass Shootings Data Set

The definition of a mass shooting varies in the literature, hence leading to different data sets pertaining to the occurrences of mass shootings. In this paper we follow the definition of a mass shooting in the magazine *Mother Jones*, which defines a mass shooting as having at least four fatalities, excluding the shooter or shooters. *Mother Jones* has kept track of the occurrences of such events in the United States since 1982 [6] and we will utilize their data set. In the later stages of their recording the occurrences of mass shootings, *Mother Jones* started including those events with at least three fatalities. However, since we are interested in the modeling of the successive occurrences of these events and since in the beginning they simply kept track of those with at least four fatalities, we exclude those with only three fatalities in the database. This data set, with some of the variables, is provided in appendix section A. This is for the period from August 20, 1982 to January 31, 2017. The number of days during this time period was $\tau = 12583$ days. This data set includes the following variables:

- `Date`: date of the occurrence of the mass shooting.
- `DayOfWeek`: the day of the week when mass shooting occurred.
- `Location`: this is the place where the mass shooting occurred.

– `Fatalities`: this is the count of the number of deaths, excluding those of the shooter or shooters, in the mass shooting.
– `NumDaysBetw`: the number of days between successive mass shootings.
– `NumDaysFromFirst`: the number of days starting from August 20, 1982, the date of the first recorded mass shooting, which we shall consider as the time origin.

We provide some descriptive summaries of this *Mother Jones* data set. Figure 1 plots the number of days between mass shootings at each of the occurrences of a mass shooting together with a distributional histogram of the inter-event times. One may observe from this plot that the inter-event times are decreasing as time increases. Figure 2 depicts the number of fatalities at each of the mass shooting events together with its distributional histogram. It is not evident that the number of event fatalities increases or decreases as time increases. Another interesting summary is the days of the week in which mass shootings occur. Table 1 provides a frequency/percentage table for the number of mass shootings for each of the seven days of the week. A chi-square goodness-of-fit test of the null hypothesis that the probabilities of mass shootings for each of the days are equal leads to a p-value of 0.1154, hence based on the observed *Mother Jones* data set, it could not be concluded that some days are more prone to mass shootings at the 5% level of significance.

Fig. 1 Plot of the number of days from the time origin of mass shooting occurrences and the inter-event times and a histogram of the inter-event times.

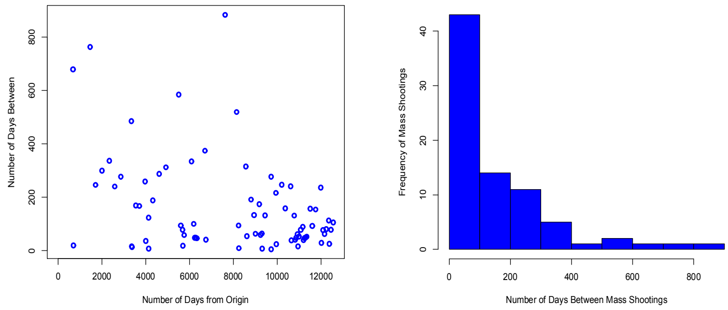
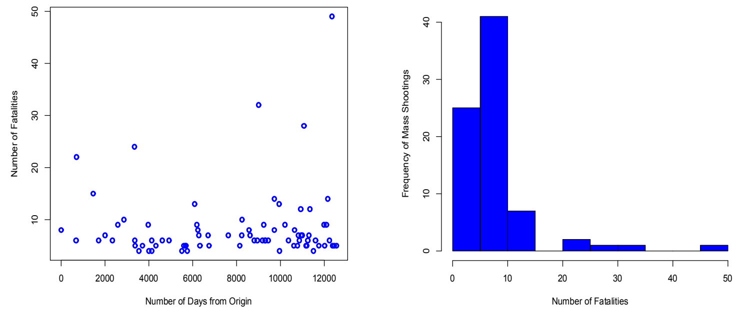


Table 1 Frequency and percentages of occurrences of mass shootings for each day of the week.

Day of Week	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
Frequency	11	15	8	16	15	10	4	79
Percentage	13.9	19.0	10.1	20.3	19.0	12.7	5.1	100

Fig. 2 Plot of the number of days from origin of mass shooting occurrences and the number of fatalities.



3 Stochastic Models

Before proceeding, we first introduce our notation to facilitate describing the models to be considered. We let $\{N(s), 0 \leq s \leq \tau\}$ be the stochastic process in which $N(s)$ is the number of mass shootings that have occurred over $(0, s]$ with $N(0) = 0$. Note that we do not count the mass shooting at the time origin. τ represents the time of the end of the monitoring period. We let $\mathfrak{F}_s$ denote the history or all the information up to time s. Associated with this stochastic process are the sequences of random variables $\{S_0 = 0, S_1, S_2, \ldots, S_K, S_{K+1} = \tau\}$ with $K = N(\tau-)$, which are the successive times in which the mass shootings occurred, and $\{T_1, T_2, \ldots, T_K, \tau - S_K\}$ which denotes the successive inter-event times. In the *Mother Jones* data set, the S_k's are given by the variable `NumDaysFromFirst`, while the $\{T_k\}$'s are given by the variable `NumDaysBetw`. We also introduce the process $\{F(s), 0 \leq s \leq \tau\}$ with $F(s)$ denoting the total number of fatalities up to time s, including the number of fatalities at the mass shooting that occurred at the time origin. Thus, at time S_k the number of fatalities is $\Delta F(S_k) = F(S_k) - F(S_k-)$, which are the values contained in the variable `Fatalities` in the *Mother Jones* data set.

At this point we describe the general specification of the model for the counting process $\{N(s), 0 \leq s \leq \tau\}$. We first introduce the backward recurrence time process $\{\mathcal{E}(s), 0 \leq s \leq \tau\}$, where $\mathcal{E}(s) = s - S_{N(s-)}$, which is the elapsed time up to s since the last mass shooting. The general stochastic model that we consider for the process $\{N(s)\}$ is of form

$$\Pr\{dN(s) \geq k | \mathcal{F}_{s-}\} = \\ \lambda_0[\mathcal{E}(s)] \exp\{\mathbf{I}(s)\kappa + \mathbf{X}(s)\beta\}(ds) I\{k = 1\} + o_p(ds) \quad (1)$$

where $I\{\cdot\}$ is the indicator function, $\mathbf{I}(s) = (I_1(s), I_2(s), \ldots, I_p(s))$ is a vector of internal covariates, and $\mathbf{X}(s) = (X_1(s), X_2(s), \ldots, X_q(s))$ is a vector of external covariates, both of which are measurable with respect to $\mathcal{F}_{s-}$. We will allow the internal covariate vector to be dependent on a parameter. See [10] for discussions of internal and external covariates. The regression coefficients are $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_p)^{\mathrm{T}}$

and $\beta = (\beta_1, \beta_2, \ldots, \beta_q)^{\intercal}$. The function $\lambda_0(\cdot)$ is a baseline hazard rate function, which could either be parametrically specified or non-parametrically specified. Observe that the effective age used in $\lambda_0(\cdot)$ is the time elapsed since the last mass shooting, the backward recurrence time. Our reason for doing so is our thinking that upon occurrence of a mass shooting, a re-start or a renewal transpires. At the same time, we include in the model the potential impact of internal covariates and external covariates which could increase or decrease the intensity of mass shootings relative to the rate $\lambda_0(\cdot)$. This model belongs to the general class of dynamic recurrent event models in [13].

If we define the process $\{A(s), s \leq s \leq \tau\}$ via

$$A(s) = \int_0^s \lambda_0[\mathcal{E}(v)] \exp\{\mathbf{I}(v)\kappa + \mathbf{X}(v)\beta\}\, dv, \tag{2}$$

then the process $\{M(s), 0 \leq s \leq \tau\}$ with $M(s) = N(s) - A(s)$ is a square-integrable zero-mean martingale with predictable quadratic variation process $\{\langle M\rangle(s), 0 \leq s \leq \tau\}$ given by $\langle M\rangle(s) = A(s)$. For theoretical background, see [2]. The model parameters are $\lambda_0(\cdot)$, κ, β, and any other parameter in the internal covariate vector. Under this model, the likelihood function based on the data $\{N(s), 0 \leq s \leq \tau\}$ is

$$L_\tau = \left[\prod_{k=1}^{K} \lambda_0(T_k) \exp\{\mathbf{I}(S_k)\kappa + \mathbf{X}(S_k)\beta\}\right] \times \exp\left\{-\int_0^\tau \lambda_0[\mathcal{E}(v)] \exp\{\mathbf{I}(v)\kappa + \mathbf{X}(v)\beta\}\, dv\right\}.$$

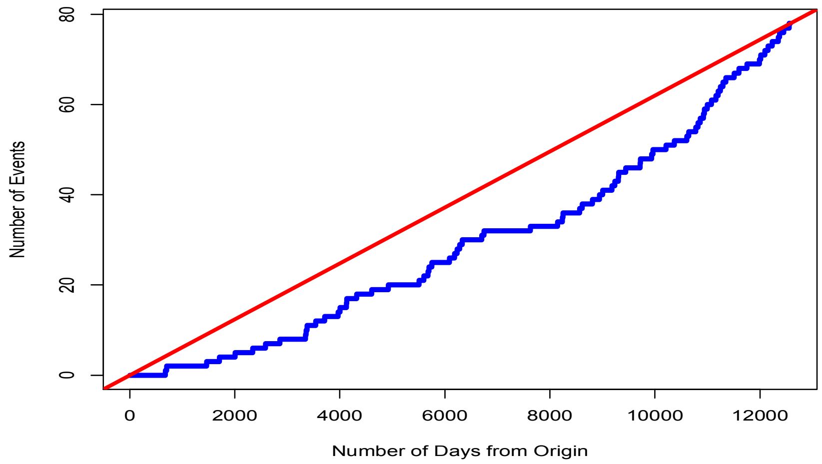
By taking specific forms of $\lambda_0(\cdot)$ and the internal covariates $\mathbf{I}$ and external covariates $\mathbf{X}$, we obtain special models. In the succeeding sections, we consider fitting simple models belonging to this general class of models. We defer consideration of models that have external covariates to future papers, but focus instead in this paper on those with an internal covariate representing the number of previous mass shootings. Likelihood-based inference for these models have been discussed in several papers. When the model is parametric in the sense that the baseline hazard $\lambda_0(\cdot)$ belongs to a parametric family, then the vector of ML estimators is the maximizer of L_τ. If the baseline hazard is non-parametrically specified, the approach using profile and/or partial likelihoods are as discussed in [13].

4 HPP Model

The homogeneous Poisson process (HPP) is typically the first model to consider when fitting recurrent event data. As such, we first consider an HPP as a possible model for the occurrences of mass shootings. The HPP model arises from the general model by taking $\lambda_0(t) = \lambda$ and excluding the internal and external covariates from the model. Thus, there is only one model parameter, λ, which is the rate of mass shooting occurrences. See the recent pedagogical paper [12] concerning the HPP model. We fitted this HPP model using the *Mother Jones* data set. The maximum likelihood estimate (MLE) of λ is

$$\hat{\lambda} = \frac{K}{\tau} = \frac{78}{12583} = 0.0062,$$

Fig. 3 Plot of the number of days from origin of mass shooting occurrence and the cumulative number of mass shootings. Time origin corresponds to August 20, 1982. The red line passing through zero has slope equal to $\hat{\lambda} = 0.0062$, which is the maximum likelihood estimate of the rate of the fitted HPP model.



where $K = 78$ is the total number of observed mass shootings over the monitoring period $(0, \tau]$. In our setting, the monitoring period is from 0 days (time origin) to $\tau = 12583$ days. The times of mass shootings are provided by the variable `NumDaysFromOrigin` in the Mother Jones data set. A plot of this data is provided in Figure 3. The last point in this plot corresponds to the pair of value $(\tau, 78)$, where 78 is the value of K. The straight line passing through zero is the line whose slope is $\hat{\lambda}$.

In [12] a procedure for testing the adequacy of the HPP model, given event occurrence times over a monitoring interval, was presented. This procedure was called the V-test. Applying this V-test, we find the value of the statistic to be $V = 224.26$ with an associated p-value of 0.0003 for testing the null hypothesis that the HPP model holds. Thus, based on the *Mother Jones* data set, the HPP model is an inadequate model for the occurrences of mass shootings in the United States. The inadequate fit could also be noted from the fact that the line $\hat{\lambda}t$ is always above the graph of $(S_k, k), k = 0, 1, 2, \ldots, K$, where S_k is the time of occurrence of the kth mass shooting. If an HPP model is adequate, we would see that the straight line and the graph of $\{(S_k, k)\}$ will be close to each other. In fact, the observed plot appears to indicate that the inter-event times of the mass shootings are ominously getting stochastically shorter as time increases.

5 Dynamic Recurrent Event Models

Noting that the HPP model does not fit well the observed data, we now consider a more general dynamic model for the occurrences of mass shootings. The simplest parametric dynamic model that utilizes $N(s-)$ as the sole dynamic covariate is the Weibull dynamic regression model, which includes as a special case the exponential dynamic regression model (cf., [10]). This specifies that

$$\lambda_0(t;\theta = (\alpha,\eta)) = (\alpha\eta)(\eta t)^{\alpha-1} \quad \text{and} \quad I(s) = N(s-).$$

When $\alpha = 1$, then this is the dynamic exponential regression model. When $\lambda_0(\cdot)$ is simply assumed to be some hazard rate function, then we obtain a dynamic Cox proportional hazards model [4]. These models could be fitted easily using the `survreg` and `coxph` object functions in the `survival` library in the `R` statistical platform [14]. The results of these model fittings are provided below.

For the exponential dynamic regression model the fitted model has

$$\hat{\eta} = 0.003319 \quad \text{and} \quad \hat{\kappa} = 0.019541.$$

It is found that κ is significantly different from zero. In our initial fittings, we also included the number of fatalities of the preceding mass shooting, but this did not turn out to be a significant predictor, hence we did not include this dynamic covariate in the Weibull and Cox PH model fittings. For the Weibull dynamic regression model, the estimates of the parameters are

$$\hat{\alpha} = 1.1285, \hat{\eta} = 0.0031, \hat{\kappa} = 0.0215.$$

It is found that α is significantly different from 1.0 (the exponential baseline hypothesis), and κ is significantly different from 0.

Fitting the semi-parametric Cox proportional hazards model with $I(s) = N(s-)$ as the internal dynamic covariate, we find the estimate of the associated regression coefficient to be

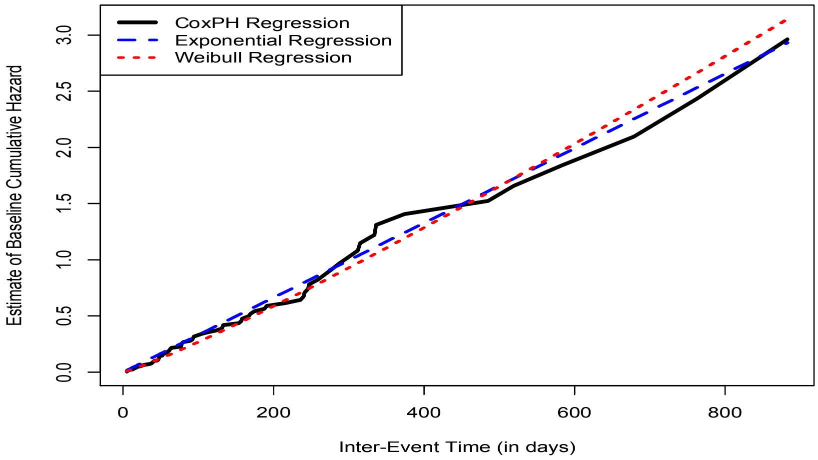
$$\hat{\kappa} = 0.02012.$$

A test of the hypothesis $\kappa = 0$ leads to the conclusion that the dynamic covariate is an important predictor.

In all of these fits, we note that the estimates of the regression coefficient of $N(s-)$ are all positive, indicating that there is an increase in the rate of occurrence of mass shootings with an increase in the number of occurrences of previous mass shootings. Of course, this could simply be that $N(s-)$ is a surrogate of calendar time and as time increases there is an increase in the rate of mass shootings, possibly due to an increase in the population of people or the higher availability of guns. See, for instance, [5,15,6].

For these three models, we also estimated the baseline cumulative hazard function, which is a functional parameter encoding the rate of mass shootings after correcting for the effect of time or prior mass shootings. Figure 4 provides the three estimates together. Observe that all three estimates are all close to being linear. However, note that the nonparametric (or semi-parametric) estimate, the so-called Aalen-Breslow-Nelson (ABN) estimate, obtained via the Cox PH model shows a bump over the linear and Weibull fits in the region from about 300 days to 500 days. This bump *may* be a manifestation of the so-called contagion effect with the fitted Weibull baseline serving as an ambient rate of the occurrence of mass shootings. However, we again point out that this is a difficult hypothesis to prove with the available data. At the same time this demonstrates a potential advantage of the semi-parametric model over the parametric Weibull model since

Fig. 4 Super-imposed plots of the estimated baseline cumulative hazard functions under the dynamic exponential, Weibull, and Cox proportional hazards models, with internal covariate being the number of prior mass shootings.



it may have the ability to potentially tease out such contagion effect. To further explore this issue, we present a plot of the estimates of the baseline hazard probabilities in Figure 5 and obtain associated kernel estimates of the baseline hazard rate function $\lambda_0(\cdot)$ based on this ABN estimate. The kernel estimator used is given by

$$\hat{\lambda}_0(t) = \int_0^\infty \frac{1}{h} K\left(\frac{v-t}{h}\right) d\hat{\Lambda}_0(v) = \sum_{l=1}^{L} \frac{1}{h} K\left(\frac{t_l - t}{h}\right) \hat{\lambda}_{0l}$$

where $t_l, l = 1, 2, \ldots, L$, are the jump times of the ABN estimate $\hat{\Lambda}_0$, which are $\hat{\lambda}_{0l} = \hat{\Lambda}_0(t_l) - \hat{\Lambda}_0(t_l-), l = 1, 2, \ldots, L$. These are the estimated baseline hazard probabilities depicted in Figure 5. $K(\cdot)$ is a kernel function which we set to be the Epanechnikov kernel

$$K(v) = (1 - v^2)^3 I\{|v| \leq 1\}.$$

In our implementation, we specified four values for the bandwidth h: 100, 200, 300, and 400. Figure 6 shows these kernel estimates of the baseline hazard rate $\lambda_0(\cdot)$. In all four estimates, we notice the increasing trend and the bump(s) in the hazard rate estimates from time zero until about 300 days, which indicate an increase in the chances of another mass shooting over this period just after a mass shooting. In fact, looking at the first estimate corresponding to the bandwidth $h = 100$, leading to the wiggliest estimate, we notice the first bump in the baseline hazard rate estimate at 71 days (hazard rate estimate of 0.002898) and the second bump at 296 days (hazard rate estimate of 0.004923). These first two bumps in the baseline hazard estimate may be an indication of the presence of a contagion effect in mass shootings.

Fig. 5 Estimates of the discrete hazard probabilities based on the ABN estimate of $\Lambda_0(\cdot)$.

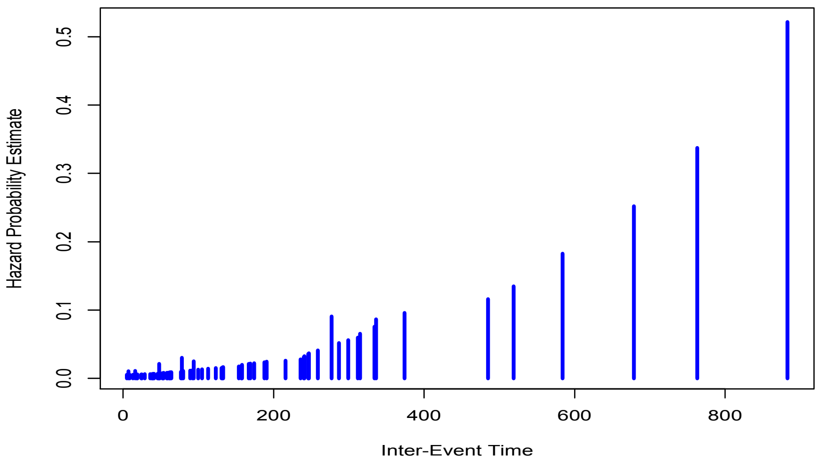
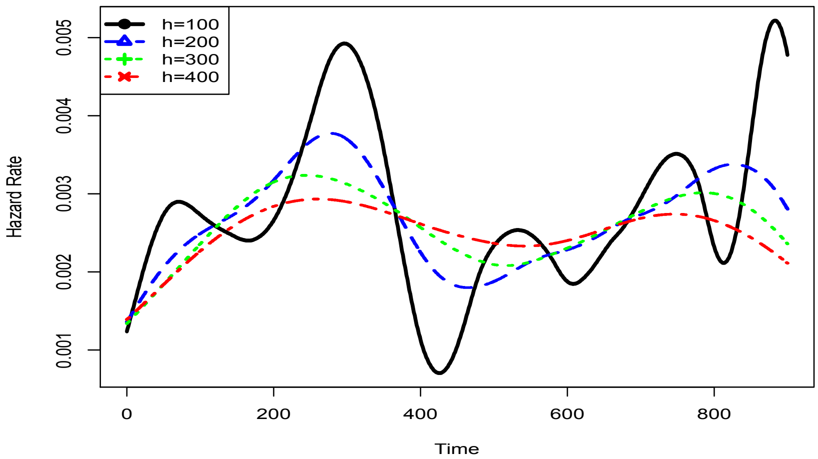


Fig. 6 Super-imposed plots of the kernel estimates of the baseline hazard rate function corresponding to four bandwidths. The bandwidths chosen were $h \in \{100, 200, 300, 400\}$ days.



There has been papers discussing the possibility of such a contagion effect. The paper [15] discussed the potential impact of media coverage after mass shootings and suicides in the context of increasing the incidence of subsequent mass shootings or suicides. The authors presented mathematical contagion models that tried to tease out the contagion effect arising from the enhanced media coverage. They also examined the impact of mental health illness and firearm availability in states. Based on the data sets that they analyzed about mass killings and school shootings they found a significant contagion effect. The *Pacific Standard* article [11] also examined the impact of intense media coverage of mass shootings in the context of an increasing incidence of mass shootings over time.

5.1 Bootstrapping Dynamic Recurrent Event Models

In this subsection we discuss the process of bootstrapping from these dynamic recurrent event models. This will enable us to estimate the standard errors of estimates, and also enable the assessment of significance in goodness-of-fit procedures discussed in the next subsection.

5.1.1 Dynamic Weibull Model

How do we generate the bootstrap samples? Such bootstrap sample generation should take into account the dynamic aspects of the event generation. Consider first the situation where the baseline hazard is Weibull. Based on the observed data, we are able to estimate the Weibull parameters (α, η) and the regression coefficient κ. Let the estimates be denoted by $(\hat{\alpha}, \hat{\eta}, \hat{\kappa})$. To generate one (parametric) bootstrap sample over the monitoring period $[0, \tau]$, we could implement the algorithm presented in appendix section B, which is coded in the `R` syntax. This algorithm incorporates the dynamicity of the event occurrences. Through this bootstrapping procedure we are able to obtain nonparametric estimates of the standard errors of estimates of the model parameters. For the fitted Weibull regression, using Mboot = 10000 bootstrap replications, we found the following bootstrap standard error (bSE's) estimates of the parameter estimates:

$$bSE(\hat{\alpha}) = 0.10888;\ bSE(\hat{\eta}) = 0.00061;\ \text{and } bSE(\hat{\kappa}) = 0.00618.$$

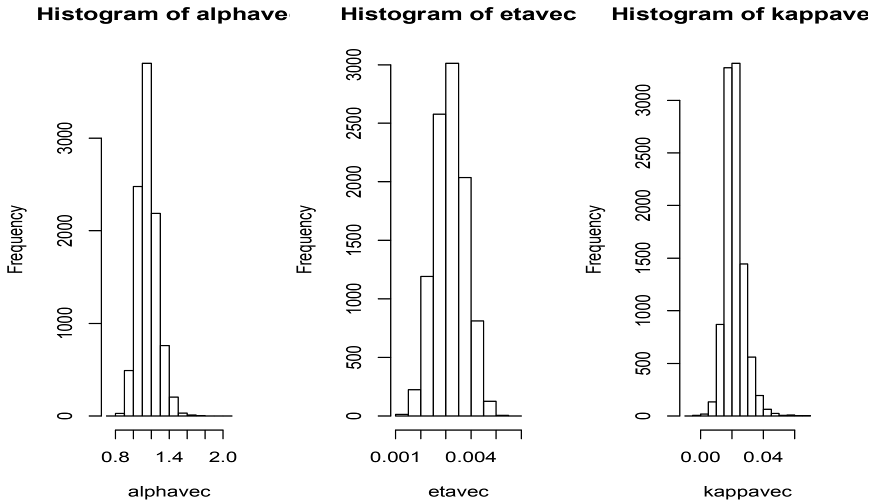
The histograms of the 10000 bootstrap estimates of α, η, and κ are provided in Figure 7. We note that in our bootstrap implementation we put an upper limit to the observed number of events in a bootstrap sample to 1000 (the `maxEvents` in the input to the algorithm). Theoretically, there is the possibility of explosion, that is, the number of events observed in a finite interval increases without bound (see, for instance, [7]) with positive probability. In practical situations, such an eventuality will not be observed, hence putting a large upper limit to the number of observed events is an acceptable solution to this 'blowing up' phenomenon, though this solution may induce some bias. However, we assess that the induced bias is negligible since the proportions of bootstrap samples reaching the upper limit of 1000 was very low out of the 10000 bootstrap samples.

5.1.2 Dynamic Cox Model

When a non-parametric baseline hazard rate function $\lambda_0(\cdot)$ is specified, the portion in the algorithm for the Weibull model containing the two code lines

```
U = runif(1)
T = (-exp(-kappa*Xcur)*log(U))^(1/alpha)/eta
```

Fig. 7 Histograms of the estimates of α, η, and κ based on the 10000 bootstrap samples under the dynamic Weibull regression model.



need to be replaced by generating a value from the observed inter-event times with probabilities induced by the ABN estimate of $\Lambda_0(\cdot)$. The replacement code line is presented after Lemma 1.

To amplify, denote the estimate of $\Lambda_0(\cdot)$ by $\hat{\Lambda}_0(\cdot)$ with jump points $v_1 < v_2 < \ldots < v_L$. For a covariate value of x, then the hazard probabilities under the dynamic model are

$$\hat{\lambda}_j(x) = \hat{\lambda}_{0j} \exp\{\hat{\kappa}x\}, j = 1, 2, \ldots, L, \tag{3}$$

where $\hat{\lambda}_{0j} = \hat{\Lambda}_0(v_j) - \hat{\Lambda}_0(v_{j-1})$ for $j = 1, 2, \ldots, L$ are the estimates of the baseline hazard probabilities at the observed complete inter-event times. It is possible that $\hat{\lambda}_j(x)$ as computed exceeds 1, so if this occurs we set the value to 1. However, these hazard probabilities need not induce a proper set of probabilities on the set $\{v_1, v_2, \ldots, v_L\}$ if $\hat{\lambda}_L(x) < 1$. To induce a proper set of probabilities, we always set $\hat{\lambda}_L(x) = 1$, equivalent to considering the largest observed gap-time as complete. That imposing this condition leads to a proper set of probabilities on $\{v_1, v_2, \ldots, v_L\}$ follows from the following lemma.

Lemma 1 *Let $\lambda_1, \lambda_2, \ldots, \lambda_L$, be such that $\lambda_j \in [0, 1], j = 1, 2, \ldots, L-1$, and $\lambda_L = 1$. Then, letting*

$$p_j = \left[\prod_{i=1}^{j-1} (1 - \lambda_i)\right] \lambda_j, j = 1, 2, \ldots, L,$$

we have $\sum_{j=1}^{L} p_j = 1$, that is, $p_1, p_2, \ldots, p_L$, determines a probability mass function.

Proof A simple inductive proof establishes the result, so we leave the proof as an exercise.

For these $\hat{\lambda}_j(x), j = 1, 2, \ldots, L$, with $\hat{\lambda}_L(x) = 1$, we then define the probabilities

$$\hat{p}_j(x) = \left\{ \prod_{l=1}^{j-1} [1 - \hat{\lambda}_j(x)] \right\} \hat{\lambda}_j(x), j = 1, 2, \ldots, L.$$

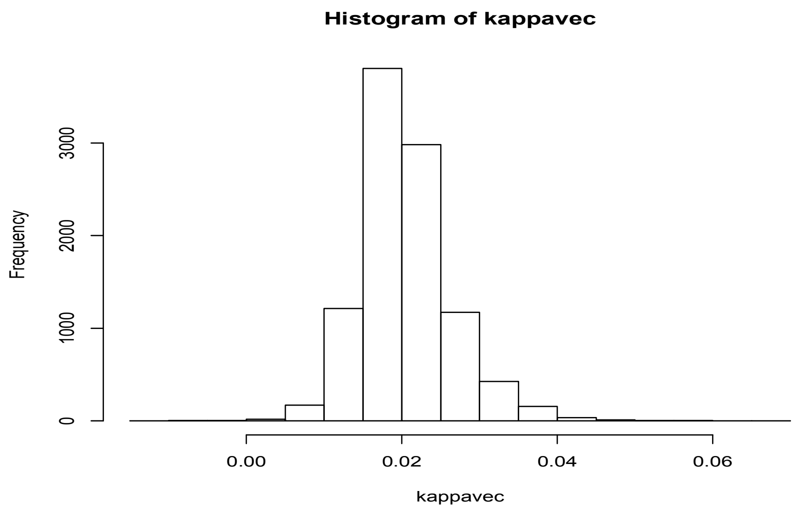
An observation is then generated from the set $\mathbf{v} = \{v_1, v_2, \ldots, v_L\}$ according to the probabilities $\mathbf{p} = (\hat{p}_1(x), \hat{p}_2(x), \ldots, \hat{p}_L(x))$ by using the `R` code

```
T = sample(v,1,prob=p)
```

with `x = Xcur` in the computation in (3). This is the command that replaces the two code lines mentioned above to generate an observation from the semi-parametric dynamic Cox model. The algorithm in `R` syntax is provided in appendix section C. It takes as input the arguments `timein` and `hazin`, which are the distinct complete inter-event times and the baseline hazard probability estimates, respectively, from the ABN estimate.

In this nonparametric bootstrap, the issue of explosion does not arise since there are just a finite number of possible positive values of the generated inter-event time. Also, as in the parametric model, we could use this bootstrap procedure to obtain an estimate of the standard error of the estimator of κ, whose estimate is $\hat{\kappa} = 0.02012$. The bootstrap estimate of its standard error, based on 10000 bootstrap replications, is $bSE(\hat{\kappa}) = 0.00582$. The estimate of the bootstrap distribution of the estimator of κ based on the 10000 bootstrap replications is displayed in Figure 8.

Fig. 8 Histogram of the estimates of κ based on the 10000 bootstrap samples under the dynamic Cox regression model.



5.2 Goodness-of-Fit Testing of Fitted Models

In the simple dynamic models that were fitted, if we partition the monitoring period $(0, \tau]$ into $L+1$ intervals given by $0 \equiv t_0 < t_1 < t_2 < \ldots < t_L < t_{L+1} \equiv \tau$, then the number of events observed in the interval $I_l = (t_{l-1}, t_l]$ is $O_l = N(t_l) - N(t_{l-1})$. If the assumed model holds, then the estimated expected number of events in the interval I_l is given by

$$\begin{aligned}
\hat{E}_l &= \int_{t_{l-1}}^{t_l} \hat{\lambda}_0[\mathcal{E}(v)] e^{\hat{\kappa} N(v-)} dv \\
&= \sum_{j=1}^{K+1} \int_{\max(t_{l-1}, S_{j-1})}^{\min(t_l, S_j)} I\{\max(t_{l-1}, S_{j-1}) < \min(t_l, S_j)\} \times \\
&\quad \hat{\lambda}_0(v - S_{N(v-)}) e^{\hat{\kappa} N(v-)} dv \\
&= \sum_{j=1}^{K+1} e^{\hat{\kappa}(j-1)} I\{\max(t_{l-1}, S_{j-1}) < \min(t_l, S_j)\} \times \\
&\quad \int_{\max(t_{l-1}, S_{j-1}) - S_{j-1}}^{\min(t_l, S_j) - S_{j-1}} \hat{\lambda}_0(w) dw \\
&= \sum_{j=1}^{K+1} e^{\hat{\kappa}(j-1)} I\{\max(t_{l-1}, S_{j-1}) < \min(t_l, S_j)\} \times \\
&\quad \left[\hat{\Lambda}_0(\min(t_l, S_j) - S_{j-1}) - \hat{\Lambda}_0(\max(t_{l-1}, S_{j-1}) - S_{j-1})\right].
\end{aligned}$$

For the Weibull baseline, we will have

$$\hat{\Lambda}_0(t) = (\hat{\eta} t)^{\hat{\alpha}} I\{t > 0\};$$

whereas, for the non-parametrically specified baseline, we will use the ABN estimate of $\Lambda_0(\cdot)$ from the `coxph` fit to evaluate $\hat{\Lambda}_0(t)$.

Analogously to the goodness-of-fit test of Akritas [1] (see also the goodness-of-fit procedure in [8,9]), we may use the Pearson-type test statistic

$$Q^2 = \sum_{l=1}^{L+1} \frac{(O_l - \hat{E}_l)^2}{\hat{E}_l}$$

for assessing the goodness-of-fit of the fitted model. Significance of the observed value of Q^2 could be assessed by comparing to a chi-square distribution with degrees-of-freedom L minus the number of estimated parameters, or by generating bootstrap samples and obtaining an estimate of the null (that is, that the assumed model is adequate) sampling distribution of the Q^2 statistic.

A bootstrap assessment of the significance of the Q^2 appears to be more reliable in this case since if we utilize the ML estimates of the parameters based on the ungrouped data, then the chi-square distribution may not be the appropriate approximate distribution to use. There is also the un-examined impact of the randomness of K, the number of events observed over the monitoring period. We believe that the bootstrap approximation of the P-value automatically incorporates these issues. There is also the subjective decision of how many intervals

Table 2 The interval upper boundaries together with the observed frequencies and estimated expected frequencies under the dynamic Weibull model and the dynamic Cox model in the implementation of the Pearson-type goodness-of-fit for the fitted models. The lower boundary of the first interval is zero.

Interval Upper Boundary	Observed Freq Frequency	Expected Freq (Weibull)	Expected Freq (Cox)
1398.11	2	4.899	4.328
2796.22	5	9.647	9.418
4194.33	10	8.387	8.240
5592.44	4	9.270	9.028
6990.55	11	10.191	10.735
8388.67	4	13.479	12.491
9786.78	12	12.989	13.537
11184.89	14	15.065	14.625
12583.00	16	19.553	19.169
		$Q^2 = 14.78$ (Bootstrap-$P = .07$)	$Q^2 = 13.00$ (Bootstrap-$P = .16$)

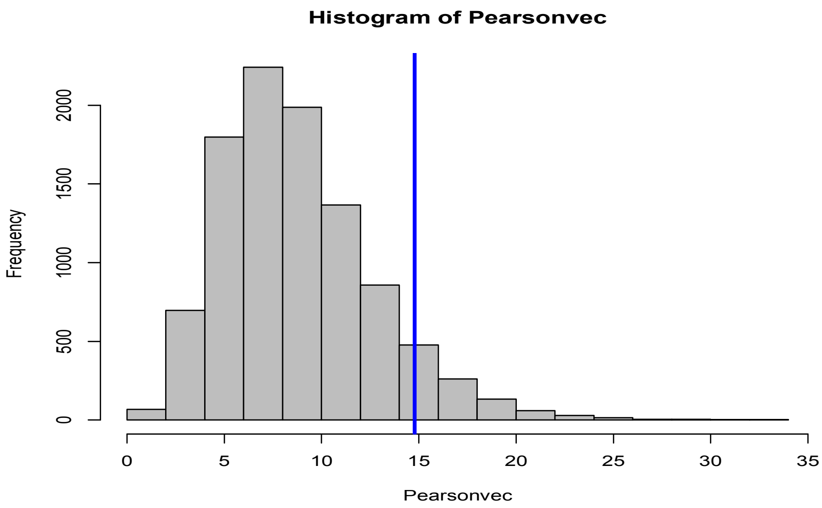
in the partition and what the boundaries should be. The simplest, but which may not be best, is to use equal length intervals, and to have $L \approx \sqrt{K}$ intervals in the partition. Clearly, these issues require more in-depth theoretical studies. Our goal in this paper is to simply utilize a simple *ad hoc* goodness-of-fit procedure to assess the adequacy of the fitted models.

We implemented these ideas by developing appropriate object functions in `R` [14]. When we applied to the mass shooting data set, we used 10000 bootstrap replications, equal-length partition, and with $L = 8$. Table 2 presents the intervals together with the observed frequency and the estimated expected frequencies under the dynamic Weibull model and dynamic Cox model. Indicated in the last two rows of this table are the values of Q^2 together with their bootstrap P-values. For the dynamic Weibull model, we find $Q^2 = 14.7870$ and the associated bootstrap P-value is 0.0769. This P-value is close to 0.05, possibly indicating that the dynamic Weibull model is not the most appropriate model. Figure 9 presents the histogram of the Q^2 goodness-of-fit statistic for the 10000 bootstrap samples under the Weibull model. For the dynamic Cox model, we find $Q^2 = 13.0051$ with an associated P-value of 0.1659. This appears to indicate that the dynamic Cox model is a better fit to the data than the dynamic Weibull model. We hypothesize that this could be due to the fact that the Weibull model will not be able to model a contagion effect or a bump in the baseline hazard, whereas the dynamic Cox model with a nonparametrically-specified baseline hazard will be able to do so. But this remains a hypothesis and this will be difficult to validate with the existing data. We also mention that among the 10000 bootstrap replications, there were some outlying values of the Q^2-statistic under the dynamic Cox model. This could be seen from the histograms of the Q^2 and the logarithm of this Q^2 statistic which are both depicted in Figure 10.

6 Forecasting Mass Shootings

An oft-quoted saying, attributed to different people (Nostradamus, Niels Bohr, Mark Twain, Yogi Berra, others), is that:

Fig. 9 Histograms of the Q^2 goodness-of-fit statistic based on the 10000 bootstrap samples under the dynamic Weibull regression model. The vertical blue line corresponds to the observed Q^2-statistic.



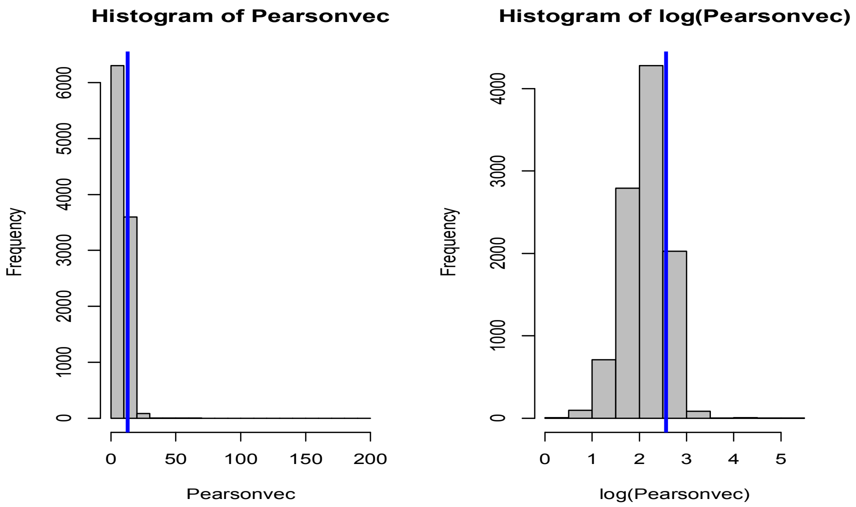
It's Difficult to Make Predictions, Especially About the Future.

And so it is the case with predicting or forecasting when the next mass shooting in the United States is going to occur, even with the aid of the stochastic models fitted to the observed data. Mass shootings are as unpredictable as earthquakes. However, these fitted models provide some guidance on future mass shooting occurrences. For instance, starting from February 1, 2017, one may inquire about the probability that at least one mass shooting will occur in the US during the next four months, that is, until May 31, 2017, which covers 120 days, given the information until January 31, 2017. Using the fitted dynamic Cox model, an estimate of this probability is

$$\begin{aligned}
&\widehat{\Pr}\{25 < T^* \leq 146|\text{data until } 1/31/2017\} \\
&= 1 - \widehat{\Pr}\{T^* > 146|\text{data until } 1/31/2017\} \\
&= 1 - \left[\hat{\bar{F}}_0(146)/\hat{\bar{F}}_0(25)\right]^{\exp\{(\hat{\kappa})(78)\}} \\
&= 1 - \left[\prod_{\{l:\ 25<t_l\leq 146\}} [1 - \hat{\lambda}_{0l}]\right]^{\exp\{(.02012)(78)\}} \\
&= 1 - 0.1780 \\
&= 0.8220,
\end{aligned}$$

where T^* represents the time-to-occurrence of the next mass shooting starting from January 6, 2017, the time of last mass shooting prior to February 1, 2017. The value of 25 is the number of days from 1/6/2017 until 2/1/2017, while the

Fig. 10 Histograms of the Q^2 and $\log(Q^2)$ goodness-of-fit statistics based on the 10000 bootstrap samples under the dynamic Cox regression model. The vertical blue line corresponds to the observed Q^2 and $\log(Q^2)$ statistics.



value of 146 is the number of days from 1/6/2017 until 5/31/2017. Note that the division by $\hat{\hat{F}}_0(25)$ in the second line is because information up to 1/31/2017 indicated that $T^* > 25$, which becomes a conditioning event.

Thus, there is a not-so-insignificant probability of about 82% of at least one mass shooting occurring during the period from February 1, 2017 to May 31, 2017, given the information up to January 31, 2017. Of course, from the perspective of helping to prevent the occurrence of a mass shooting, the probability estimate above will not be of direct help since it does not pinpoint when and where a mass shooting will occur nor could it help in identifying potential mass shooter(s). [**Note:** As of May 1, 2017, the date of initial draft of this manuscript, there has indeed been at least one event since February 1, 2017 that qualifies as a mass shooting. The first occurred last February 6, 2017 in Yazoo City, Mississippi claiming four victims, and another one at Toomsuba, Mississippi last February 21, 2017, also with four victims.]

7 Concluding Remarks

In this paper we analyzed data pertaining to the commission of a mass shooting in the United States spanning the period from August 20, 1982 to January 31, 2017. The data was compiled by the magazine *Mother Jones*, and it includes mass shootings with at least four fatalities, excluding the shooter or shooters. The advent of a mass shooting is one of the most unnerving events in our society. The analysis performed in this paper utilized dynamic event-time models, dynamic in the sense that the occurrence of a mass shooting depends to some extent on what has transpired before. We fitted four models: the homogeneous Poisson process, a dynamic exponential regression model, a dynamic Weibull regression model, and a dynamic Cox model, the latter being a semi-parametric model. We found that the first two models did not fit the data well, while the Weibull-based and the Cox-based models offer better fits to the observed data, with the Cox-based model having the advantage in that it potentially detects a contagion effect through the nonparametric baseline hazard. A contagion effect is one when the rate of occurrence of a mass shooting bumps up a certain period after the last mass shooting, and this has been a topic of interest in several papers dealing with mass shootings. However, we emphasize that it is difficult to validate the presence of this contagion effect on the basis of available data. Both the dynamic Weibull regression and dynamic Cox regression fitted models indicate that the number of prior mass shootings has an effect in terms of the waiting-time for the next occurrence of a mass shooting, with this effect being to stochastically shorten such waiting time. It is certainly conceivable that the number of prior mass shootings serves as a surrogate to calendar time and that as time progresses, the rate of incidence of mass shootings also increases owing to an increase in availability of guns or due to an increasing population size.

We also proposed methods for validating the fitted models through Pearson-type goodness-of-fit tests. However, the determination of the significance of the observed values of these Pearson-type test statistics is done via bootstrapping methods. This led to an examination of the proper way in which to generate bootstrap samples that incorporates the dynamic mechanisms in which events occur. Using the proposed goodness-of-fit methods, we found that the Cox-based model best fits the observed data, though the procedure did not lead to rejection (at level of significance 5%) of the dynamic Weibull model. We also discussed the issue of forecasting the advent of a mass shooting using the fitted Cox dynamic model, with the caveat that the fitted models will not truly be of practical value in terms of preventing when and where a mass shooting will occur nor in pinpointing potential mass shooter(s).

Further studies are warranted regarding the modeling and analysis of mass shootings. More elaborate dynamic models incorporating information from external covariates and possibly with additional information other than those provided by the *Mother Jones* data set will clearly be of interest. It would also be interesting to change the internal covariate from the number of mass shootings that have occurred *since* the time origin to the number of mass shootings that have occurred *within* a certain period, say two years, prior to the time under consideration.

References

1. Akritas, M.: Pearson-type goodness-of-fit tests: the univariate case. Journal of the American Statistical Association **83**, 222–230 (1988)
2. Andersen, P., Borgan, O., Gill, R., Keiding, N.: Statistical Models Based on Counting Processes. Springer-Verlag, New York (1993)
3. Berkowitz, B., Gamio, L., Lu, D., Uhrmacher, K., LIndeman, T.: The math of mass shootings. The Washington Post (January 8, 2017)
4. Cox, D.: Regression models and life tables (with discussion). Journal of the Royal Statistical Society **34**, 187–220 (1972)
5. Duwe, G.: Mass Murder in the United States: A History. McFarland & Company, Inc., North Carolina (2007)

6. Follman, M., Aronsen, G., Pan, D.: A guide to mass shootings in America. Mother Jones (January 8, 2017)
7. Gjessing, H.K., Røysland, K., Peña, E.A., Aalen, O.O.: Recurrent events and the exploding Cox model. Lifetime Data Anal. **16**(4), 525–546 (2010)
8. Hjort, N.: Goodness-of-fit tests in models for life history data based on cumulative hazard rates. Annals of Statistics **18**, 1221–1258 (1990)
9. Hollander, M., Peña, E.A.: A chi-squared goodness-of-fit test for randomly censored data. J. Amer. Statist. Assoc. **87**(418), 458–463 (1992)
10. Kalbfleisch, J.D., Prentice, R.L.: The statistical analysis of failure time data, second edn. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ (2002)
11. Keller, J.: Does the media cause mass shootings? Pacific Standard (October 3, 2016)
12. Liu, P., Peña, E.A.: Sojourning With the Homogeneous Poisson Process. Amer. Statist. **70**(4), 413–423 (2016)
13. Peña, E.A., Slate, E.H., González, J.R.: Semiparametric inference for a general class of models for recurrent events. J. Statist. Plann. Inference **137**(6), 1727–1747 (2007)
14. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013). ISBN 3-900051-07-0
15. Towers, S., Gomez-Lievano, A., Khan, M., Mubayi, A., Castillo-Chavez, C.: Contagion in mass killings and school shootings. PLOS One **10**(7), 1–12 (2015)

A *Mother Jones* Mass Shootings Data Set

In the data set below, the variable name abbreviations are: F = Fatalities; T = NumDaysBetw; and S = NumDaysFromFirst. The time origin coincides with 8/20/82, so the value of S corresponds to the number of days elapsed since this date.

	Date	DayOfWeek	Location	F	T	S
1	6/29/84	Thursday	Dallas, Texas	6	679	679
2	7/18/84	Tuesday	San Ysidro, California	22	19	698
3	8/20/86	Tuesday	Edmond, Oklahoma	15	763	1461
4	4/23/87	Wednesday	Palm Bay, Florida	6	246	1707
5	2/16/88	Monday	Sunnyvale, California	7	299	2006
6	1/17/89	Monday	Stockton, California	6	336	2342
7	9/14/89	Wednesday	Louisville, Kentucky	9	240	2582
8	6/18/90	Sunday	Jacksonville, Florida	10	277	2859
9	10/16/91	Tuesday	Killeen, Texas	24	485	3344
10	11/1/91	Thursday	Iowa City, Iowa	6	16	3360
11	11/14/91	Wednesday	Royal Oak, Michigan	5	13	3373
12	5/1/92	Thursday	Olivehurst, California	4	169	3542
13	10/15/92	Wednesday	Watkins Glen, New York	5	167	3709
14	7/1/93	Wednesday	San Francisco, California	9	259	3968
15	8/6/93	Thursday	Fayetteville, North Carolina	4	36	4004
16	12/7/93	Monday	Garden City, New York	6	123	4127
17	12/14/93	Monday	Aurora, Colorado	4	7	4134
18	6/20/94	Sunday	Fairchild Air Force Base, Washington	5	188	4322
19	4/3/95	Sunday	Corpus Christi, Texas	6	287	4609
20	2/9/96	Thursday	Fort Lauderdale, Florida	6	312	4921
21	9/15/97	Sunday	Aiken, South Carolina	4	584	5505
22	12/18/97	Wednesday	Orange, California	5	94	5599
23	3/6/98	Thursday	Newington, Connecticut	5	78	5677
24	3/24/98	Monday	Jonesboro, Arkansas	5	18	5695
25	5/21/98	Wednesday	Springfield, Oregon	4	58	5753
26	4/20/99	Monday	Littleton, Colorado	13	334	6087
27	7/29/99	Wednesday	Atlanta, Georgia	9	100	6187
28	9/15/99	Tuesday	Fort Worth, Texas	8	48	6235
29	11/2/99	Monday	Honolulu, Hawaii	7	48	6283
30	12/30/99	Wednesday	Tampa, Florida	5	46	6329

31	12/26/00	Tuesday	Wakefield, Massachusetts	7	374	6703
32	2/5/01	Monday	Melrose Park, Illinois	5	41	6744
33	7/8/03	Tuesday	Meridian, Mississippi	7	883	7627
34	12/8/04	Wednesday	Columbus, Ohio	5	519	8146
35	3/12/05	Saturday	Brookfield, Wisconsin	7	94	8240
36	3/21/05	Monday	Red Lake, Minnesota	10	9	8249
37	1/30/06	Monday	Goleta, California	8	315	8564
38	3/25/06	Saturday	Seattle, Washington	7	54	8618
39	10/2/06	Monday	Lancaster County, Pennsylvania	6	191	8809
40	2/12/07	Monday	Salt Lake City, Utah	6	133	8942
41	4/16/07	Monday	Blacksburg, Virginia	32	63	9005
42	10/7/07	Sunday	Crandon, Wisconsin	6	174	9179
43	12/5/07	Wednesday	Omaha, Nebraska	9	59	9238
44	2/7/08	Thursday	Kirkwood, Missouri	6	64	9302
45	2/14/08	Thursday	DeKalb, Illinois	6	7	9309
46	6/25/08	Wednesday	Henderson, Kentucky	6	132	9441
47	3/29/09	Sunday	Carthage, North Carolina	8	277	9718
48	4/3/09	Friday	Binghamton, New York	14	5	9723
49	11/5/09	Thursday	Fort Hood, Texas	13	216	9939
50	11/29/09	Sunday	Parkland, Washington	4	24	9963
51	8/3/10	Tuesday	Manchester, Connecticut	9	247	10210
52	1/8/11	Saturday	Tucson, Arizona	6	158	10368
53	9/6/11	Tuesday	Carson City, Nevada	5	241	10609
54	10/14/11	Friday	Seal Beach, California	8	38	10647
55	2/22/12	Wednesday	Norcross, Georgia	5	131	10778
56	4/2/12	Monday	Oakland, California	7	40	10818
57	5/20/12	Sunday	Seattle, Washington	6	48	10866
58	7/20/12	Friday	Aurora, Colorado	12	61	10927
59	8/5/12	Sunday	Oak Creek, Wisconsin	7	16	10943
60	9/27/12	Thursday	Minneapolis, Minnesota	7	53	10996
61	12/14/12	Friday	Newtown, Connecticut	28	78	11074
62	3/13/13	Wednesday	Herkimer County, New York	5	89	11163
63	4/21/13	Sunday	Federal Way, Washington	5	39	11202
64	6/7/13	Friday	Santa Monica, California	6	47	11249
65	7/26/13	Friday	Hialeah, Florida	7	49	11298
66	9/16/13	Monday	Washington, D.C.	12	52	11350
67	2/20/14	Thursday	Alturas, California	4	157	11507
68	5/23/14	Friday	Santa Barbara, California	6	92	11599
69	10/24/14	Friday	Marysville, Washington	5	154	11753
70	6/17/15	Wednesday	Charleston, South Carolina	9	236	11989
71	7/16/15	Thursday	Chattanooga, Tennessee	5	29	12018
72	10/1/15	Thursday	Roseburg, Oregon	9	77	12095
73	12/2/15	Wednesday	San Bernardino, California	14	62	12157
74	2/20/16	Saturday	Kalamazoo County, Michigan	6	80	12237
75	6/12/16	Sunday	Orlando, Florida	49	113	12350
76	7/7/16	Thursday	Dallas, Texas	5	25	12375
77	9/23/16	Friday	Burlington, WA	5	78	12453
78	1/6/17	Friday	Fort Lauderdale, Florida	5	105	12558
79	<NA>	<NA>	<NA>	NA	25	12583

B Bootstrap Sampling Algorithm: Dynamic Weibull Model

```
BootSampWeibull <-
 function(alpha=1.1287,eta=0.0031,kappa=0.02146,tau=12583,maxEvents=1000)
{

S = 0
Svec = 0
Tvec = NULL
```

```
Deltavec = NULL
Xvec = 0
Xcur = 0
K = 0

ok = TRUE

while(ok) {

 U = runif(1)
 T = (-exp(-kappa*Xcur)*log(U))^(1/alpha)/eta

 if((S+T) < tau) {
  K = K + 1
  if(K > maxEvents) {ok=FALSE; print("Explosion Occurring!")} #cut-off explosion
  S = S + T
  Xcur = Xcur + 1
  Svec = c(Svec,S)
  Xvec = c(Xvec,Xcur)
  Tvec = c(Tvec,T)
  Deltavec = c(Deltavec,1)
 }
 else {
  ok = FALSE
  Svec = c(Svec,tau)
  Xvec = c(Xvec,Xcur)
  Tvec = c(Tvec,tau-S)
  Deltavec = c(Deltavec,0)
 }
}

return(list(alpha=alpha,eta=eta,kappa=kappa,tau=tau,
 K=K,Svec=Svec,Tvec=Tvec,Deltavec=Deltavec,Xvec=Xvec))
}
```

C Bootstrap Sampling Algorithm: Dynamic Cox Model

```
BootSampCoxPH <-
 function(timein=time,hazin=haz,kappa=0.02012,tau=12583)
{

L = length(timein)
probs = rep(0,L)

S = 0
Svec = 0
Tvec = NULL
Deltavec = NULL
Xvec = NULL
Xcur = 0
K = 0

ok = TRUE

while(ok) {

curhaz = hazin*exp(kappa*Xcur)
for(l in 1:L) {curhaz[l] = min(c(1,curhaz[l]))}
curhaz[L] = 1
```

```
probs[1] = curhaz[1]
for(l in 2:L) {probs[l] = probs[l-1]*(1/curhaz[l-1] - 1)*curhaz[l]}

T = sample(timein,1,prob=probs)

 if((S+T) < tau) {
  K = K + 1
  S = S + T
  Xcur = Xcur + 1
  Svec = c(Svec,S)
  Xvec = c(Xvec,Xcur)
  Tvec = c(Tvec,T)
  Deltavec = c(Deltavec,1)
 }
 else {
  ok = FALSE
  Svec = c(Svec,tau)
  Xvec = c(Xvec,Xcur)
  Tvec = c(Tvec,tau-S)
  Deltavec = c(Deltavec,0)
 }
}
return(list(kappa=kappa,tau=tau,K=K,Svec=Svec,Tvec=Tvec,
 Deltavec=Deltavec,Xvec=Xvec,tau=tau))
}
```