Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

# Models for surveillance data under reporting delay: applications to US veteran first-time suicide attempters

Y. Xia, N. Lu, I. Katz, R. Bossarte, J. Arora, H. He, J.X. Tu, B. Stephens, A. Watts & X.M. Tu

**To cite this article:** Y. Xia, N. Lu, I. Katz, R. Bossarte, J. Arora, H. He, J.X. Tu, B. Stephens, A. Watts & X.M. Tu (2015) Models for surveillance data under reporting delay: applications to US veteran first-time suicide attempters, Journal of Applied Statistics, 42:9, 1861-1876, DOI: 10.1080/02664763.2015.1014885

**To link to this article:** http://dx.doi.org/10.1080/02664763.2015.1014885

Published online: 27 Feb 2015.

*Journal of Applied Statistics*, 2015
Vol. 42, No. 9, 1861–1876, http://dx.doi.org/10.1080/02664763.2015.1014885



# Models for surveillance data under reporting delay: applications to US veteran first-time suicide attempters

Y. Xia[a,b*], N. Lu[b,c], I. Katz[b], R. Bossarte[b,d], J. Arora[a,b], H. He[a,b], J.X. Tu[a,d], B. Stephens[b], A. Watts[a,b] and X.M. Tu[a,b,d]

[a]*Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA;* [b]*Center of Excellence for Suicide Prevention, Canandaigua VA Medical Center, Canandaigua, NY 14424, USA;* [c]*Department of Management, Harbin Institute of Technology, Harbin, People's Republic of China;* [d]*Department of Psychiatry, University of Rochester, Rochester, NY 14642, USA*

(*Received 5 October 2013; accepted 30 January 2015*)

Surveillance data provide a vital source of information for assessing the spread of a health problem or disease of interest and for planning for future health-care needs. However, the use of surveillance data requires proper adjustments of the reported caseload due to underreporting caused by reporting delays within a limited observation period. Although methods are available to address this classic statistical problem, they are largely focused on inference for the reporting delay distribution, with inference about caseload of disease incidence based on estimates for the delay distribution. This approach limits the complexity of models for disease incidence to provide reliable estimates and projections of incidence. Also, many of the available methods lack robustness since they require parametric distribution assumptions. We propose a new approach to overcome such limitations by allowing for separate models for the incidence and the reporting delay in a distribution-free fashion, but with joint inference for both modeling components, based on functional response models. In addition, we discuss inference about projections of future disease incidence to help identify significant shifts in temporal trends modeled based on the observed data. This latter issue on detecting 'change points' is not sufficiently addressed in the literature, despite the fact that such warning signs of potential outbreak are critically important for prevention purposes. We illustrate the approach with both simulated and real data, with the latter involving data for suicide attempts from the Veteran Healthcare Administration.

**Keywords:** functional response models; inverse probability weighting; prediction interval; truncation; Veteran Healthcare Administration

## 1. Introduction

The US Department of Veteran Affairs has recently stepped up its effects to develop and implement diverse strategies to reduce deaths from suicide. One major component of the prevention

*Corresponding author. Email: yinglin.xia2007@gmail.com

efforts is to identify Veterans at elevated risk for self-harm to ensure that such Veterans immediately receive needed mental health services by collecting data every month for Veteran first-time suicide attempters that had recent, prior Veteran Healthcare Administration (VHA) service utilization. However, the use of surveillance data such as this monthly updated VHA database requires adjustments of the underreported caseload caused by the delay in reporting within an observation time frame. This 'truncation' issue is well known and the topic is widely researched, especially in the late 1980s and early 1990s AIDS research [1–3,5,6,8,10,13,18]. Various methods have been developed to address this classic problem. However, these approaches mostly focus on inference for the reporting delay distribution, with estimates of caseloads based on some ad hoc methods. As a result, adjusted estimates of caseload for the observation period and projections of future incidence are limited by the models used for the reporting delay. In addition, many available methods assume parametric distributions such as the multinomial and Poisson, further limiting their utility when applied real surveillance systems.

In this paper, we discuss a new approach to address such issues by allowing for separate models for the caseload and the reporting delay distributions. This new approach not only permits a more complex model for the disease incidence, but also accommodates the situation when the time of reporting is only available for a subsample, a common scenario in dealing with surveillance data. Furthermore, this approach requires no parametric assumption for either modeling component, enabling robust inference for a much wider class of data distributions in practice.



## 2. Models for suicide attempters

We discuss our approach within the context of US Veteran first-time suicide attempters. But the same considerations apply to other disease surveillance systems as well.

There are 139 VHA Facilities, which are administratively grouped into 21 Veterans Integrated Service Networks (VISN), based primarily on the geographic locations. Because of high levels of security for data use within the VHA, only facility-level information is available for our study. Thus, we focus on Veteran first-time suicide attempters within each facility, although the same considerations apply to modeling such cases at the VISN-level as well.

Let $i$ index the month, $y_i$ denote the number of Veteran first-time suicide attempters, $m_i$ the size, and $\mathbf{x}_i$ the facility-level covariates for a facility in the $i$th month. Since $y_i$ is the number of new attempters and $m_i$ is much larger than $y_i$, the $y_i$'s may be viewed as independent observations within each facility. Under this assumption, log-linear models for count responses may be used to model $y_i$. A major limitation of parametric methods such as the Poisson and negative binomial (NB) log-linear models is the lack of robust inference when assumed mathematical distributions such as the Poisson do not fit the data well [20]. For example, the Poisson model does not apply to over-dispersed count data, while the NB, although addressing overdispersion, does not provide valid inference if the over-dispersed data do not follow the NB model [20].

For robust inference, the most popular approach is to only specify the first moment, i.e.

$$\mu_i = E(y_i \mid \mathbf{x}_i), \quad \log(\mu_i) = \log m_i + \mathbf{x}_i^\top \boldsymbol{\beta}, \quad 1 \le i \le n, \tag{1}$$

where $\log m_i$ is the offset term. Indeed, the above is unaltered whether $y_i$ follows a Poisson, NB or any other distribution, so long as the conditional mean $\mu_i$ of $y_i$ given $\mathbf{x}_i$ satisfies Equation (1). Thus, this distribution-free, or semi-parametric, model provides valid inference for a much wider class of data distributions than their parametric counterparts [14,20].

It follows from Equation (1) that $\mu_i = m_i \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$, where $r_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ is the 'rate' of first-time suicide attempt per individual within the facility in the $i$th month ($1 \le i \le n$). Thus, by controlling for facility size using the offset term, we can interpret $\boldsymbol{\beta}$ as the effect of $\mathbf{x}_i$ on the rate of incidence for the facility.

One popular approach for inference for the model in Equation (1) is to solve for $\boldsymbol{\beta}$ the following estimating equations (EE):

$$\mathbf{w}(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i V_i^{-1}(\mathbf{x}_i)(y_i - \mu_i) = \sum_{i=1}^{n} D_i V_i^{-1}(\mathbf{x}_i) S_i = \mathbf{0}, \tag{2}$$

where $D_i = (\partial/\partial\boldsymbol{\beta})\mu_i$, $S_i = y_i - \mu_i$, and $V_i(\mathbf{x}_i)$ is some function of $\mathbf{x}_i$ only. If $V_i(\mathbf{x}_i) = \mu_i$, the EE above is actually identical to the score equations used by the ML, thereby yielding the same estimate as the MLE [20]. However, unlike the MLE, $\hat{\boldsymbol{\beta}}$ remains consistent and asymptotically normal, regardless of the distribution of $y_i$ [14,20].



## 3. Reporting delays and models for delay distributions

Like most surveillance systems, delay in reporting is also a serious issue for the VHA surveillance database. Shown in Figure 1 are the percentages ($y$-axis) of Veteran first-time suicide attempters reported by 2 (short dashed line), 3 (medium dashed line) and 4 (long dashed line) months after the incident occurred, starting with the facility with most severe reporting delay to the one with most efficient reporting ($x$-axis). The patterns show a high degree of heterogeneity in reporting delays across the facilities; about 20 facilities reported less than 70% of the cases 2 months after the attempts took place. Among these facilities, about half reported even less than 80% of the cases 4 months after the attempt occurred.

Various approaches have been proposed for estimating the reporting delay distribution and using the estimate to adjust for underreporting when modeling the incidence. Available methods

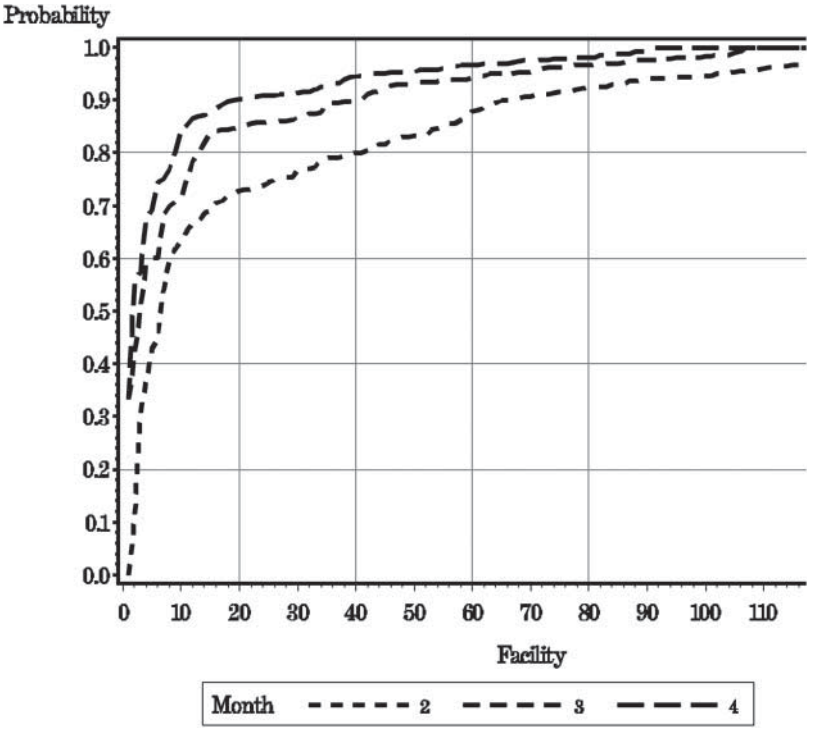


Figure 1. The percent of Veteran first-time suicide attempters reported in 2, 3 and 4 months after the attempt occurred by the number of facilities across all VISNs.

can be largely grouped into two major categories. One is to treat the reporting delay as the time between the occurrence of the event and the reporting of the case and utilize survival methods to estimate the delay distribution [3,6,10,13]. The other is to view the observed cases as a discrete outcome and use models for discrete responses such as the Poisson and multinomial to model the delay distribution [1,2,5,8,18]. The former typically focuses on inference about the reporting delay distribution, while the latter on disease incidence. We give a brief review of each approach below.

Within our context, let $d$ denote the lag time in months between the occurrence of the suicide attempt and reporting of the case. Thus, if the attempt in the $i$th month does not get reported until $d$ months later, the reporting time is the $(i+d)$th month $(1 \le i \le n)$. The attempter is observed only if $i + d \le n$, or equivalently $0 \le d \le n - i$.

Let $F_i(t \mid \mathbf{u}_i) = \Pr(d \le t \mid \mathbf{u}_i)$ denote the cumulative probability function (CDF) of $d$, conditional on some facility-level covariates $\mathbf{u}_i$. We assume that the maximum delay is $n_0$ $(< n-1)$ months, i.e. $F_i(n_0 \mid \mathbf{u}_i) = 1$. Under this assumption, all attempts that occur prior to the $(n-n_0)$th month will have been reported by month $n$, while for those who attempt suicide between the $(n-n_0)$th and $n$th month, only a fraction of those with reporting delays $\le n - i$ will have been reported. The first approach treats $d$ as a 'survival time' and models $F_i(t \mid \mathbf{u}_i)$ using survival methodology. We focus on one particular implementation of this approach to show how it works and highlight its key differences from the second alternative. Note that $\mathbf{u}_i$ is generally different from $\mathbf{x}_i$ used in Section 1 for modeling the incidence in the absence of reporting delay, although they may share some variables in common.

To model $F_i(t \mid \mathbf{u}_i)$, first note that it can be written as [3]

$$F_i(t \mid \mathbf{u}_i) = \begin{cases} \prod_{l=t+1}^{n_0} (1 - p_{il}(\mathbf{u}_i)) & \text{if } n - n_0 + 1 \le i \le n \\ 1 & \text{if } 1 \le i \le n - n_0 \end{cases}, \quad 0 \le t \le n_0 - 1, \tag{3}$$

where $p_{il}(\mathbf{u}_i) = \Pr(d = l \mid d \le l, \mathbf{u}_i)$. To model $p_{il}(\mathbf{u}_i)$, let $m_{il}$ be the number of subjects whose attempts occur in the $i$th month and get reported $l$ months later and $c_{il}$ be the number of such cases with a reporting delay equal to $l$ $(1 \le l \le n_0)$. Then,

$$c_{il} \overset{\text{i.d.}}{\sim} Bi(m_{il}, p_{il}(\mathbf{u}_i)), \quad \text{with } \begin{matrix} 1 \le l \le n_0 & \text{if } 1 \le i \le n - n_0, \\ 1 \le l \le n - i & \text{if } n - n_0 + 1 \le i \le n - 1, \end{matrix} \tag{4}$$

where $Bi(n, p)$ denotes a binomial distribution with mean $p$ and size $n$. Although the probability $p_{il}(\mathbf{u}_i)$ may be modeled using any model for binary responses, a particularly popular choice within the context of survival analysis is the complementary log–log link function [3,20]:

$$\log[-\log(1 - p_{il}(\mathbf{u}_i))] = \gamma_{0l} + \mathbf{u}_i^\top \boldsymbol{\gamma}_1, \quad \text{with } \begin{matrix} 1 \le l \le n_0 & \text{if } 1 \le i \le n - n_0, \\ 1 \le l \le n - i & \text{if } n - n_0 + 1 \le i \le n - 1. \end{matrix} \tag{5}$$

Under the assumptions in Equations (4) and (5), $F_i(t \mid \mathbf{u}_i, \boldsymbol{\gamma})$ has a particularly simple form:

$$F_i(t \mid \mathbf{u}_i, \boldsymbol{\gamma}) = F(t \mid \boldsymbol{\gamma})^{\exp(\mathbf{u}_i^\top \boldsymbol{\gamma})}, \quad F(t \mid \boldsymbol{\gamma}) = \prod_{l=t+1}^{n_0} \exp(-\exp(\gamma_{0l})),$$
$$0 \le t \le n_0 - 1, \quad n - n_0 \le i \le n, \quad \boldsymbol{\gamma} = (\gamma_{01}, \ldots, \gamma_{0n_0}, \boldsymbol{\gamma}_1^\top)^\top. \tag{6}$$

Inference about $\boldsymbol{\gamma}$ is readily made using either maximum likelihood and EE. Upon estimating $\boldsymbol{\gamma}$, we can estimate $F_i(t \mid \mathbf{u}_i)$ by $\hat{F}_i(t \mid \mathbf{u}_i) = F_i(t \mid \mathbf{u}_i, \hat{\boldsymbol{\gamma}})$.

Since $c_{il}$ ($m_{il}$) was quite small for most of the facilities within our context, we assumed a homogenous $p_l$ to ensure stable estimates, in which case (6) reduces to

$$F(t \mid \boldsymbol{\gamma}) = \prod_{l=t+1}^{n_0} \exp(-\exp(\gamma_{0l})) = \prod_{l=t+1}^{n_0} (1-p_l), \quad n-n_0 \le i \le n, \;\; 0 \le t \le n_0 - 1. \tag{7}$$

In this special case, the ML estimates of $p_l$ are in closed form [3]:

$$\hat{p}_l = \frac{c_l}{r_l}, \quad 1 \le l \le n_0, \tag{8}$$

where $r_l$ is the number of suicide attempters reported by $l$ months and $c_l$ is the number of those with a delay equal to $l$ ($1 \le l \le n_0$). Note that although the $c_{il}$'s are not stochastically independent, the variance estimates from the ML are still valid [3,6].

The second approach treats reported cases as observed responses from a multinomial model, applies models for such responses and estimates parameters using methods for missing data. Let $a_i(d)$ denote the number of cases reported with a lag of $d$ months and $q_i(t \mid \mathbf{u}_i)$ denote the probability distribution function of the reporting delay, i.e.

$$q_i(t \mid \mathbf{u}_i) = F_i(t \mid \mathbf{u}_i) - F_i(t-1 \mid \mathbf{u}_i), \quad 1 \le i \le n, \;\; 1 \le t \le n_0,$$

where the CDF $F_i(d \mid \mathbf{u}_i)$ is defined the same way as in the first approach. Let

$$\begin{aligned} \mathbf{a}_i &= (a_i(0), \ldots, a_i(n_0-1))^\top, \quad \mathbf{q}_i(\mathbf{u}_i, \boldsymbol{\zeta}) = (q_i(0 \mid \mathbf{u}_i, \boldsymbol{\zeta}), \ldots, q_i(n-1 \mid \mathbf{u}_i, \boldsymbol{\zeta}))^\top, \\ \mathbf{q}_i^o(\mathbf{u}_i, \boldsymbol{\zeta}) &= (q_i^o(0 \mid \mathbf{u}_i, \boldsymbol{\zeta}), \ldots, q_i^o(n-i \mid \mathbf{u}_i, \boldsymbol{\zeta}))^\top, \quad q_i^o(d \mid \mathbf{u}_i, \boldsymbol{\zeta}) = \frac{q_i(d \mid \mathbf{u}_i, \boldsymbol{\zeta})}{\sum_{j=0}^{n-i} q_i(j \mid \mathbf{u}_i, \boldsymbol{\zeta})}. \end{aligned} \tag{9}$$

Then $\mathbf{a}_i$ is observed if $1 \le i \le n - n_0$. For $n - n_0 + 1 \le i \le n$, only the first $(n-i)$ components, $a_i(0), \ldots, a_i(n-i)$, are observed. Let $y_i^o$ denote the observed number of attempters in the $i$th month. Then,

$$y_i^o = \begin{cases} \displaystyle\sum_{d=0}^{n-i} a_i(d) & \text{if } n - n_0 + 1 \le i \le n, \\ \displaystyle\sum_{d=0}^{n_0} a_i(d) & \text{if } 1 \le i \le n - n_0. \end{cases} \tag{10}$$

In other words, $y_i^o = y_i$ for $1 \le i \le n - n_0$, but $y_i^o \le y_i$ for $n - n_0 + 1 \le i \le n$.

For each $1 \le i \le n - n_0$, $\mathbf{a}_i \sim \mathrm{MN}(\mathbf{q}_i(\mathbf{u}_i, \boldsymbol{\zeta}), y_i^o)$, where $\mathrm{MN}(\mathbf{q}, m)$ denotes a multinomial with mean $\mathbf{q}$ and sample size $m$. We can model the cell counts $\mathbf{a}_i$ using a generalized linear model such as the generalized logit or proportional odds model [17]. However, for $n - n_0 + 1 \le i \le n$, $\mathbf{a}_i$ has missing components, but the observed components $\mathbf{a}_i^o = (a_i(0), \ldots, a_i(n-i-1))^\top$ still follows a multinomial $\mathrm{MN}(\mathbf{q}_i^o(\mathbf{u}_i, \boldsymbol{\zeta}), y_i^o)$, albeit with a different mean $\mathbf{q}_i^o(\mathbf{u}_i, \boldsymbol{\zeta})$ defined in Equation (9). Thus, the observed likelihood is given by

$$l = \prod_{i=1}^{n-n_0} \prod_{d=0}^{n_0} (q_i(d \mid \mathbf{u}_i, \boldsymbol{\zeta}))^{a_i(d)} \prod_{i=n-n_0+1}^{n} \prod_{d=0}^{n-i} (q_i^o(d \mid \mathbf{u}_i, \boldsymbol{\zeta}))^{a_i(d)}, \tag{11}$$

where

$$a_i(n-i) = y_i^o - \sum_{d=0}^{n-i-1} a_i(d) \quad \text{if } n - n_0 + 1 \le i \le n,$$

$$a_i(n_0) = y_i^o - \sum_{d=0}^{n_0-1} a_i(d) \quad \text{if } 1 \le i \le n - n_0.$$

Since this likelihood depends on the length of the observation period $n$ and maximal delay $n_0$, it is unique to each application, which in particular precludes applications of standard fitting methods. To avoid maximizing this application-specific likelihood, one may use the EM algorithm to take advantage of available software for fitting (truncated) multinomial responses [2,18].

Both approaches focus primarily on inference for the reporting delay. For inference about caseloads, ad hoc methods may be used such as the Delta method [2,18]. Also, the second approach requires parametric assumptions, which may yield biased inference if data fail to meet the posited distributional assumptions. Most important, inference about caseloads is based on the model for the reporting delay, limiting the complexity of models for estimating and projecting disease incidence. Next, we discuss a distribution-free approach to allow for separate models for the delay distribution and disease incidence to provide flexibility for modeling disease incidence, which is particularly important for our context.

## 4. A new approach for suicide attempters and reporting delays

### 4.1 *Models for reported caseload under reporting delay*

In the presence of reporting delay, only a subset of those who attempt suicide in the $i$th month are observed in the limited time interval $[0, n]$. Thus, the observed cases $y_i^o$ generally underestimate the true number of attempters. We can use the inverse probability weighting (IPW) technique to correct the underreporting by $y_i^o$ before applying the methods in Section 2. The IPW has a long history in the analysis of sample survey data [9]. Within our context, the basic idea is to treat each reported suicide attempt as a representative of a group of all those who attempt suicide in the same month and use the inverse of the selection probability to account for the unreported cases.

To illustrate, suppose that the maximum delay in reporting suicide attempters is 12 months and the reporting delay distribution is uniform over the 12-month period. Then, at the end of 12 months, all attempters that occur in the first month will have been reported. However, only $\frac{11}{12} = 92\%$ of those who attempt in the second month will have been reported by month 12; only $\frac{10}{12} = 83\%$ of the ones who attempt in the third month will have been reported and so on. Thus, each reported suicide attempter in the second month actually represents $(\frac{11}{12})^{-1} = 1.1$ attempters, and each reported case in the third month represents a total of $(\frac{10}{12})^{-1} = 1.2$ attempters, etc.

For each facility, let $\pi_i$ denote the probability that a suicide attempter in the $i$th month is reported by the end of observation period (month $n$). Under the assumption of a maximum delay $n_0$, we have

$$\pi_i(\mathbf{u}_i) = \begin{cases} F_i(n-i \mid \mathbf{u}_i) & \text{if } n - n_0 + 1 \le i \le n, \\ 1 & \text{if } 1 \le i \le n - n_0, \end{cases} \tag{12}$$

where $F_i(t \mid \mathbf{u})$ and $\mathbf{u}_i$ have the same interpretation as in the preceding section. With $\pi_i$, we can statistically account for the 'truncated' attempters in observation period $[0, n]$.

In the presence of reporting delay, we do not observe $y_i$, but rather a smaller number $y_i^o$, representing a subset of individuals whose suicide attempts get reported in the observation period

$[0, n]$. We can estimate $y_i$ by $y_i^o/\pi_i$ and substitute $y_i^o/\pi_i$ in place of $y_i$ in Equation (1) to obtain

$$\mu_i = E\left(\frac{y_i^o}{\pi_i(\mathbf{u}_i)}\middle|\mathbf{x}_i\right) = m_i r_i = m_i \exp(\mathbf{x}_i^\top \boldsymbol{\beta}). \tag{13}$$

Inference about $\boldsymbol{\beta}$ is again based on the EE in Equation (2) by substituting $y_i^o/\pi_i(\mathbf{u}_i)$ for $y_i$. If the $\pi_i$'s are known, the above revised model is readily fit using the methods discussed in Section 2. By Equation (12), we can readily estimate $\pi_i$ using the methods discussed in Section 3, and then proceed with fitting the model in Equation (13). In most applications, the $\pi_i$'s are unknown and must be estimated. If $\pi_i$ is replaced by an estimate $\hat{\pi}_i$, the responses $y_i^o/\hat{\pi}_i$ are no longer independent. Although the EE in Equation (2) still provides consistent estimates, variance estimates generally underestimate the variability of the EE estimate [12,15,19]. Although the dependence issue may be addressed by accounting for the sampling variability of $\hat{\pi}_i$ [12,15,19], the resulting approach can be quite complex. A more convenient alternative is to use functional response models (FRMs) to facilitate inference.

### 4.2 *An FRM approach*

Like all regression models, the distribution-free Poisson log-linear regression in Equation (1) model the linear $y_i$, or the conditional mean, or first moment, of $y_i$ given $\mathbf{x}_i$. Many relationships of interest arising in practice also require additional moments or even between-subject interactions. For example, to distinguish the Poisson and NB, we may model the second moment $y_i^2$ in addition to the linear $y_i$ [12,23]. To model non-parametric statistics such as the Mann–Whitney–Wilcoxon rank sum statistic, we need to model pairs of subject responses [4,22].

The FRM addresses the aforementioned limitations of traditional regression:

$$E[f(y_{i_1}, \ldots, y_{i_q}; \boldsymbol{\gamma}) \mid \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}] = h(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}; \boldsymbol{\beta}), \quad (i_1, \ldots, i_q) \in C_q^n, \tag{14}$$

where $f(\cdot)$ is some functional, $h(\cdot)$ some smooth functional (e.g. continuous second-order derivatives), $C_q^n$ denotes the set of $\binom{n}{q}$ combinations of $q$ distinct elements $(i_1, \ldots, i_q)$ from the integer set $\{1, \ldots, n\}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ a vector of parameters. By generalizing the single-subject based linear response in standard regression to complex functions of responses from multiple subjects, Equation (14) is uniquely positioned to model higher order moments as well as between-subject interactions [4,7,11,16,21–24].

To apply the FRM in integrating the reporting delay model in Equations (4) and (5) with the incidence model in Equation (2), first consider a distribution-free alternative for estimating $F_i(t \mid \mathbf{u}_i, \boldsymbol{\gamma})$. Following the notation in Section 3, we have

$$E[a_i(d) \mid \mathbf{u}_i] = q_i(d \mid \mathbf{u}_i, \boldsymbol{\gamma}), \quad 0 \le d \le n_0 - 1. \tag{15}$$

In the presence of reporting delay, $\mathbf{a}_i$ will have missing data in its last $(n - i)$ components for $n - n_0 + 1 \le i \le n$. Define a set of indicators for missing data for each $\mathbf{a}_i$ as follows:

$$r_{id} = \begin{cases} 1 & \text{if } 1 \le i \le n - n_0, \\ 1 & \text{if } n - n_0 + 1 \le i \le n \quad \text{and } 0 \le d \le n - i, \\ 0 & \text{if otherwise.} \end{cases}$$

$$\zeta_d = E(r_{id}), \quad \mathbf{r}_i = (r_{i1}, \ldots, r_{in_0})^\top, \quad \boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_{n_0})^\top.$$

Thus, $r_{id} = 1$ if $a_i(d)$ is observed and 0 otherwise.

Since missing data in $\mathbf{a}_i$ follow the missing at random condition, we can estimate $\boldsymbol{\gamma}$ in Equation (15) by the following weighted estimating equations (WEE) [12,15,19,23].

$$\mathbf{w}(\boldsymbol{\gamma}, \boldsymbol{\zeta}) = \sum_{i=1}^{n} D_i V_i^{-1} \Psi_i S_i = \mathbf{0}, \tag{16}$$

where

$$S_i = \mathbf{a}_i - \mathbf{q}_i, \quad D_i = \frac{\partial}{\partial \boldsymbol{\gamma}} \mathbf{q}_i, \quad V_i = \mathrm{Var}(\mathbf{a}_i \mid \mathbf{u}_i), \quad \Psi_i = \mathrm{diag}_d(\Psi_{id}), \quad \Psi_{id} = \frac{r_{id}}{\zeta_d}.$$

We can readily estimate $\zeta_d$ by $\hat{\zeta}_d = (1/n)\sum_{i=1}^{n} r_{id}$. By substituting $\hat{\zeta}_d$ in place of $\zeta_d$ in Equation (16) and solving the resulting WEE for $\boldsymbol{\gamma}$, we obtain the WEE estimate $\hat{\boldsymbol{\gamma}}$. With such an estimate, we can estimate $\pi_i(\boldsymbol{\gamma})$ by

$$\pi_i(\hat{\boldsymbol{\gamma}}(\hat{\boldsymbol{\zeta}})) = \begin{cases} \sum_{d=0}^{n-i} q_i(d \mid \mathbf{u}_i, \hat{\boldsymbol{\gamma}}) & \text{if } n - n_0 + 1 \le i \le n, \\ 1 & \text{if } 1 \le i \le n - n_0, \end{cases}$$

where $\hat{\boldsymbol{\gamma}}(\hat{\boldsymbol{\zeta}})$ denotes the dependence of $\hat{\boldsymbol{\gamma}}$ on $\hat{\boldsymbol{\zeta}}$. By substituting $\pi_i(\hat{\boldsymbol{\gamma}}(\hat{\boldsymbol{\zeta}}))$ for $\pi_i(\boldsymbol{\gamma})$ in Equation (13), we can estimate $\boldsymbol{\beta}$ using the EE discussed earlier in Section 3.

A major drawback of the above procedure is the need to adjust the asymptotic variance [23] of the estimate $\hat{\boldsymbol{\beta}}$ from the EE for the sampling variability in the estimated $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\zeta}}$ from Equation (16) [12,15,19,23]. Alternatively, we can integrate all these modeling steps into a single FRM as follows:

$$\begin{aligned}
&\mathbf{f}_i = \mathbf{f}(y_i^o, \mathbf{a}_i, \mathbf{r}_i, \pi_i(\boldsymbol{\gamma}), \boldsymbol{\eta}) = (f_{i1}, \mathbf{f}_{i2}^\top, \mathbf{f}_{i3}^\top)^\top, \quad \mathbf{h}_i = (h_{i1}, \mathbf{h}_{i2}^\top, \mathbf{h}_{i3}^\top)^\top, \quad k = 2, 3, \\
&\mathbf{f}_{ik} = (f_{ik1}, \ldots, f_{ikn_0})^\top, \quad \mathbf{h}_{ik} = (h_{ik1}, \ldots, h_{ikn_0})^\top, \quad k = 2, 3, \\
&f_{i1} = \frac{y_i^o}{\pi_i(\boldsymbol{\gamma})}, \quad h_{i1} = m_i \exp(\mathbf{x}_i^\top \boldsymbol{\beta}), \\
&f_{i2l} = \frac{a_i(l-1)}{y_i^o}, \quad h_{i2l} = q_i(l-1 \mid \mathbf{u}_i, \boldsymbol{\gamma}), \quad f_{i3l} = r_{i(l-1)}, \quad h_{i3l} = \zeta_{(l-1)}, \quad 1 \le l \le n_0, \\
&\pi_i(\boldsymbol{\gamma}) = \begin{cases} \sum_{l=0}^{n-i} q_i(l \mid \mathbf{u}_i, \boldsymbol{\gamma}) & \text{if } n - n_0 + 1 \le i \le n, \\ 1 & \text{if } 1 \le i \le n - n_0, \end{cases}
\end{aligned} \tag{17}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\zeta}^\top)^\top$ denotes the collection of parameters. The above is not a generalized linear or a nonlinear model, since the response $\mathbf{f}_i$ is quite a complex function of outcomes $y_i^o$, $c_{il}$ and $n_l$ and the unknown quantity $\pi_i(\boldsymbol{\gamma})$. It is an FRM, with one component consisting of $E(f_{i1} \mid \mathbf{x}_i, \mathbf{u}_i; \boldsymbol{\theta}) = h_{i1}$ for modeling the suicide attempter, a second component $E(\mathbf{f}_{i2} \mid \mathbf{x}_i, \mathbf{u}_i; \boldsymbol{\theta}) = \mathbf{h}_{i2}$ for modeling the reporting delay and a third one $E(\mathbf{f}_{i3} \mid \mathbf{x}_i, \mathbf{u}_i; \boldsymbol{\theta}) = \mathbf{h}_{i3}$ for modeling the missing data indicator. The model in Equation (17) is distribution free, as no parametric model is assumed for any component of the FRM. We may model $q_i(l \mid \mathbf{u}_i, \boldsymbol{\gamma})$ using either approach in Section 3.

Inference about $\boldsymbol{\theta}$ is based on the following weighted generalized estimating equations (WGEE) for FRM:

$$\mathbf{w}(\boldsymbol{\theta}) = \sum_{i=1}^{n} D_i V_i^{-1} \Phi_i S_i = \mathbf{0}, \quad S_i = \mathbf{f}_i - \mathbf{h}_i, \quad D_i = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{h}_i, \quad V_i = A_i^{1/2} R(\boldsymbol{\alpha}) A_i^{1/2},$$

$$A_i = \begin{pmatrix} \mathrm{Var}(f_{i1} \mid \mathbf{x}_i, \mathbf{u}_i) & 0 & 0 \\ & \mathrm{Var}(\mathbf{f}_{i2} \mid \mathbf{x}_i, \mathbf{u}_i) & 0 \\ & & \mathrm{Var}(\mathbf{f}_{i3} \mid \mathbf{x}_i, \mathbf{u}_i) \end{pmatrix},$$

$$\Phi_i = \begin{pmatrix} 1 & 0 & 0 \\ & \Psi_i & 0 \\ & & \mathbf{I}_{n_0} \end{pmatrix}, \quad R_i = \begin{pmatrix} 1 & R_{12}(\boldsymbol{\alpha}) & R_{13}(\boldsymbol{\alpha}) \\ & \mathbf{I}_{n_0} & R_{23}(\boldsymbol{\alpha}) \\ & & \mathbf{I}_{n_0} \end{pmatrix}, \tag{18}$$

where $R(\boldsymbol{\alpha})$ denotes a choice of working correlation matrix parameterized by $\boldsymbol{\alpha}$. In the above, $D_i$ and $A_i$ are readily computed if we assume some parametric assumptions about the distribution of $\mathbf{f}_i$. For example, if we assume a Poisson with the mean in Equation (1), $A_i$ is given by

$$\mathrm{Var}(f_{i1} \mid \mathbf{x}_i, \mathbf{u}_i) = h_{i1},$$

$$\mathrm{Var}(\mathbf{f}_{i2} \mid \mathbf{x}_i, \mathbf{u}_i) = \begin{pmatrix} h_{i21}(1-h_{i21}) & -h_{i21}h_{i22} & \cdots & -h_{i21}h_{i2n_0} \\ \vdots & h_{i22}(1-h_{i22}) & \vdots & -h_{i22}h_{i2n_0} \\ & & \ddots & \vdots \\ & & & h_{i2n_0}(1-h_{i2n_0}) \end{pmatrix},$$

$$\mathrm{Var}(\mathbf{f}_{i3} \mid \mathbf{x}_i, \mathbf{u}_i) = \begin{pmatrix} h_{i31}(1-h_{i31}) & 0 & \cdots & 0 \\ \vdots & h_{i32}(1-h_{i32}) & \vdots & 0 \\ & & \ddots & \vdots \\ & & & h_{i3n_0}left(1-h_{i3n_0}) \end{pmatrix}. \tag{19}$$

However, the WGEE estimate $\hat{\boldsymbol{\theta}}$ obtained by solving the equations in (18) remains consistent and asymptotically normal, regardless of the parametric models assumed so long as the specification in Equation (17) is correct (see below). For example, the choice of $A_i$ in Equation (18) still yields valid inference if the number of suicide attempters follows an NB.

As in the standard WGEE, we can choose $R_{12}(\boldsymbol{\alpha})$, $R_{13}(\boldsymbol{\alpha})$ and $R_{23}(\boldsymbol{\alpha})$ in a variety of ways to model correlations among the components of $\mathbf{f}_i$. The choice of $R(\boldsymbol{\alpha})$ and properties associated with GEE estimates have been extensively discussed in the literature. In particular, the WGEE estimate may not be consistent under working correlation structures other than the working independence model. Thus, the working independence model may be used in general to ensure valid inference, in which case (18) corresponds to $R_{kl}(\boldsymbol{\alpha}) = \mathbf{0}$ $(1 \le k < l \le 3)$.

Let $\hat{\boldsymbol{\theta}}$ denote the WGEE estimate of $\boldsymbol{\theta}$ by solving the EEs in (18). Under mild regularity conditions, $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal (see the appendix):

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to {}_dN(\mathbf{0}, \Sigma_\theta = B^{-1}\Sigma_U B^{-\top}),$$
$$B = E\left[D_i V_i^{-1} \Phi_i \frac{\partial}{\partial \boldsymbol{\theta}} S_i\right], \quad \Sigma_U = E(D_i V_i^{-1} \Phi_i S_i S_i^\top \Phi_i V_i^{-1} D_i^\top), \tag{20}$$

A consistent estimate of $\Sigma_\theta$ is given by

$$\hat{\boldsymbol{\Sigma}}_\theta = \hat{B}^{-1}\hat{\Sigma}_U\hat{B}^{-1}, \quad \hat{B} = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_i \hat{V}_i^{-1} \hat{\Phi}_i \frac{\partial}{\partial \boldsymbol{\theta}} \hat{S}_i, \quad \hat{\Sigma}_U = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_i \hat{V}_i^{-1} \hat{\Phi}_i \hat{S}_i \hat{S}_i^\top \hat{\Phi}_i \hat{V}_i^{-1} \hat{D}_i^\top, \tag{21}$$

where $\hat{A}$ denotes $A$ with $\boldsymbol{\theta}$ substituted by the GEE estimate $\hat{\boldsymbol{\theta}}$.

## 5. Interval estimates for mean response and individual responses

One of the primary objectives of the VHA surveillance system is to identify facilities that show a significant increase of number of suicide attempters so that such facilities can be notified in a timely fashion. Interval estimates are required to detect such 'change points'.

### 5.1 *Confidence intervals for mean response*

Once we have an estimate $\hat{\boldsymbol{\theta}}$ for the FRM in Equation (17), we can obtain an estimate of the mean response in any given month $i$ by

$$\hat{\mu}_i = \exp(\log m_i + \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) = m_i \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}), \quad 1 \le i \le n. \tag{22}$$

To find $100(1-\alpha)\%$ confidence intervals for the mean $\mu_i$, we first find $100(1-\alpha)\%$ confidence intervals for the linear predictor $\hat{\eta}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$. Since

$$\mathrm{Var}(\hat{\eta}_i) = \mathrm{Var}(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) = \mathbf{x}_i^\top \Sigma_\beta \mathbf{x}_i, \tag{23}$$

we can estimate $\mathrm{Var}(\hat{\eta}_i)$ by $\widehat{\mathrm{Var}}(\hat{\eta}_i) = \mathbf{x}_i^\top \hat{\Sigma}_\beta \mathbf{x}_i$, where $\hat{\Sigma}_\beta$ denotes a consistent estimate of $\Sigma_\beta$ such as the robust sandwich variance estimate in Equation (21). A $100(1-\alpha)\%$ confidence interval for $\eta_i$ is given by

$$(L_{\eta i}, R_{\eta i}) = (\hat{\eta}_i - q_{1-(1/2)\alpha}\sqrt{\widehat{\mathrm{Var}}(\hat{\eta}_i)}, \hat{\eta}_i + q_{1-(1/2)\alpha}\sqrt{\widehat{\mathrm{Var}}(\hat{\eta}_i)}), \tag{24}$$

where $q_\alpha$ denotes the $\alpha$th percentile of the standard normal. By changing the above intervals into the scale of $\mu_i$, we obtain a $100(1-\alpha)\%$ confidence interval for $\mu_i$:

$$(L_{\mu i}, R_{\mu i}) = (m_i \exp(L_{\eta i}), m_i \exp(R_{\eta i})). \tag{25}$$

The confidence interval captures the variability of $\hat{\mu}_i$ in estimating the (population) mean $\mu_i$ of number of suicide attempters in the $i$th month. The interval in Equation (25) cannot be used to predict the variability of new number of suicide attempters $y_i$ $(i > n)$ beyond the observation period $[1, n]$. To ascertain whether $y_i$ in a future month $i$ $(i > n)$ signals a significant departure from the posited model, we need prediction intervals that also account for the variability of the random $y_i$.

### 5.2 *Prediction intervals for a new individual observation*

Given a statistical model such as the Poisson or a distribution-free version in Equation (1), we can estimate a new observation $y_i$ by $\hat{y}_i = m_i \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})$, the same point estimate as for the mean response $\mu_i$ in Equation (22). However, the variance estimate in Equation (23) generally underestimates the variability of $\hat{y}_i$, since it only accounts for the sampling variability of $\hat{\boldsymbol{\beta}}$. When predicting $y_i$ in a future month $i$ $(i > n)$, we must also consider the variability of $y_i$ about the mean $\mu_i$. For this purpose, we must assume a distribution that governs the random behavior of $y_i$ such as the Poisson or NB. The distribution-free version in Equation (1) does not provide sufficient information for constructing prediction intervals for $y_i$.

Suppose that $y_i$ follows a Poisson model with the mean given in Equation (1). If $\mu_i$ is known, a $100(1-\alpha)\%$ prediction interval for a future $y_i$ is: $(L_{yi}(\mu_i), R_{yi}(\mu_i))$, where $L_i$ $(R_i)$ is the $\alpha/2$th $((1-\alpha/2)$th) percentile of the Poisson$(\mu_i)$. In practice, $\mu_i$ is unknown and estimated by $\hat{\mu}_i$. This sampling variability must be accounted for to provide more accurate prediction intervals.

Following the discussion in Section 5.1, $\hat{\mu}_i = m_i \exp(\hat{\eta}_i)$, where $\hat{\eta}_i$ follows an approximate normal, $N(\hat{\eta}_i, \mathrm{Var}(\hat{\eta}_i))$. By treating $\eta_i$ in as random, $\eta_i$ follows approximately $N(\hat{\eta}_i, \mathrm{Var}(\hat{\eta}_i))$.

Thus, we can integrate out $\hat{\mu}_i$ in $F_i(y \mid \hat{\mu}_i)$, the CDF of Poisson$(\hat{\mu}_i)$, conditional on $\hat{\eta}_i$ and $\text{Var}(\hat{\eta}_i)$ to obtain the following CDF:

$$F_i(y \mid \hat{\eta}_i, \text{Var}(\hat{\eta}_i)) = \sum_{t=0}^{y} \int_{-\infty}^{\infty} \frac{\exp(-m_i \exp(\eta))(m_i \exp(\eta))^t}{t!} \phi(\eta \mid \hat{\eta}_i, \text{Var}(\hat{\eta}_i))\, \mathrm{d}\eta. \tag{26}$$

The above is not a Poisson, but incorporates the variability of $\hat{\mu}_i$ based on the estimate $\hat{\mu}_i$ and associated sampling distribution. Although the CDF above is not in closed form, it is readily evaluated numerically.

Let $L_{yi}$ and $R_{yi}$ be the $(1/2)\alpha$ and $(1-(1/2)\alpha)$ percentiles of the distribution in Equation (26):

$$\begin{aligned} L_{yi} &= \max\{y : F_i(y \mid \hat{\eta}_i, \text{Var}(\hat{\eta}_i)) \le \tfrac{1}{2}\alpha\}, \\ R_{yi} &= \min\{y : F_i(y \mid \hat{\eta}_i, \text{Var}(\hat{\eta}_i)) \ge 1 - \tfrac{1}{2}\alpha\}. \end{aligned} \tag{27}$$

Then, a $100(1-\alpha)\%$ prediction interval for $y_i$ is given by $(L_{yi}, R_{yi})$. Note that because of the discrete nature of $y_i$, the interval $(L_{yi}, R_{yi})$ may not have exact $100(1-\alpha)\%$ coverage, but the definition in Equation (27) ensures at least $100(1-\alpha)\%$ coverage.

The above is readily modified to provide prediction intervals for NB and other models for count responses. We again emphasize that although a specification of the mean response such as Equation (1) suffices to estimate $\mu_i$ and associated confidence intervals, a fully specified distribution model such as the Poisson is needed to predicate the variability of a new $y_i$.

## 6. Application

We illustrate the methodology with both simulated and real study data. We start with a simulation study.

### 6.1 *A simulation study*

We conducted a simulation study to examine the performance of the proposed FRM model in Equation (17) for incidence of suicide attempt in the presence of reporting delay. Since asymptotic behaviors of estimates of Poisson models are determined by the mean of the Poisson distribution, with larger means indicating larger sample sizes [20], we considered three scenarios with the mean averaged to 5, 20 and 50 in our simulations.

For notational brevity, we considered a period of $n = 14$ months, assumed a maximum reporting delay of $n_0 = 12$ months, and set $\mathbf{x}_i = (1, i)^\top$ with no offset term ($m_i = 1$). Thus, in the absence of reporting delay, $y_i$ was simulated according to:

$$y_i \mid \mathbf{x}_i \overset{\text{i.d.}}{\sim} \text{Poisson}(\mu_i), \quad \mu_i = E(y_i \mid \mathbf{x}_i), \quad \log(\mu_i) = \beta_0 + i\beta_1, \quad 1 \le i \le n. \tag{28}$$

We set $\beta_1 = 0.2$. To ensure that $\mu_i$ average to 5, 20 and 50, we set $\bar{\mu}_\cdot = (1/n)\sum_{i=1}^{n} \mu_i$ to each of the desired values and then solved for $\beta_0$, yielding $\beta_0 = -0.18$, 1.19 and 2.11.

To simulate observed counts $a_i(d)$ in the presence of reporting delay, we assumed a multinomial MN$(\mathbf{q}(\gamma), y_i)$, where $\mathbf{q}(\gamma)$ was independent of $i$ given by

$$q(d \mid \gamma) = \begin{cases} \dfrac{\exp((n_0 - d)\gamma)}{1 + \sum_{j=0}^{n_0-1} \exp((n_0 - j)\gamma)} & \text{if } 0 \le d \le n_0 - 1, \\ \dfrac{1}{1 + \sum_{j=0}^{n_0-1} \exp((n_0 - j)\gamma)} & \text{if } d = n_0, \end{cases} \tag{29}$$

Table 1. Comparison of estimates and standard errors from FRM and EE for modeling incidence and reporting delay for the Simulation Study.

| | Estimates of Parameters and Standard Errors (s.e.) for Simulation Study | | |
|---|---|---|---|
| Estimate | $\boldsymbol{\beta} = (-0.18, 0.2)$ | $\boldsymbol{\beta} = (1.19, 0.2)$ | $\boldsymbol{\beta} = (2.11, 0.2)$ |
| EE | | | |
| $\beta_0$ | −0.2 | 1.18 | 2.11 |
| s.e. | 0.10 | 0.14 | 0.11 |
| Empirical s.e. | 0.12 | 0.17 | 0.13 |
| $\beta_1$ | 0.19 | 0.2 | 0.2 |
| s.e. | 0.046 | 0.021 | 0.012 |
| Empirical s.e. | 0.051 | 0.024 | 0.015 |
| Type 1 error for $\beta_1$ | 0.08 | 0.07 | 0.08 |
| $\gamma$ | 0.49 | 0.5 | 0.5 |
| s.e. | 0.16 | 0.14 | 0.13 |
| FRM | | | |
| $\beta_0$ | −0.19 | 1.19 | 2.11 |
| s.e. | 0.13 | 0.16 | 0.14 |
| Empirical s.e. | 0.14 | 0.17 | 0.15 |
| $\beta_1$ | 0.21 | 0.2 | 0.2 |
| s.e. | 0.053 | 0.025 | 0.016 |
| Empirical s.e. | 0.052 | 0.026 | 0.017 |
| Type 1 error for $\beta_1$ | 0.06 | 0.04 | 0.05 |
| $\gamma$ | 0.48 | 0.5 | 0.5 |
| s.e. | 0.17 | 0.15 | 0.16 |



We set $\gamma = 0.5$ so that the proportion of cases reported over the 12 months after the occurrence of the incident changed from 39% for cases reported within the same month to 0.1% for those reported in the last month.

We fit the FRM in Equation (17) to the simulated $y_i$ and $a_i(d)$ from (28) and (29). As discussed in Section 4, we used the working independence model for the correlations among $f_i$, $\mathbf{f}_{i2}$ and $\mathbf{f}_{i3}$. For comparison purposes, we also fit the data using an ad hoc approach by first estimating $\mathbf{q}_i(\gamma)$ based on $a_i(d)$ and then fitting the observed $y_i^o$ using the model in Equation (13), with $y_i^o$ and $\pi_i$ given by Equations (10) and (12), respectively. To be consistent with the FRM, we estimated $\gamma$ for the $\mathbf{q}(\gamma)$ in Equation (29) using the EEs, rather than maximum likelihood as in Equation (11) to provide robust inference about $\gamma$. As discussed in Section 4, the FRM addresses non-independence among $y_i^o/\hat{\pi}_i$ with an estimated $\pi_i$, thereby providing valid inference for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \gamma, \boldsymbol{\zeta}^\top)^\top$.

Shown in Table 1 are the estimates of $\boldsymbol{\beta}$ and $\gamma$ and associated standard errors and $p$-values averaged over 1000 Monte Carlo replications. Note that estimates of $\boldsymbol{\zeta}$ are not shown in the table, since our primary interest centers on $\boldsymbol{\beta}$ and $\gamma$. Both EE and FRM provided accurate estimates even for the smallest $\bar{\mu}_\cdot = 5$. However, the standard errors were different, with FRM yielding larger standard errors than the EE for the estimates of $\boldsymbol{\beta}$. This is expected since the standard errors by EE did not take into account the variability of the estimated $\gamma$, thereby underestimating the variability of the EE estimate of $\boldsymbol{\beta}$.

The standard errors of the estimates of $\boldsymbol{\theta}$ also generally decreased as $\beta_0$ increased from $-0.18$ to 1.19–2.11. This is also expected since increased $\beta_0$ led to larger means, which in turn yielded more efficient estimates. Also, as expected, the standard errors of the FRM estimates of $\boldsymbol{\beta}$ were slightly larger than their EE counterparts, again reflecting the added sampling variability in the estimate of $\gamma$.

The simulation study shows that (1) without correcting for the sampling variability in the estimate of $\gamma$, the EE underestimates the variability of estimates of $\boldsymbol{\beta}$ and (2) the approach works well even for relatively small sample sizes.

### 6.2 *A case study*

We applied the approach to the VHA surveillance database discussed in Section 1. The data used for the Case Study was first-time, non-fatal suicide attempters who also had recent VHA service utilization prior to the date of the suicide attempt, with a total of 14,182 across the 139 VHA facilities. We first modeled number of such suicide attempters over a period of time and then used the modeled incidence to project further incidence beyond the observation period. The first analysis was based on the attempts that occurred and were reported between August 2010 and November 2011, while the second was based on those that occurred in the next 3 months, December 2011, January and February 2012, but reported by March 2013. The longer reporting time in the second analysis (14 months) was to ensure that the observed incidence in the projected 3 months was not influenced by reporting delays, since the first analysis showed that a very small percent of cases was reported beyond this maximal delay.

For the first analysis, we set $i = 1$ for August 2010 and thus $i = 14$ designates the last month of the period, November 2011. Since the facility size was the only covariate available, we just included month $i$ in the model for $y_i^o$ in Equation (13), i.e. $\mathbf{x}_i = (1, i)^\top$. For $a_i(d)$, we set $u_i = 1$ for the model in Equation (15), since as noted earlier, most of the facilities had relatively small $a_i(d)$ and a homogeneous reporting distribution, i.e. a constant $q(d \mid \boldsymbol{\gamma})$, would provide a more stable estimate of the reporting delay distribution. For setting the maximum length $n_0$ of reporting delay for each facility, we inspected the distribution of observed $a_i(d)$ for beginning months of the period and found that except for a few facilities, $a_i(14) = 0$, for such months. For such facilities, $n_0$ ($< 14$) was set equal to the largest $d$ for which $a_i(d) = 0$. Even for the few facilities for which $a_i(14) \neq 0$, $a_i(d)$ accounted for quite a small proportion of reported cases over the period for month $i = 1$ or 2. Thus, we set $n_0 = 14$ for these remaining facilities. Although it is possible that the maximal delay for such facilities may exceed 14 months, the percent of cases reported beyond this maximal delay should be quite small with no major effect on the estimated and projected incidence.

Although more complex temporal trends could be entertained, such as piecewise linear or higher order polynomials, an examination of the observed and model-estimated (corrected for reporting delays) incidence seemed to indicate that the linear function (in the log rate $\log(r_i)$) modeled the trend reasonably well. Furthermore, for $i$ near the end of the study period $n$, $y_i^o/\pi_i$ becomes more dependent on $\pi_i$ and more complex patterns such as higher order polynomials may over-extrapolate the observed data. By applying the model above to each of the 139 facilities, 21 (15%) showed a significant increase, 32 (23%) exhibited a significant decrease, while the remaining 86 (62%) had no change over the 14-month period monitored. Also, 3 (14%) VISNs showed a significant difference in incidence rate across the facilities.

The fitted model also provides a basis for projecting incidence of first-time suicide attempters for each facility beyond the end of the study period. As noted in Section 5.2, distribution assumptions are needed to model the variability of incidence beyond the end of the study. For the second analysis, we considered both Poisson and NB models, with the latter to accommodate potential overdispersion. To determine which model to use for a facility, we refit the data from each facility and compared the two using the goodness-of-fit statistics associated with the Poisson and NB [20]. To increase robustness, however, the estimate of $\boldsymbol{\beta}$ for constructing the prediction interval was still based on the proposed FRM approach.

Shown in Figure 2 are the projected incidence and 90% prediction intervals over 3 months beyond the end of the study, that is, Months 15, 16 and 17, along with the estimated mean

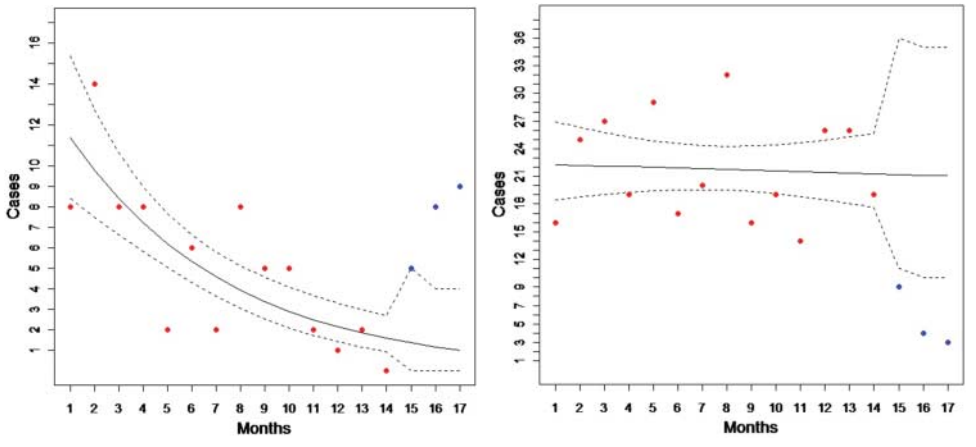


Figure 2. Observed (dots) and projected incidence (solid line), along with 90% confidence (Months 1–14) and prediction (Months 15–17) intervals, for two VHA Facilities based on NB model; only the first 14 dots (Months 1–14) were used for model fitting.



incidence and associated 90% confidence intervals (Months 1–14), for 2 of the 139 facilities based on the NB. The model-based mean incidence seemed to fit the observed incidence well. The projected incidence was the same as the mean incidence, but extrapolated to the last 3 months. The left plot shows that the observed incidence (the last three dots) in the projected Month 16 and 17 exceeded the upper bound of the prediction interval, indicating that the numbers of first-time suicide attempters observed were significantly higher than expected based on the trend modeled over the past 14 months for this facility. Facilities with a significant shift in incidence like this one may need more careful monitoring in the future for potential outbreak or intervention. Note that we selected three months for the projection time frame because the VHA surveillance system was updated every 3 months. Note also that we used the NB for both facilities because it provided improved fit over its Poisson counterpart for these 2 particular facilities.

As expected, the prediction intervals are much wider than the confidence intervals, since the latter describe the variability of incidence in each month, rather than just the sampling variability of the estimate of mean incidence. If the model for the mean in Equation (1) and the models selected for predicting incidence (Poisson or NB) were both correct, each monthly incidence observed in the projection period would have about 90% chance of falling inside the prediction interval. Given only three incidence observations in the projected period from each facility, it is not possible to evaluate the quality of prediction intervals for each facility. We assessed the accuracy of prediction intervals by pooling results across all facilities. Out of the 417 ($139 \times 3$) incidence observations in the projected 3-month period, about 12% fell outside the interval. Thus, the prediction intervals seemed to capture the variability of incidence in the projected time period well.

## 7. Discussion

Unlike available methods, the approach developed allows one to model disease incidence separately from the reporting delay to take advantage of the often larger amount of data for the former model. This approach enables one to entertain more complex models for disease incidence to improve precision and reliability of estimates and projections of disease incidence. Furthermore, by framing all such considerations within the framework of FRM, we are able to integrate the separate modeling components under a single unified model to provide joint inference about all

parameters of interest. In addition, we also addressed projections of future disease incidence, a problem understudied in the current literature, despite its important research and clinical implications. Thus, the novelties of our approach include (1) modeling disease incidence and reporting delay separately to allow for accommodation of a subsample of data for modeling the reporting delay; (2) use of the FRMs to frame the two modeling components (disease incidence and reporting delay) within the context of a single model; (3) use of a set of weighted generalized EEs adapted to the FRM to provide consistent parameter estimates and valid inference and (4) new methods for projecting disease incidence.

We examined the performance of the proposed approach through simulated data. The simulation results indicate good performance, even for relatively small sample sizes (small mean). Since incidence of first-time suicide attempters was small for some of the facilities, this robustness feature ensures reliable estimates of incidence for such facilities.

We also illustrated the approach with a real VHA surveillance database for US Veteran first-time suicide attempters. Because of high levels of security for data use within the VHA, we only had information about facility sizes. Despite the limited information, we still obtained useful information about the variability of incidence of first-time suicide attempters over time across the facilities. The estimated reporting delay distribution also helps identify the facilities with long reporting delays to improve incidence reporting in these facilities.

In the real study application, we also assessed the performance of fitted models for incidence of first-time suicide attempters beyond the end of the study period. The fitted model seems to work well, with the observed type I error rates (percent of incidence exceeding the prediction bands) well approximating the nominal type I error level based on the fitted model. Unlike modeling incidence, distribution assumptions such as the Poisson or NB are needed to model the variability of projected incidence. The prediction intervals provide short-term incidence projections useful for warning signs of potential outbreak of such incidence in a facility.

We developed an SAS macro to implement the proposed FRM in SAS. The macro not only fits the proposed FRM to data, but also selects appropriate models for projections and plots point and interval estimates as discussed in Section 6.2. The software is available from the authors upon request.



## Acknowledgments

We like to thank two anonymous reviewers for their constructive comments that led to a significantly improved manuscript. The work was supported in part by research funding from the Center of Excellence for Suicide Prevention, Canandaigua VA Medical Center, Canandaigua, NY 14424.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## References

[1] P. Bacchetti, *Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns*, J. Am. Statist. Assoc. 85 (1990), pp. 530–539.

[2] R. Brookmeyer and M.H. Gail, *A method for obtaining short-term projections and lower bounds on the size of AIDS epidemic*, J. Am. Statist. Assoc. 83 (1988), pp. 301–308.

[3] R. Brookmeyer and J. Liao, *The analysis of delays in disease reporting: Methods and results for the acquired immunodeficiency syndrome*, Am. J. Epidemiol. 132(2) (1990), pp. 355–365.

[4] R. Chen, T. Chen, N. Lu, H. Zhang, P. Wu, C. Feng, and X.M. Tu, *Extending the Mann–Whitney–Wilcoxon rank sum test to longitudinal data analysis with covariates*, Appl. Stat. 41(12) (2014), pp. 2658–2675.

[5] V. DeGruttola, X.M. Tu, and M. Pagano, *Pediatric AIDS in New York city: Estimating the distributions of infection, latency, and reporting delay and projecting future incidence*, J. Am. Statist. Assoc. 87 (1992), pp. 633–640.
[6] B. Efron, *Logistic regression, survival analysis, and the Kaplan–Meier curve*, J. Am. Statist. Assoc. 83 (1988), pp. 414–425.
[7] D. Gunzler, N. Lu, W. Tang, P. Wu, and X.M. Tu, *A class of distribution-free models for longitudinal mediation analysis*, Psychometrika 79(4) (2014), pp. 543–568.
[8] J.E. Harris, *Reporting delays and the incidence of AIDS*, J. Am. Statist. Assoc. 85 (1990), pp. 915–924.
[9] D.G. Horvitz and D.J. Thompson, *A generalization of sampling without replacement from a finite universe*, J. Am. Statist. Assoc. 47 (1952), pp. 663–685.
[10] J.D. Kalbfleisch and J.F. Lawless, *Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS*, J. Am. Statist. Assoc. 84 (1989), pp. 360–372.
[11] J. Kowalski and J. Powell, *Nonparametric inference for stochastic linear hypotheses: Application to high-dimensional data*, Biometrika 91 (2004), pp. 393–408.
[12] J. Kowalski and X.M. Tu, *Modern Applied U Statistics*, Wiley, New York, 2007.
[13] S.W. Lagakos, L.M. Barraj, and V. De Gruttola, *Nonparametric analysis of truncated survival data, with application to AIDS*, Biometrika 75 (1988), pp. 515–523.
[14] K.-Y. Liang and S.L. Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika 73(1) (1986), pp. 13–22.
[15] N. Lu, W. Tang, H. He, Q. Yu, P. Crits-Christoph, H. Zhang, and X.M. Tu, *On the impact of parametric assumptions and robust alternatives for longitudinal data analysis*, Biomet. J. 51 (2009), pp. 627–643.
[16] Y. Ma, W. Tang, C. Feng, and X.M. Tu, *Inference for Kappas for longitudinal study data: Applications to sexual health research*, Biometrics 64 (2008), pp. 781–789.
[17] P. Mccullagh and J.A. Nelder, *Generalized Linear Models*, 2nd ed., Chapman & Hall/CRC Press, Boca Raton, FL, 1989.
[18] M. Pagano, X.M. Tu, V. DeGruttola, and S. MaWhinney, *Analysis of censored and truncated data: Estimating the reporting delay distribution and current AIDS incidence*, Biometrics 50 (1994), pp. 1203–1214.
[19] J.M. Robins, A. Rotnitzky, and L.P. Zhao, *Analysis of semiparametric regression models for repeated outcomes in the presence of missing data*, J. Am. Statist. Assoc. 90 (1995), pp. 106–121.
[20] W. Tang, H. He, and X.M. Tu, *Applied Categorical and Count Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL, 2012.
[21] P. Wu, D. Gunzler, N. Lu, T. Chen, P. Wymen, and X.M. Tu, *Causal inference for community-based multi-layered intervention study*, Stat. Med. 33(22) (2014), pp. 3905–3918.
[22] P. Wu, Y. Han, T. Chen, and X.M. Tu, *Causal inference for Mann–Whitney–Wilcoxon rank sum and other nonparametric statistics*, Stat. Med. 33(8) (2014), pp. 1261–1271.
[23] Q. Yu, R. Chen, W. Tang, H. He, R. Gallop, P. Crits-Christoph, and X.M. Tu, *Distribution-free inference of negative and zero-inflated Poisson for longitudinal data*, Stat. Med. 32 (2013), pp. 2390–2405.
[24] Q. Yu, W. Tang, J. Kowalski, and X.M. Tu, *Multivariate U-Statistics: A tutorial with applications*, Wiley Interdiscip. Rev: Comput. Stat. 3 (2011), pp. 457–471. DOI: 10.1002/wics.178.



## Appendix

The FRM in Equation (17) differs from conventional distribution-free (semi-parametric) models in that the response function $\mathbf{f}_i$ also involves parameters such as $\pi_i(\boldsymbol{\gamma})$. Below we justify the asymptotic normality of the WGEE estimate $\hat{\boldsymbol{\theta}}$ stated in Equation (20).

Consider the normalized $(1/n)\mathbf{w}(\boldsymbol{\theta})$, but for notational brevity, we continue to denote this normalized quantity as $\mathbf{w}_n$. By applying a Taylor expansion of $\mathbf{w}(\boldsymbol{\theta})$ [12], we have

$$\sqrt{n}\mathbf{w}_n = -\left(\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{w}_n\right)^{\top}\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) + \mathbf{o}_p(1), \tag{A1}$$

where $\mathbf{o}_p(\cdot)$ denotes the stochastic version $\mathbf{o}(\cdot)$ [12]. It follows from the (weak) law of large numbers that

$$\left(\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{w}_n\right)^{\top} = E\left[\frac{\partial}{\partial\boldsymbol{\theta}}(D_i V_i^{-1}\Phi_i S_i)\right]^{\top} \to_p E\left[D_i V_i^{-1}\Phi_i\frac{\partial}{\partial\boldsymbol{\theta}}S_i\right] = B, \tag{A2}$$

where $\to_p$ denotes convergence in probability. It follows from Equations (30) and (31) that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) = -B^{-1}\sqrt{n}\mathbf{w}_n + \mathbf{o}_p(1) \to_d N(\mathbf{0}, \Sigma_\theta), \tag{A3}$$

where $\Sigma_\theta$ is given in Equation (20).

Note that if $\mathbf{f}_i$ is free of any parameter, then $(\partial/\partial\boldsymbol{\theta})S_i = -D_i$ and $B$ in Equation (32) simplifies to $B = -E(D_i V_i^{-1}\Phi_i D_i^{\top})$ as in standard WGEE [12,15,19].