The Annals of Applied Statistics
2014, Vol. 8, No. 2, 726–746
DOI: 10.1214/14-AOAS725

MAXIMUM LIKELIHOOD AND PSEUDO SCORE APPROACHES FOR PARAMETRIC TIME-TO-EVENT ANALYSIS WITH INFORMATIVE ENTRY TIMES[1]

BY BRIAN D. M. TOM, VERNON T. FAREWELL AND AND SHEILA M. BIRD

MRC Biostatistics Unit

We develop a maximum likelihood estimating approach for time-to-event Weibull regression models with outcome-dependent sampling, where sampling of subjects is dependent on the residual fraction of the time left to developing the event of interest. Additionally, we propose a two-stage approach which proceeds by iteratively estimating, through a pseudo score, the Weibull parameters of interest (i.e., the regression parameters) conditional on the inverse probability of sampling weights; and then re-estimating these weights (given the updated Weibull parameter estimates) through the profiled full likelihood. With these two new methods, both the estimated sampling mechanism parameters and the Weibull parameters are consistently estimated under correct specification of the conditional referral distribution. Standard errors for the regression parameters are obtained directly from inverting the observed information matrix in the full likelihood specification and by either calculating bootstrap or robust standard errors for the hybrid pseudo score/profiled likelihood approach. Loss of efficiency with the latter approach is considered. Robustness of the proposed methods to misspecification of the referral mechanism and the time-to-event distribution is also briefly examined. Further, we show how to extend our methods to the family of parametric time-to-event distributions characterized by the generalized gamma distribution. The motivation for these two approaches came from data on time to cirrhosis from hepatitis C viral infection in patients referred to the Edinburgh liver clinic. We analyze these data here.

1. Introduction. The modeling of the time from disease onset or infection (i.e., initiating event) to an outcome of relevance is of considerable importance in studies of the natural history of a disease and in projection of disease burden. Prospective studies which recruit and follow an appropriate cohort of subjects from disease onset to the event of interest are ideal for this purpose. However, these studies are inefficient in terms of resources if the event of interest tends to occur well after disease onset, as is the case for hepatitis C virus (HCV) studies of progression to cirrhosis from initial infection. The alternative is to follow a prevalent cohort of cross-sectionally sampled subjects who, prior to recruitment, have already expe-

Received March 2013; revised September 2013.

[1]Supported in part by the Medical Research Council (Unit Programme numbers U105261167, U105260794).

Key words and phrases. Biased data, generalized gamma distribution, outcome-dependent sampling, pseudo score, robust standard error, survival analysis, Weibull distribution.

rienced the initiating event (e.g., HCV infection) but not yet the event of interest (e.g., cirrhosis). The left truncated time-to-event data obtained from such a study provide a length-biased sample of the incident population, if sampling is such that an assumption of stationarity over calendar time for the occurrence of the initiating event can be made. Methods for handling both incidence data and such length-biased prevalence data have been well described in the (bio)statistics literature [Andersen et al. (1993), Wang, Brookmeyer and Jewell (1993), Kalbfleisch and Prentice (2002), Brookmeyer (2005), Keiding (2005), Wang (2005), Tsai (2009), Qin and Shen (2010)].

A less explored situation is the analysis of prevalence data arising from a referral cohort where entry into the cohort is dependent on a subject's residual fraction of time remaining to the event of interest, and inference on the incident population is required. Such data are believed to occur in HCV studies conducted in tertiary care settings, where HCV patients are more likely to be referred to specialist clinics at later stages of disease [Fu et al. (2007)]. The conventional truncation likelihood approach which simply conditions on the time of entry into the cohort does not work here, as the referral time and the time to the event are correlated. The ignoring of this referral bias has led to higher rates of progression to cirrhosis being reported in studies in specialist clinics compared to those in community-based settings [Freeman et al. (2001)]. As cirrhosis linked to HCV infection is a major epidemic of the 21st century, it is extremely important to get an accurate picture of the present and future disease burden facing affected regions in order to inform public health decisions and actions.

The aforementioned type of referral or outcome-dependent sampling bias is particularly difficult to deal with unless a full specification (up to unknown parameters) of the probability sampling generating mechanism is provided. In practice, this mechanism will rarely be known and, instead, an approximate formulation of the sampling distribution, which is reasonably robust to misspecification, would be sought.

Previously, Fu, Tom and Bird (2009) proposed a weighted pseudo score [Lawless (1997), Cook and Lawless (2007)] or inverse probability weighted method for estimating the parameters of a Weibull regression model for the incubation period from infection to cirrhosis for the community of hepatitis C virus-infected individuals, when there is cirrhosis-related referral bias to the studied prevalent cohort. The method assumed that everyone in the community would come to clinical attention at or before cirrhosis, so that cirrhosis events are not missed. Therefore, the target community population was assumed "immortal" (in the sense of no competing events), and individuals observed in the study sample to have experienced a cirrhotic event were associated with a weight of one in the estimation procedure. However, for other individuals, Fu, Tom and Bird (2009) used approximate weights and, therefore, consistency of these estimated weights, and, consequently, the regression parameter estimates of interest, was, in general, not guaranteed.

Here we outline a full likelihood approach to this outcome-dependent referral problem in which the likelihood for the joint distribution of the time to referral and the time to outcome of interest, both from the initiating event, is fully specified. In practice, depending on the dimensionality of the joint parameter space, the full likelihood may be difficult to maximize over both the regression parameters of interest and the parameters associated with the time-to-entry process. Therefore, we also investigate another strategy based on a hybrid two-stage approach that iteratively alternates between estimating the parameters associated with the time-to-outcome distribution (i.e., regression and shape parameters) from a pseudo score with fixed weights and then estimating the parameters associated with the time-to-entry/referral process from the profiled full likelihood assuming the regression and shape parameters are known. We retain the assumption of an immortal cohort, although this can be relaxed [Copas and Farewell (2001)]. Primarily, we describe the approaches where the time-to-event distribution is assumed Weibull. However, we show how the methods can be extended to the family of parametric time-to-event distributions characterized by the generalized gamma distribution [Stacy (1962), Stacy and Mihram (1965), Prentice (1974), Farewell and Prentice (1977), Lawless (1980), Cox et al. (2007)], for which the Weibull is an important special case.

2. Notation, framework and assumptions. For individuals in the target/incident population, let the calendar time of the initiating event be Y and the calendar period of interest for inference on this population be between calendar times d_1 and d_2. Therefore, $d_1 \leq Y \leq d_2$. Clinical observation of an individual will be left truncated at their time of referral to the clinic which is the time of entry into the cohort for those referred before d_2. Let the time intervals from Y to potential referral and to the event of interest be R and T, respectively, and denote by Z the $p \times 1$ vector of explanatory variables. We assume that the time-to-event T from Y in the incident population comes from a Weibull distribution with support on the positive real line and with positive shape and scale parameters, γ and λ, respectively, where $\lambda = \exp(\beta^T z)$ for given $Z = z$ and β is a vector of regression parameters associated with z. More explicitly, the density and distribution functions of T from an initiating event calendar time $Y = y$, and given the vector of explanatory variables $Z = z$, are $f_T(t|y,z) = \{\gamma \exp(-\gamma\beta^T z)\} \exp[-\{t/\exp(\beta^T z)\}^\gamma] t^{\gamma-1}$ and $F_T(t|y,z) = 1 - \exp[-\{t/\exp(\beta^T z)\}^\gamma]$, respectively. As there is no dependence on the actual value of y in these functions, we simplify the notation for the density and distribution functions of T to $f_T(t|z)$ and $F_T(t|z)$, respectively. Additionally, we assume, as is done for length-biased sampling problems, that within the calendar period $[d_1, d_2]$, the rate of occurrence of the initiating event remains constant. The independence of the distribution of T from when its initiating event occurred and the stationarity of the initiating event process within the calendar period of interest are together referred to as the steady state or equilibrium condition [Wang (2005)].

An individual is assumed to be included in the studied prevalent cohort if $0 < R < d_2 - Y$, with $S = I(0 < R < d_2 - Y)$ the indicator variable denoting selection/inclusion. In addition to the assumption that selected patients will experience the event of interest and be referred prior to the time of the event, we assume the following for the individuals in the target population.

ASSUMPTION 1 (Truncation before outcome). The truncation (or potential referral or entry) time of an individual is always less than the time to outcome and so $R < T$.

ASSUMPTION 2 (Conditional truncation time). For a known vector $\nu = (\nu_0, \ldots, \nu_{m+1})^T$, with $\nu_0 = 0$, $\nu_{m+1} = 1$ and $\nu_j < \nu_{j+1}$ $(j = 0, \ldots, m)$, and unknown mixture probability vector $\pi' = (\pi_0, \ldots, \pi_m)^T$ with $\sum_{j=0}^m \pi_j = 1$, the distribution of R given $T = t$ (for $t > 0$) is a mixture of independent uniform random variables with support in the interval $[0, t)$, density function

$$f_{R|T}(r|t) = \sum_{j=0}^{m} \frac{\pi_j}{(\nu_{j+1} - \nu_j)t} I(\nu_j < r/t \le \nu_{j+1})$$

and cumulative distribution function

$$F_{R|T}(r|t) = \sum_{j=0}^{m} \frac{\pi_j \{\min(r, \nu_{j+1}t) - \max(0, \nu_j t)\}}{(\nu_{j+1} - \nu_j)t} I(\nu_j < r/t).$$

The form chosen for this conditional density reflects the belief that the residual fraction, $1 - r/t$, of time remaining to the event of interest (or, alternatively, the fraction, r/t, of event time elapsed) drives whether a subject is referred [Fu, Tom and Bird (2009)]. It is constructed as a mixture of uniforms so as to allow flexibility in the shape of distribution that can be captured. A notable feature of the random variable $V = R/T$ (for $T > 0$), corresponding to the fraction of time elapsed to the event of interest, is its independence from T (see theorem in the supplementary material [Tom, Farewell and Bird (2014)]). We will subsequently investigate the impact of misspecifying the partitioning of ν on results obtained.

For selected subjects ($S = 1$), denote by C the censoring time from entry into the cohort, and let $X = \min(T, R + C)$ be the observed follow-up time until the outcome event or censoring, with $\Delta = I(T - R < C)$ the "right censoring" indicator taking the value 1 when uncensored. As the calendar period of interest for inference on this population is between d_1 and d_2, then for selected subjects, $d_2 - Y \ge X$. That is, follow-up beyond d_2 is not planned. Additionally, we assume that (T, R) is independent of C [conditional on either Z or (Z, Y)] and that the parameters governing the distribution of C are distinct from those governing the joint distribution of (T, R). That is, the censoring process is ignorable.

To proceed with estimation, we make the following further simplifying assumption:

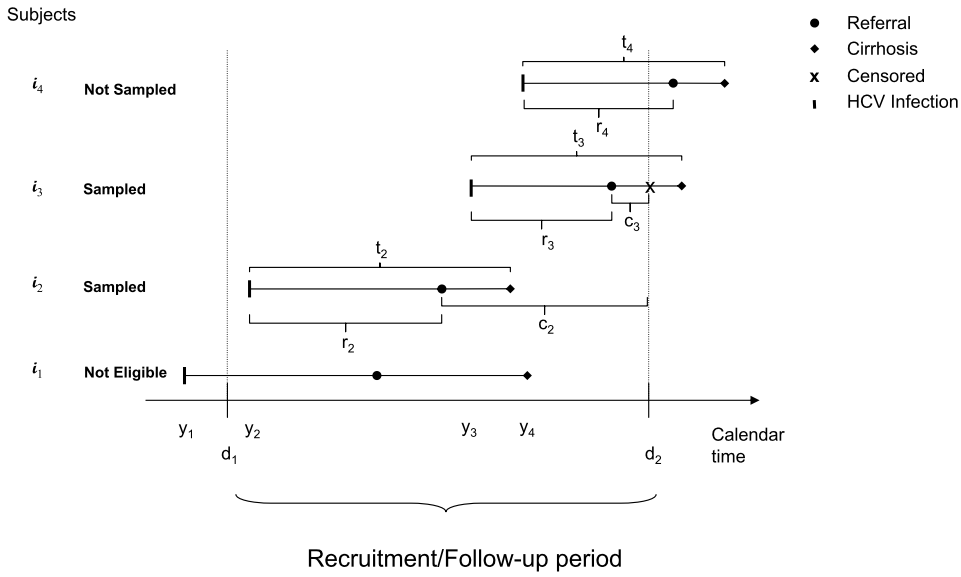

FIG. 1. *Prevalent referral cohort sampling setup.*

ASSUMPTION 3 (Known initiation time). The calendar time of the initiating event can be determined for those subjects selected for inclusion in the cohort.

In Section 4 we discuss how one would proceed if the time of the initiating event is best known to within an interval. Figure 1 presents pictorially the salient features of our prevalent referral cohort design setup.

3. Estimation methods.

3.1. *Maximum likelihood approach.* Let n be the number of individuals who have been selected into the cohort. For an included individual $i \in \{1, \ldots, n\}$, let the observed data be $(r_i, x_i, \delta_i, y_i, z_i)$, which are assumed to be independent realizations of $(R_i, X_i, \Delta_i, Y_i, Z_i)$. Under Assumptions 1 and 3, the ignorability of the censoring process and conditional on $\{Z_i\}$ and $\{Y_i\}$, the full likelihood for $\theta^T = (\gamma, \beta^T, \pi^T)$, where $\pi = (\pi_1, \ldots, \pi_m)^T$, can be written (with, for conciseness, some abuse of notation for continuous variables) as

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} \{\mathrm{pr}(R_i = r_i, T_i = x_i | Y_i = y_i, Z_i = z_i, S_i = 1)^{\delta_i} \\
&\qquad \times \mathrm{pr}(R_i = r_i, T_i \geq x_i | Y_i = y_i, Z_i = z_i, S_i = 1)^{1-\delta_i}\} \qquad (1) \\
&= \prod_{i=1}^{n} L_i(\theta).
\end{aligned}
$$

The first term in the product is the likelihood contribution if x_i corresponds to the true time-to-event t_i (i.e., $\delta_i = 1$) and the second when a right censored event time is observed (i.e., $\delta_i = 0$).

When $\delta_i = 1$ and setting $u_i = d_2 - y_i$, it can be shown that

$$\begin{aligned}&\operatorname{pr}(R_i = r_i, T_i = x_i | Y_i = y_i, Z_i = z_i, S_i = 1)\\ &\quad = \frac{f_{R|T}(r_i|x_i) f_T(x_i|z_i)}{\operatorname{pr}(0 < R_i < u_i)}.\end{aligned}$$

In the situation where $\gamma > 1$ (i.e., the hazard rate of T increases over time), and defining $\varphi = (\gamma - 1)/\gamma$, the denominator, $\operatorname{pr}(0 < R_i < u_i)$, can be analytically evaluated and is found to be

$$\begin{aligned}&\sum_{j=0}^{m} \frac{\pi_j}{(\nu_{j+1} - \nu_j)} \big[\{\nu_{j+1} F_T(u_i/\nu_{j+1}|z_i) - \nu_j F_T(u_i/\nu_j|z_i)\}\\ &\qquad + u_i e^{-\beta^T z_i} \Gamma(\varphi) \{F_G\big((u_i/\nu_j)^\gamma; e^{-\gamma\beta^T z_i}, \varphi\big)\\ &\qquad\qquad - F_G\big((u_i/\nu_{j+1})^\gamma; e^{-\gamma\beta^T z_i}, \varphi\big)\}\big]\\ &= \sum_{j=0}^{m} \frac{\pi_j}{(\nu_{j+1} - \nu_j)} \big[\{\nu_{j+1} F_T(u_i/\nu_{j+1}|z_i) - \nu_j F_T(u_i/\nu_j|z_i)\}\\ &\qquad + u_i e^{-\beta^T z_i} \{\Gamma\big(\varphi, e^{-\gamma\beta^T z_i}(u_i/\nu_{j+1})^\gamma\big)\\ &\qquad\qquad - \Gamma\big(\varphi, e^{-\gamma\beta^T z_i}(u_i/\nu_j)^\gamma\big)\}\big]\end{aligned} \tag{2}$$

with $F_G(u; r, s) = \gamma(s, ru)/\Gamma(s) = \{\Gamma(s) - \Gamma(s, ru)\}/\Gamma(s)$ the cumulative distribution function of a gamma random variable with rate $r > 0$ and shape $s > 0$, evaluated at u $(0 < u < \infty)$, where $\gamma(s, u) = \int_0^u t^{s-1} e^{-t}\, dt$ and $\Gamma(s, u) = \int_u^\infty t^{s-1} e^{-t}\, dt$ denote the lower and upper incomplete gamma functions and $\Gamma(s) = \int_0^\infty t^{s-1} e^{-t}\, dt$ the ordinary gamma function. Details of the derivation are provided in the supplementary material [Tom, Farewell and Bird (2014)] for the family of parametric time-to-event distributions characterized by the generalized gamma distribution with either monotonically increasing or arc shaped (upside-down bathtub) hazards [Glaser (1980), Cox et al. (2007)].

For selected individuals with $\delta_i = 0$, the likelihood contribution in (1), $\operatorname{pr}(R_i = r_i, T_i \geq x_i | Y_i = y_i, Z_i = z_i, S_i = 1)$, can be written as

$$\frac{\operatorname{pr}(R_i = r_i, T_i \geq x_i | Y_i = y_i, Z_i = z_i)}{\operatorname{pr}(0 < R_i < u_i)},$$

where it can be shown (see the supplementary material [Tom, Farewell and Bird (2014)]) that when $\gamma > 1$, the numerator, $\operatorname{pr}(R_i = r_i, T_i \geq x_i | Y_i = y_i, Z_i = z_i) =$

$\operatorname{pr}(R_i = r_i, T_i \geq x_i | Z_i = z_i)$, takes the closed form

$$\begin{aligned}
&\sum_{j=0}^{m} \frac{\pi_j}{(\nu_{j+1} - \nu_j)} \big[\Gamma(\varphi) e^{-\beta^T z_i} I\big(\nu_j < \min(r_i/x_i, \nu_{j+1})\big) \\
&\qquad \times \big\{ F_G\big((r_i/\nu_j)^\gamma; e^{-\gamma\beta^T z_i}, \varphi\big) \\
&\qquad\quad - F_G\big((r_i/\min(r_i/x_i, \nu_{j+1}))^\gamma; e^{-\gamma\beta^T z_i}, \varphi\big) \big\} \big] \\
&= \sum_{j=0}^{m} \frac{\pi_j}{(\nu_{j+1} - \nu_j)} \big[e^{-\beta^T z_i} I\big(\nu_j < \min(r_i/x_i, \nu_{j+1})\big) \\
&\qquad \times \big\{ \Gamma\big(\varphi, e^{-\gamma\beta^T z_i} (r_i/\min(r_i/x_i, \nu_{j+1}))^\gamma\big) \\
&\qquad\quad - \Gamma\big(\varphi, e^{-\gamma\beta^T z_i} (r_i/\nu_j)^\gamma\big) \big\} \big].
\end{aligned} \tag{3}$$

For the case where $\gamma < 1$ (i.e., the hazard rate of T is monotonically decreasing over time), similar closed-form expressions for $\operatorname{pr}(0 < R_i < u_i)$ and $\operatorname{pr}(R_i = r_i, T_i \geq x_i | Z_i = z_i)$ can be obtained but with the upper incomplete gamma function of the form $\Gamma(\varphi, (u/\lambda)^\gamma)$ replaced with $(u/\lambda)^{\gamma\varphi} E_{1-\varphi}((u/\lambda)^\gamma)$ in (2) and (3), where $E_p(z)$ denotes the generalized exponential integral with $p > 1$ and $z \geq 0$. However, for this present paper, we consider only $\gamma > 1$, as it is difficult to envisage in our context a situation where an initially decreasing hazard rate over time would arise.

The maximum likelihood estimates, $\hat{\theta}$, for θ can now be obtained by substituting these various expressions for the terms in (1) into $L(\theta) = \prod_{i=1}^{n} L_i(\theta)$ and then maximizing $l(\theta) = \log L(\theta) = \sum_{i=1}^{n} l_i(\theta)$ over θ. Estimates of the standard errors for $\hat{\theta}$ are obtained from inverting the observed information matrix, $-\partial^2 l(\theta)/\partial\theta\, \partial\theta^T$, evaluated at $\hat{\theta}$.

3.2. *Hybrid pseudo score/profile likelihood approach.* As an alternative to the full likelihood approach, a pseudo score method based on inverse probability weights can be developed [Cook and Lawless (2007)]. We assume that the incident population has N individuals with initiating event times occurring in the period d_1 to d_2. The weighted pseudo score, $U_1(\psi, \pi)$, with $\psi^T = (\gamma, \beta^T)$, is constructed by weighting the Weibull score contributions, $\partial l_i^W/\partial\psi$ for selected subjects by $w_i = 1/p_i$ $(i = 1, \ldots, n)$, where p_i is the selection probability for subject i. This weighted pseudo Weibull score, which has expectation zero, takes the form

$$\begin{aligned}
U_1(\psi, \pi) &= \sum_{i=1}^{N} S_i w_i \frac{\partial l_i^W}{\partial \psi} \\
&= \sum_{i=1}^{N} \frac{S_i}{p_i} \frac{\partial}{\partial \psi} \big[\delta_i \log f_T(x_i|z_i) + (1 - \delta_i) \log\big(1 - F_T(x_i|z_i)\big) \big].
\end{aligned}$$

For a selected study subject i (i.e., $S_i = 1$), p_i is either $\text{pr}(0 < R_i < u_i = d_2 - y_i | T_i = x_i)$ if $\delta_i = 1$ or $\text{pr}(0 < R_i < u_i | T_i \geq x_i)$ if $\delta_i = 0$, with $x_i \leq u_i$. The former probability expression evaluates to 1, as a subject who is observed to have experienced the event of interest would have $t_i = x_i \leq u_i$ and since $T_i > R_i$ (by Assumption 1), then, with probability 1, $R_i < u_i$. The latter probability expression is shown in the supplementary material [Tom, Farewell and Bird (2014)] to be $\{\text{pr}(0 < R_i < u_i) - F_T(x_i|z_i)\}/\{1 - F_T(x_i|z_i)\}$, which is a function of θ. These expressions are derived under the supposition that no further follow-up information on referred individuals beyond d_2, the close of the study, is available. This reflects the situation in our application. However, these expressions can be easily modified to take account of further follow-up information beyond the close of study, as shown in the supplementary material [Tom, Farewell and Bird (2014)] for selected individuals with $\delta_i = 0$ and $x_i > u_i$. The former probability expression for an uncensored selected individual i is trivially $F_{R_i|T_i}(u_i|x_i)$, where x_i can now be greater than u_i.

Estimation of $\theta^T = (\gamma, \beta^T, \pi^T)$ under this second approach proceeds in two stages. First, ψ, the vector of Weibull shape and regression parameters, is estimated by setting the pseudo score, $U_1(\psi, \pi)$, to zero and solving for ψ with given $\{p_i\}$ to get the maximum weighted pseudo score estimates of ψ. Next, the inclusion probabilities $\{p_i\}$ for selected subjects with $\delta_i = 0$ are reevaluated at these maximum weighted pseudo score estimates and at the maximum profile likelihood estimate of π obtained after maximizing $l(\theta)$ over π with ψ set in (1) to its maximum weighted pseudo score estimates. These two steps are iterated until convergence of the estimates for θ to $\tilde{\theta}$. Initially the inclusion probabilities $\{p_i\}$ are all assumed to take the value 1 and, therefore, the initial estimate of ψ is from the standard (unweighted) Weibull regression model. This iterative estimation procedure is similar to that used by Hardin and Hilbe (2003) for longitudinal data, although, to minimize efficiency loss, we do not adopt their assumption of orthogonality of the estimating equations.

Estimated standard errors based on this approach can be obtained either through a standard bootstrap procedure or determined based on Taylor series expansion arguments applied to the set of unbiased estimating equations $U_1(\psi, \pi) = 0$ and $U_2(\psi, \pi) \equiv \partial l(\theta)/\partial \pi = 0$. Under appropriate regularity conditions, the asymptotic joint distribution of $((\tilde{\psi} - \psi)^T, (\tilde{\pi} - \pi)^T)$ is Gaussian with expectation zero and variance–covariance matrix consistently estimated by the robust sandwich matrix $\Sigma \Lambda \Sigma^T$ evaluated at $\tilde{\theta}$, where Σ^{-1} is

$$-\begin{pmatrix} \dfrac{\partial U_1}{\partial \psi^T} & \dfrac{\partial U_1}{\partial \pi^T} \\ \dfrac{\partial U_2}{\partial \psi^T} & \dfrac{\partial U_2}{\partial \pi^T} \end{pmatrix}$$

and $\Lambda = \sum_{\{i\,:\,S_i=1\}} U_{0i} U_{0i}^T$, where $U_{0i}^T = (U_{1i}^T, U_{2i}^T) = (w_i(\theta)\, \partial l_i^W/\partial \psi^T, \partial l_i/\partial \pi^T)$ for $S_i = 1$, with the dependency of $w_i(\theta)$ on θ explicitly shown. With this extra

notation, it is easily seen that $U_1(\psi,\pi)=\sum_{i=1}^n U_{1i}$ and $U_2(\psi,\pi)=\sum_{i=1}^n U_{2i}$, and

$$\frac{\partial U_1}{\partial \psi^T}=\sum_{i=1}^n\left(\frac{\partial l_i^W}{\partial \psi}\frac{\partial w_i}{\partial \psi^T}+w_i\frac{\partial^2 l_i^W}{\partial \psi\,\partial \psi^T}\right),\qquad \frac{\partial U_1}{\partial \pi^T}=\sum_{i=1}^n\frac{\partial l_i^W}{\partial \psi}\frac{\partial w_i}{\partial \pi^T},$$

$$\frac{\partial U_2}{\partial \psi^T}=\sum_{i=1}^n\frac{\partial^2 l_i}{\partial \pi\,\partial \psi^T}\quad\text{and}\quad \frac{\partial U_2}{\partial \pi^T}=\sum_{i=1}^n\frac{\partial^2 l_i}{\partial \pi\,\partial \pi^T}.$$

3.3. *Simulation study*: *Consistency, efficiency and robustness considerations*. To illustrate the performance of the proposed methods, in particular, with regard to efficiency, bias and robustness, we conducted a small-scale simulation with a design similar to that in Fu et al. (2007), Fu, Tom and Bird (2009). We performed 500 simulation runs and generated, in each of the runs, a community sample size of $N=5000$. We considered three different time-to-event distributions from which to simulate our data. These were (i) the Weibull, (ii) the gamma and (iii) the log-normal. The parameter configurations for these three distributions were (i) $\psi_W^T=(\gamma_W,\beta_0,\beta_1,\beta_2)=(4,4.6,-0.03,-0.4)$, (ii) $\psi_G^T=(\gamma_G,\beta_0,\beta_1,\beta_2)=(12.71,1.96,-0.03,-0.4)$, and (iii) $\psi_{LN}^T=(\sigma_{LN},\beta_0,\beta_1,\beta_2)=(0.275,4.464,-0.03,-0.4)$, corresponding to the shape parameters, γ_W and γ_G, scale parameter, σ_{LN}, and the regression parameters $\beta=(\beta_0,\beta_1,\beta_2)^T$ associated with the covariate vector z comprising of an intercept, a continuous variable, z_1, generated from a log-normal distribution with location and scale parameters taking the values 3 and 0.3, respectively, and a binary variable, z_2, generated from a Bernoulli distribution with success probability of $1/3$. These covariates are included through a log-linear regression model on the Weibull's and gamma's scale parameters and a linear model on the log-normal's location parameter. The parameters β_0, γ_W, γ_G and σ_{LN} were chosen to make the log baseline means from the regression models corresponding to the three distribution all equal to 4.50. The truncation times (in years), which are entry times for those selected, are generated from the conditional distribution proposed earlier with $\pi'=(0.1,0.06,0.12,0.24,0.48)^T$ and $\nu=(0,0.5,0.625,0.75,0.875,1)^T$. For simplicity in interpretation of the various simulation results to be presented, we assume everyone in the community experienced the initiating event at the same calendar date and those whose truncation time was less than $d_0=15$ years entered the referral cohort. Administrative right censoring of sampled subjects occurred at $c_0=15$ years from the calendar date of the initiating event. This was the only type of censoring considered here. The parameters of the Weibull distribution were informed by the data that arose from the Edinburgh Royal Infirmary's hepatitis C virus liver clinic, which are analyzed later.

Correct specification of the time-to-event distribution. Table 1 presents the findings from the aforementioned simulation. Both approaches produce consis-

TABLE 1
Full and hybrid pseudo score/profiled Weibull likelihood simulation results

True distribution	Parameters	Full likelihood			Hybrid pseudo score/profile likelihood			
		Mean	$\overline{\text{SE}}$	ESE	Mean	$\overline{\text{RSE}}$	ESE	RE
Weibull	β_0	4.635	0.289	0.287	4.631	0.386	0.397	0.525
	β_1	−0.031	0.005	0.006	−0.030	0.007	0.007	0.604
	β_2	−0.405	0.093	0.090	−0.404	0.109	0.112	0.645
	γ_W	4.041	0.345	0.342	4.062	0.361	0.401	0.729
	π_1	0.060	0.020	0.020	0.061	0.018	0.019	1.011
	π_2	0.120	0.030	0.031	0.120	0.029	0.031	0.998
	π_3	0.241	0.041	0.040	0.240	0.040	0.040	1.011
	π_4	0.480	0.047	0.047	0.479	0.050	0.049	0.929
Gamma	β_0	4.378	0.276	0.307	4.325	0.277	0.335	0.838
	β_1	−0.028	0.005	0.006	−0.027	0.005	0.006	0.839
	β_2	−0.405	0.091	0.091	−0.402	0.103	0.111	0.677
	γ_W	6.658	0.772	0.779	6.821	0.646	0.883	0.779
	π_1	0.059	0.022	0.022	0.060	0.019	0.022	1.066
	π_2	0.114	0.039	0.041	0.113	0.037	0.039	1.080
	π_3	0.233	0.054	0.056	0.231	0.053	0.054	1.063
	π_4	0.482	0.066	0.071	0.475	0.066	0.070	1.046
Log-normal	β_0	4.344	0.269	0.308	4.338	0.276	0.392	0.617
	β_1	−0.028	0.004	0.005	−0.028	0.005	0.007	0.671
	β_2	−0.402	0.094	0.101	−0.416	0.113	0.136	0.554
	γ_W	7.919	0.993	1.071	8.027	0.767	1.230	0.757
	π_1	0.062	0.023	0.025	0.062	0.020	0.025	1.007
	π_2	0.108	0.042	0.043	0.109	0.040	0.041	1.088
	π_3	0.225	0.061	0.062	0.224	0.059	0.060	1.053
	π_4	0.492	0.076	0.078	0.484	0.076	0.075	1.073

Mean, average of the estimate; $\overline{\text{SE}}$, average of the estimated standard error; ESE, empirical standard error; $\overline{\text{RSE}}$, average of the estimated robust standard error; RE, the empirical variance of the maximum likelihood estimator divided by the empirical variance of the maximum hybrid pseudo score/profile likelihood estimator.

tent estimates of the parameters from the Weibull model and the sampling mechanism when the time-to-event distribution was Weibull. As expected, more efficient estimates of the shape and regression parameters were obtained from the full likelihood approach than the hybrid pseudo score/profile likelihood approach. Similar estimated standard errors were obtained from both approaches for the corresponding estimates of π. This perhaps reflects the near optimality of the hybrid approach when estimating π since the relevant part of the full likelihood is being used.

Misspecification of the time-to-event distribution. Table 1 also shows the results when the true time-to-event distributions are gamma and log-normal but the

full likelihood and hybrid pseudo score/profile likelihood approaches were fitted assuming the time-to-event distribution was Weibull. Here we see that the impact of this incorrect assumption for the time-to-event distribution is negligible for the estimation of the regression parameters β_1 and β_2 and minor for the estimation of π. The estimates of the log baseline mean under misspecification of the true gamma and log-normal distributions by the Weibull [i.e., $\log(\Gamma(1+1/\gamma_W))+\beta_0$] were approximately 4.31 and 4.28, respectively. As earlier mentioned, the true log baseline mean is 4.50. Therefore, for these particular cases of misspecification there is underestimation of the log baseline mean. This underestimation of the log baseline mean would result in an underestimation of the tail probabilities of the marginal population time-to-event distribution, which would lead to underestimation of the population size.

Misspecification of the referral mechanism. To investigate the relative robustness of the proposed mixture of uniforms for the conditional distribution of the truncation times given the time to event, we began by rerunning our simulation study as before, except with the conditional distribution of the truncation time now generated from a single uniform distribution in the interval zero to the true time to event instead of the five-component mixture of uniforms. However, the *less parsimonious* five-component mixture of uniforms, with $\nu = (0, 0.5, 0.625, 0.75, 0.875, 1)^T$, was assumed as the working conditional distribution of the truncation times when fitting the full likelihood and hybrid approaches to the simulated data at each simulation run. The results are shown in Table 2. The less parsimonious working conditional distribution for the truncation times has no

TABLE 2
Full and hybrid pseudo score/profiled likelihood Weibull simulation results under a less parsimonious representation of the truncation time conditional distribution

	Full likelihood			Hybrid pseudo score/ profile likelihood		
Parameters	**Mean**	$\overline{\text{SE}}$	**ESE**	**Mean**	$\overline{\text{RSE}}$	**ESE**
β_0	4.606	0.255	0.258	4.604	0.238	0.252
β_1	−0.030	0.005	0.005	−0.030	0.005	0.005
β_2	−0.400	0.085	0.087	−0.400	0.084	0.086
γ	4.057	0.394	0.385	4.056	0.363	0.380
π_1	0.125	0.027	0.026	0.125	0.026	0.026
π_2	0.125	0.030	0.030	0.125	0.030	0.031
π_3	0.127	0.032	0.031	0.127	0.031	0.031
π_4	0.125	0.032	0.032	0.125	0.032	0.031

Mean, average of the estimate; $\overline{\text{SE}}$, average of the estimated standard error; ESE, empirical standard error; $\overline{\text{RSE}}$, average of the estimated robust standard error.

apparent impact, as would be expected, on consistent estimation of the regression and shape parameters of the Weibull distribution. This suggests that finer partitions of ν than needed do not impact on consistency of estimated regression and shape parameters, although may inflate the standard errors of these estimates and the mixture probability estimates.

To explore the impact of misspecification due to the incorrect partitioning of ν, we again repeated the simulation study but now allowing the truncation times to be generated from an eight-component mixture of uniform conditional distribution, with $\nu = (0, 0.125, \ldots, 0.875, 1)^T$ and $\pi' = (0.025, 0.05, 0.1, 0.1, 0.125, 0.15, 0.2, 0.25)^T$, mimicking a strong preference for referrals to occur in the last half of individuals' incubation period. Additionally, we considered three scenarios for recruitment and administrative censoring, which reflected an increasing number of individuals referred and observed experiencing the event of interest and thus providing more information to the analysis: (i) $c_0 = d_0 = 15$; (ii) $c_0 = d_0 = 20$; and (iii) $c_0 = d_0 = 30$. We fitted the simulated data sets assuming the working five-component conditional truncation time distribution mentioned earlier, which is based on a *coarser* partitioning of ν than the true generating mechanism. The results are shown in Table 3. Here, we see a noticeable negative impact of misspecification, from a cruder partitioning of ν, on the estimates of the regression parameters and shape parameter (and mixture probabilities) when using the full likelihood approach (i.e., bias), which diminishes as the amount of information from the sample increases (as reflected by the diminishing standard errors). There is, however, no noticeable bias observed in the estimates of the regression parameters and shape parameter obtained using the hybrid approach. Moreover, there is no apparent bias, under this hybrid approach, in the mixture probabilities, except for π_1, where the bias decreases as the information content increases. This result suggests that the hybrid approach is significantly more robust to misspecification than the full likelihood approach under various data scenarios, but at the cost of being less efficient in general. This perhaps is due to the hybrid approach being a two-stage method.

3.4. *Application to Edinburgh Royal Infirmary's hepatitis C virus liver clinic.* The hepatitis C virus epidemic is a major public health concern in the UK and across the world. To project national hepatitis C virus burden, unbiased estimation of the progression rate from infection to liver cirrhosis is required for the whole community of hepatitis C viral infected individuals. Often, however, the available data on progression to cirrhosis are from a biased sample of the population of interest. In the application we consider here, the data on 387 individuals infected with the hepatitis C virus prior to 2000 (i.e., within the calendar period 1950 to 2000) arose from the Edinburgh Royal Infirmary's hepatitis C virus liver clinic, a tertiary referral hospital clinic whereby patients with more rapid disease progression, or symptomatic disease, would be preferentially referred, with referral

TABLE 3
Full and hybrid pseudo score/profiled Weibull likelihood simulation results under misspecification of the truncation time conditional distribution by a cruder partitioning

	Full likelihood			Hybrid pseudo score/ profile likelihood		
Parameters	**Mean**	$\overline{\text{SE}}$	**ESE**	**Mean**	$\overline{\text{RSE}}$	**ESE**
$c_0 = d_0 = 15$						
β_0	3.841	0.133	0.156	4.790	0.383	0.388
β_1	−0.018	0.003	0.004	−0.033	0.007	0.007
β_2	−0.211	0.052	0.056	−0.427	0.107	0.108
γ	5.231	0.391	0.470	4.048	0.341	0.392
π_1	0.244	0.030	0.028	0.252	0.032	0.037
π_2	0.132	0.030	0.029	0.156	0.035	0.035
π_3	0.169	0.032	0.033	0.193	0.037	0.038
π_4	0.214	0.036	0.038	0.244	0.041	0.043
$c_0 = d_0 = 20$						
β_0	4.224	0.100	0.112	4.654	0.158	0.174
β_1	−0.023	0.002	0.003	−0.031	0.003	0.004
β_2	−0.290	0.039	0.041	−0.417	0.059	0.062
γ	4.565	0.227	0.253	4.020	0.212	0.230
π_1	0.199	0.018	0.018	0.196	0.019	0.019
π_2	0.146	0.019	0.019	0.153	0.020	0.020
π_3	0.184	0.020	0.020	0.194	0.022	0.022
π_4	0.230	0.023	0.023	0.247	0.024	0.025
$c_0 = d_0 = 30$						
β_0	4.527	0.051	0.048	4.603	0.056	0.054
β_1	−0.028	0.001	0.001	−0.030	0.002	0.001
β_2	−0.374	0.023	0.023	−0.402	0.025	0.025
γ	4.114	0.112	0.111	4.013	0.109	0.112
π_1	0.151	0.009	0.010	0.150	0.009	0.010
π_2	0.152	0.010	0.010	0.152	0.010	0.011
π_3	0.197	0.011	0.011	0.199	0.011	0.011
π_4	0.245	0.012	0.013	0.249	0.013	0.013

Mean, average of the estimate; $\overline{\text{SE}}$, average of the estimated standard error; ESE, empirical standard error; $\overline{\text{RSE}}$, average of the estimated robust standard error.

increasingly likely to be closer to onset of cirrhosis. Thus, it is important to account for this outcome-dependent recruitment when analyzing these data so as to provide realistic estimates of the progression rates and the effects of risk factors on time to cirrhosis from infection for the Edinburgh's community (of unknown size) of hepatitis C virus-infected individuals.

To investigate the pattern of referral over patients' cirrhosis incubation period, we model the referral time R given the cirrhosis time T as coming from the prob-

ability density function

$$f_{R|T}(r|t) = \sum_{j=0}^{1} \frac{4\pi_j}{t} I\left(\frac{j}{4} < \frac{r}{t} \le \frac{j+1}{4}\right) + \sum_{j=2}^{5} \frac{8\pi_j}{t} I\left(\frac{j+2}{8} < \frac{r}{t} \le \frac{j+3}{8}\right),$$

where $\{\pi_j : j = 0, \ldots, 5\}$ are the unknown mixture probabilities, summing to 1, that are required to be estimated. This distribution is chosen because of the clinical belief that patients are more likely to be referred in the last half of their cirrhosis incubation period and we therefore decided to model this half in more detail.

Furthermore, we assume that, for the ith hepatitis C virus patient in the community, the time to cirrhosis from known infection time follows a Weibull distribution with unknown shape and scale parameters, γ and λ_i, respectively. The scale parameter, λ_i, is related to the ith patient's continuous and binary explanatory variables, age at hepatitis C viral infection (x_{1i}) and excessive alcohol consumption (x_{2i}), through the relationship $\log \lambda_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$, where $\beta^T = (\beta_0, \beta_1, \beta_2)$ is the vector of regression parameters.

Table 4 shows the results obtained on fitting the Edinburgh Royal Infirmary data using both the full likelihood and hybrid pseudo score/profile likelihood approaches. The bootstrap standard errors are obtained from a bootstrap sample of 500. Relatively similar regression parameter estimates and corresponding estimated standard errors are obtained from the two approaches. The belief by clinicians that referral was more likely in the last half of the cirrhosis period is borne out

TABLE 4

Full and hybrid pseudo score/profile likelihood results for time to cirrhosis from hepatitis C virus (HCV) infection data from Edinburgh Royal Infirmary's liver clinic prior to 2000

	Full likelihood		**Hybrid pseudo score/ profile likelihood**		
Parameters	**Estimate**	**s.e.**	**Estimate**	**Robust s.e.**	**Bootstrap s.e.**
β_0	4.410	0.123	4.380	0.174	0.181
β_1	−0.022	0.004	−0.023	0.004	0.004
β_2	−0.521	0.082	−0.494	0.109	0.111
γ	4.948	0.408	5.256	0.441	0.467
π_1	0.096	0.020	0.093	0.014	0.015
π_2	0.065	0.055	0.060	0.050	0.046
π_3	0.126	0.076	0.139	0.070	0.065
π_4	0.064	0.036	0.062	0.035	0.036
π_5	0.639	0.055	0.635	0.051	0.041
Log-likelihood	−1259.78				
N (bootstrap IQR)			4196	(3414–5139)	

s.e., standard error; N, estimated size of Edinburgh's hepatitis C virus community prior to 2000; IQR, inter-quartile range; β_1 and β_2, regression coefficients corresponding to age at HCV infection and excessive alcohol consumption status.

with about 90% of infected individuals estimated to be referred then. Strikingly, about 64% of infected individuals are estimated to have been referred in the last one eighth of their cirrhosis period. On repeating the analysis with a cruder representation of the referral mechanism based on partitioning ν only into halves produced fairly similar regression estimates under both approaches (data not shown) to those in Table 4. However, for the more variable Weibull shape parameter, there are noticeable differences in the estimates from this cruder referral mechanism to those previously obtained, in particular, for the full likelihood approach. The estimates (with standard errors) of the shape parameter are now 3.833 (0.360) and 4.755 (0.400) for the full likelihood and hybrid approaches, respectively. Additionally, the estimates of the probability of being referred in the last half of the incubation period, obtained assuming the cruder referral mechanism, are roughly equal under the two approaches but now calculated to be approximately 98% as opposed to the 90% previously estimated.

From the hybrid method, we obtain an estimate (bootstrap inter-quartile range) of 4196 (3414–5139) infected individuals in Edinburgh's hepatitis C virus community prior to 2000, through the summation of the inverse probability weights. Both older age at onset of infection and excessive alcohol consumption are found statistically significantly to increase the rate of progression to cirrhosis. The corresponding relative risk estimates (with 95% confidence intervals) for age at infection onset and for excessive alcohol consumption are 1.13 (1.09, 1.18) and 13.4 (5.1, 35.3), respectively. The (inverse probability weighted) estimates of the mean age at HCV infection (with standard deviation) and the proportion consuming excessive alcohol in the Edinburgh HCV community prior to 2000 are 20.3 (6.3) years and 6.7%, respectively. For comparison, the mean age at HCV infection (with standard deviation) and the proportion consuming excessive alcohol from the Edinburgh Royal Infirmary clinic data are 22.4 (9.8) years and 30%, respectively. There is a striking difference in the community's and clinic's estimates of the proportion consuming excessive alcohol.

An estimated marginal 30-year progression rate (with sampling uncertainty) to cirrhosis from infection in the Edinburgh HCV community can also be calculated through a fast parametric bootstrap-like approach [Aalen et al. (1997)]. Here we repeatedly sample $\theta^T = (\gamma, \beta^T, \pi^T)$ from the asymptotic distribution of $\tilde{\theta}$, specified by a multivariate normal distribution with mean vector and variance–covariance matrix given by $\tilde{\theta}$ and the robust sandwich matrix, $\Sigma\Lambda\Sigma^T$ evaluated at $\tilde{\theta}$. For each of these sampled parameter vectors, we define a corresponding hypothetical Edinburgh HCV community (prior to 2000) that can be entirely followed up to cirrhosis. The size, mean and standard deviation of the age at HCV infection and the excessive alcohol consumption proportion for each of these hypothetical communities are calculated by applying the inverse probability weights, calculated using the sampled θ's, to the Edinburgh Royal Infirmary data. The communities' cirrhosis data can be constructed by generating the times to cirrhosis from the proposed

Weibull model using the sampled regression and shape parameters, after first simulating the explanatory variables, age at HCV infection and alcohol consumption status, for the created communities. We assume that the age at HCV infection and alcohol consumption status distributions are independent from one another and are log-normal and Bernoulli, respectively, with mean and standard deviation (for the log normal) and excessive alcohol consumption proportion parameters arising from the application of the sampled θ, through the generated inverse probability weights, to the Edinburgh Royal Infirmary data. Our assumption of marginal independence is based on an estimated Pearson's correlation between age at HCV infection and excessive alcohol consumption status of 0.012 in the actual collected data, which we do not anticipate to change dramatically when translated to the community.

The application of this fast parametric bootstrap-like approach, over 500 runs, gave a mean 30-year progression rate to cirrhosis in the hypothetical communities of 14% with a standard deviation of 6% and a 95% range of 6% to 29%. Figure 2 provides an example of a marginal Kaplan–Meier curve for one hypothetical Edinburgh community generated at the maximum weighted pseudo score estimates.

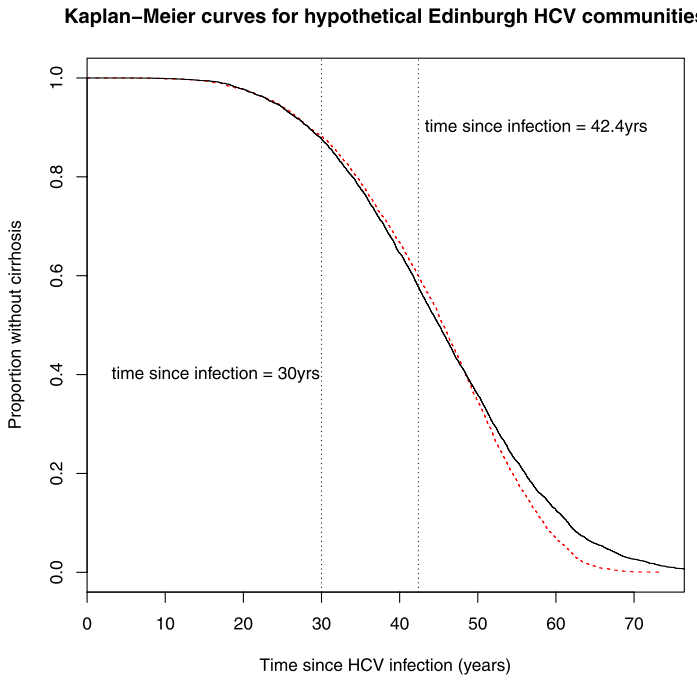

FIG. 2. *Marginal Kaplan–Meier curves for hypothetical Edinburgh HCV communities derived under the assumption that the time-to-cirrhosis distribution is either Weibull* (*solid line*) *or generalized gamma* (*dashed line*). *The vertical dotted lines correspond to time from infection of* 30 *years and the last observed time in the Edinburgh liver clinic data of* 42.4 *years, which corresponds to an* (*uncensored*) *cirrhotic event.*

The estimated 30-year Kaplan–Meier progression rate (with 95% confidence interval) to cirrhosis based on the actual Edinburgh Royal Infirmary data, ignoring the outcome-dependent referral and left truncation, is 42% (31%, 52%). The corresponding conditional Kaplan–Meier estimate (conditioning on not experiencing cirrhosis at least roughly 1 year after infection), assuming that the Edinburgh Royal Infirmary data is a length-biased sample of the Edinburgh HCV community, is 86% with a 95% confidence interval of (75%, 92%). Both of these standard estimates dramatically overestimate the 30-year progression rate, as they do not account for the correlation between the referral time and the time to cirrhosis of a referred patient.

To check the robustness of our findings for the Edinburgh data, we implemented our two approaches replacing the assumption of a Weibull time-to-event distribution with the generalized gamma distribution (see the supplementary material [Tom, Farewell and Bird (2014)] for its formulation), which has one extra parameter and includes the Weibull, gamma and log-normal all as special cases. Although a likelihood ratio test on 1 degree of freedom ($p = 0.02$) rejected the Weibull in favor of the generalized gamma, the maximum likelihood estimates for the regression parameters of interest were similar to those previously obtained ($\hat{\beta}_1 = -0.021$ and $\hat{\beta}_2 = -0.546$) and the estimate of the proportion of infected individuals referred to in the last one eighth of their cirrhosis period was again 64%. Furthermore, the estimated mean 30-year progression rate to cirrhosis was similar. On closer inspection, we found that the differences between the assumption of a generalized gamma and that of a Weibull for the time-to-event distribution was noticeable only in the upper tails of the estimated marginal time-to-event distributions of the Edinburgh HCV community, past the actual largest observed time to cirrhosis (i.e., an uncensored event of 42.4 years) seen from the Edinburgh Royal Infirmary's liver clinic. Similar to the marginal Kaplan–Meier curve presented under the assumption that the time-to-event distribution is Weibull, Figure 2 also displays the equivalent Kaplan–Meier curve under the assumption of the generalized gamma, and thus provides an illustration of the discrepancy between curves being evident in the upper tail beyond 42.4 years.

4. Discussion. A weighted pseudo score method is commonly suggested for handling response-biased observations, where specifying the full likelihood is difficult. Provided that the inverse probability weights can be consistently estimated, then consistency of regression parameter estimates will be achieved using this approach. However, if the full likelihood is available, then it is generally preferable to use it to estimate the parameters of interest, as these estimates will be more efficient than those obtained from the weighted pseudo score method. This preference for the full likelihood over the weighted pseudo score method also holds when the time-to-event distribution is misspecified. In the context of misspecification, we would advocate fitting the more flexible generalized gamma time-to-event distribution (or an alternative such as a semi-parametric piecewise exponential-type

distribution), instead of the Weibull, in order to get better estimates of the marginal progression rates. Nevertheless, it is worth noting that a by-product of the weighted pseudo score approach, which makes it appealing, is the straightforward estimation of the total incident population size. This is not directly available (although calculable) from the full likelihood approach. In public health terms, estimation of the total number of HCV carriers and the "true" impact of covariates on HCV progression are key.

In the informative entry time problem addressed here, we were able to develop both a full likelihood approach and a hybrid two-stage pseudo score/profile likelihood approach for outcome-dependent referral where sampling is dependent on the residual fraction of time remaining to develop the event. Under correct specification of the referral mechanism, we found that the full likelihood approach was indeed more efficient than the hybrid approach in the estimation of the regression parameters of interest and the shape parameter. The former approach, however, appeared to be more susceptible to bias if the outcome-dependent referral mechanism was misspecified through a coarser representation and the "information content" of the data (in terms of number of referrals, number of events and length of follow-up) was low. In the situation where the "information content" is considered to be relatively high, it perhaps would be more appealing to adopt the hybrid method over the full likelihood approach, as it could be significantly more robust and the decrease in the resulting efficiency may still be acceptable. In general, we would recommend that when using either approach, and, in particular, the full likelihood one, analysts should begin by specifying a reasonably fine partitioning of the ν which can then be refined to obtain a more parsimonious representation of the referral mechanism. This strategy would allow checking for sensitivity/robustness to misspecification of the referral mechanism. However, convergence issues may arise if the partitioning is too fine or if the selection probability for a subject is too small. We have not investigated these convergence issues here.

The application of these methods to data from the Edinburgh Royal Infirmary's hepatitis C virus liver clinic allowed us to characterize realistically the extent of Edinburgh's HCV epidemic prior to 2000 in terms of progression rate to cirrhosis and the impact of alcohol consumption and age at HCV infection on this progression. Standard survival analysis methods severely overestimated the 30-year progression rate and underestimated the relative risks for the explanatory variables.

In our present analysis of the Edinburgh HCV data, we assumed that the time of infection was known. This simplifying assumption was thought reasonable in our case, since even when the times of HCV infection in the Edinburgh liver clinic were uncertain, this uncertainty tended to be only in the determination of the exact date of infection within a calendar year or two. As the mean incubation period to cirrhosis is several orders of magnitude greater than the size of this interval, we expect that, for the analysis of our data set, the added uncertainty in estimation due to this imprecision in timing of infection will be inconsequential. However,

in other applications where the timing of the initiation event (e.g., cancer or HIV infection onset) is known only to within an interval, which may be quite large, and where either the mean time to the event of interest (e.g., death or AIDS) or the mean follow-up time are not of an appreciably long enough length compared to the mean width of these intervals, the implications for analyses of assuming-known initiation time (e.g., by choosing the mid-point of the interval) can be major. Struthers and Farewell (1989) discuss an approach to account for unknown onset times, when the time of onset is known only to be in an interval, say, (a, b). This approach requires the specification of a density, say, g, for the time of infection (e.g., a uniform distribution) over the interval (a, b). The likelihood to be optimized over the parameters then takes the form $\prod_{i=1}^{n} \int_{a_i}^{b_i} L_i(\theta) g_i(y; \tau)\, dy$, where $L_i(\theta)$ is the likelihood contribution from the ith subject, given known infection time, and the density, $g_i(\cdot)$, for this subject's time of infection may be specified up to an unknown parameter vector, τ. Therefore, it can be seen that this approach can be adapted to our situation where the sampling is dependent on the residual fraction of time left to developing the event of interest and the onset time is known only to within an interval. However, careful thought is required on the most appropriate form for g. For example, in HCV studies where the majority of subjects are injecting drug users and when time of HCV infection is unknown, there is evidence to suggest that infection occurs earlier in a subject's injecting career [Hutchinson, Bird and Goldberg (2005), Hagan et al. (2008), De Angelis et al. (2009)].

Future application of these methods to the HCV epidemic in Scotland, more generally, is planned with Health Protection Scotland. Health Protection Scotland has developed a clinical database on referrals of HCV patients to liver clinics across all regions of Scotland. Application of our methodology should provide regional estimates for the number of HCV carriers in Scotland and will allow us to examine if the "true" impact of covariates (such as age at HCV infection and heavy alcohol use) are stable across regions although the covariate distribution may differ between regions.

Acknowledgments. We would like to acknowledge the anonymous referees and Editor for their insightful comments and suggestions which have helped improve the paper.

SUPPLEMENTARY MATERIAL

Appendix: Derivations of the expressions based on the generalized gamma and mixture of uniforms (DOI: 10.1214/14-AOAS725SUPP; .pdf). Proofs of the various expressions required in the constructing of the likelihood and pseudo score based on the assumption that the time-to-event distribution is from a generalized gamma distribution and the conditional referral distribution is a mixture of independent uniforms.

REFERENCES

AALEN, O. O., FAREWELL, V. T., ANGELIS, D. D., DAY, N. E. and GILL, O. N. (1997). A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: Application to AIDS prediction in England and Wales. *Stat. Med.* **16** 2191–2210.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. MR1198884

BROOKMEYER, R. (2005). Biased sampling of cohorts. In *Encyclopedia of Biostatistics*, 2nd ed. (P. Armitage and T. Colton, eds.) 427–439. Wiley, New York.

COOK, R. J. and LAWLESS, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer, Berlin.

COPAS, A. J. and FAREWELL, V. T. (2001). Incorporating retrospective data into an analysis of time to illness. *Biostatistics* **2** 1–12.

COX, C., CHU, H., SCHNEIDER, M. F. and MUÑOZ, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Stat. Med.* **26** 4352–4374. MR2405358

DE ANGELIS, D., SWEETING, M., ADES, A. E., HICKMAN, M., HOPE, V. and RAMSAY, M. (2009). An evidence synthesis approach to estimating hepatitis C prevalence in England and Wales. *Stat. Methods Med. Res.* **18** 361–379. MR2750101

FAREWELL, V. T. and PRENTICE, R. L. (1977). A study of distributional shape in life testing. *Technometrics* **19** 69–75.

FREEMAN, A. J., DORE, G. J., LAW, M. G., THORPE, M., OVERBECK, J. V., LLOYD, A. R., MARINOS, G. and KALDOR, J. M. (2001). Estimating progression to cirrhosis in chronic hepatitis C virus infection. *Hepatology* **34** 809–816.

FU, B., TOM, B. D. M. and BIRD, S. M. (2009). Re-weighted inference about hepatitis C virus-infected communities when analysing diagnosed patients referred to liver clinics. *Stat. Methods Med. Res.* **18** 303–320. MR2750070

FU, B., TOM, B. D. M., DELAHOOKE, T., ALEXANDER, G. J. M. and BIRD, S. M. (2007). Event-biased referral can distort estimation of hepatitis C virus progression rate to cirrhosis, and of prognostic influences. *J. Clin. Epidemiol.* **60** 1140–1148.

GLASER, R. E. (1980). Bathtub and related failure rate characterizations. *J. Amer. Statist. Assoc.* **75** 667–672. MR0590699

HAGAN, H., POUGET, E. R., DES JARAIS, D. C. and LELUTIU-WEINBERGER, C. (2008). Meta-regression of hepatitis C virus infection in relation to time since onset of illicit drug injection: The influence of time and place. *Am. J. Epidemiol.* **168** 1099–1109.

HARDIN, J. W. and HILBE, J. M. (2003). *Generalized Estimating Equations*. Chapman & Hall, London. MR2000388

HUTCHINSON, S. J., BIRD, S. M. and GOLDBERG, D. J. (2005). Modelling the current and future disease burden of hepatitis C among injecting drug users in Scotland. *Hepatology* **42** 711–723.

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley, New York. MR1924807

KEIDING, N. (2005). Delayed entry. In *Encyclopedia of Biostatistics*, 2nd ed. (P. Armitage and T. Colton, eds.) 1404–1409. Wiley, New York.

LAWLESS, J. F. (1980). Inference in the generalized gamma and log gamma distributions. *Technometrics* **22** 409–419. MR0585636

LAWLESS, J. F. (1997). Likelihood and pseudo likelihood estimation based on response-biased observation. In *Selected Proceedings of the Symposium on Estimating Functions* (*Athens*, *GA*, 1996) (V. B. Ishwar, V. P. Godambe and R. L. Taylor, eds.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **32** 43–55. IMS, Hayward, CA. MR1837796

PRENTICE, R. L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika* **61** 539–544. MR0378212

QIN, J. and SHEN, Y. (2010). Statistical methods for analyzing right-censored length-biased data under Cox model. *Biometrics* **66** 382–392. MR2758818

STACY, E. W. (1962). A generalization of the gamma distribution. *Ann. Math. Statist.* **33** 1187–1192. MR0143277

STACY, E. W. and MIHRAM, G. A. (1965). Parameter estimation for a generalized gamma distribution. *Technometrics* **7** 349–358. MR0192586

STRUTHERS, C. A. and FAREWELL, V. T. (1989). A mixture model for time to AIDS data with left truncation and an uncertain origin. *Biometrika* **76** 814–817.

TOM, B. D. M., FAREWELL, V. T. and BIRD, S. M. (2014). Supplement to "Maximum likelihood and pseudo score approaches for parametric time-to-event analysis with informative entry times." DOI:10.1214/14-AOAS725SUPP.

TSAI, W. Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika* **96** 601–615. MR2538760

WANG, M.-C. (2005). Length bias. In *Encyclopedia of Biostatistics*, 2nd ed. (P. Armitage and T. Colton, eds.) 2756–2759. Wiley, New York.

WANG, M.-C., BROOKMEYER, R. and JEWELL, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics* **49** 1–11. MR1221402

MRC BIOSTATISTICS UNIT
ROBINSON WAY
CAMBRIDGE CB2 0SR
UNITED KINGDOM
E-MAIL: brian.tom@mrc-bsu.cam.ac.uk
vern.farewell@mrc-bsu.cam.ac.uk
sheila.brid@mrc-bsu.cam.ac.uk