Title

Weighted hurdle regression method for joint modeling of cardiovascular events likelihood and rate in the US dialysis population.

Journal

Statistics in Medicine, 33(25)

Authors

Sentürk, Damla

Dalrymple, Lorien

Mu, Yi

et al.

Publication Date

2014-11-10

DOI

10.1002/sim.6232

Peer reviewed

Published in final edited form as:
Stat Med. 2014 November 9; 33(25): 4387–4401. doi:10.1002/sim.6232.

Weighted Hurdle Regression Method for Joint Modeling of Cardiovascular Events Likelihood and Rate in the U.S. Dialysis Population

Damla Şentürk[a], **Lorien S. Dalrymple**[b], **Yi Mu**[c], and **Danh V. Nguyen**[d,e,*,†]

[a]Department of Biostatistics, University of California, Los Angeles, California 90095, U.S.A.

[b]Division of Nephrology, Department of Medicine, University of California, Sacramento, California 95817, U.S.A.

[c]Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, California 95616, U.S.A.

[d]Department of Medicine, UC Irvine School of Medicine, Orange, CA 92868-3298, U.S.A.

[e]Institute for Clinical and Translational Science, University of California, Irvine, California 92687-1385, U.S.A.

SUMMARY

We propose a new weighted hurdle regression method for modeling count data, with particular interest in modeling cardiovascular events in patients on dialysis. Cardiovascular disease remains one of the leading causes of hospitalization and death in this population. Our aim is to jointly model the relationship/association between covariates and (a) the probability of cardiovascular events, a binary process and (b) the rate of events once the realization is positive - when the 'hurdle' is crossed - using a zero-truncated Poisson distribution. When the observation period or follow-up time, from the start of dialysis, varies among individuals the estimated probability of positive cardiovascular events during the study period will be biased. Furthermore, when the model contains covariates, then the estimated relationship between the covariates and the probability of cardiovascular events will also be biased. These challenges are addressed with the proposed weighted hurdle regression method. Estimation for the weighted hurdle regression model is a weighted likelihood approach, where standard maximum likelihood estimation can be utilized. The method is illustrated with data from the United States Renal Data System. Simulation studies show the ability of proposed method to successfully adjust for differential follow-up times and incorporate the effects of covariates in the weighting.

Keywords

cardiovascular outcomes; dialysis; end stage renal disease; hurdle model; infection; Poisson regression; United States Renal Data System

*Correspondence to: Danh V. Nguyen, 843 Health Sciences Road, Institute for Clinical and Translational Science, Irvine, CA 92697-1385.
†danhvn1@uci.edu

1 Introduction

As of 2010, end-stage renal disease (ESRD) affected more than 570,000 adults in the United States. Of these, more than 400,000 were on dialysis, a life-sustaining treatment [1]. Annual mortality for patients on maintenance dialysis is approximately 20–25% with overall 5-year survival lower than most malignancies [1]. ESRD is associated with accelerated mortality, and cardiovascular (CV) disease is the leading cause of death, accounting for nearly half of all deaths [1]. Furthermore, CV disease and infection remain the leading causes of hospitalization and death in these patients [1]. Prior studies have shown an increased risk of CV events following infections in the general population [2] based on the United Kingdom General Practice Research Database [3] and similarly in the U.S. dialysis population using the United States Renal Data System (USRDS) database [4] – [7]. Mechanistically, infections may have acute effects on the vascular endothelium and may contribute to a chronic sub-clinical inflammatory state that influences atherogenesis and/or progression of atherosclerosis. To date, studies have not examined the association between patient-level risk factors, including infection, jointly with (a) the likelihood of CV events and (b) the subsequent occurrence (rate) of CV events. Thus, factors associated with (a) or (b) or both are not clear.

Therefore, in this study, we propose a joint model to examine factors associated with CV likelihood (CV “onset”) and subsequent CV occurrence (CV “recurrence/progression”). More specifically, we propose a new weighted hurdle regression method for zero-inflated count data to *jointly* model the association between (a) the likelihood/probability of cardiovascular events, such as myocardial infarction or stroke, during a fixed study period (e.g., five years from the start of dialysis) and (b) the rate of cardiovascular events as a function of individual covariates, including infection-related hospitalization, demographics and comorbidities among other factors. More specifically, we consider a hurdle regression model, which simultaneously models the binary process (presence or absence of cardiovascular events) and a zero-truncated count process for the positive cardiovascular event counts. The standard hurdle model/distribution for the number of cardiovascular events for person i, denoted Y_i, under equal follow-up time for all individuals is

$$\Pr(Y_i=y_i;\lambda_i,\pi_i)=\begin{cases} 1-\pi_i, & y_i=0 \\ \pi_i \frac{e^{\lambda_i}\lambda_i^{y_i}}{(1-e^{-\lambda_i})y_i!} & y_i>0 \end{cases}, \quad (1)$$

where $\pi_i = \Pr(Y_i > 0)$ and distribution of the positive counts is taken to be a zero-truncated Poisson distribution with rate λ_i in (1). Both π_i and λ_i may depend on covariates as detailed in Section 2. The underlying motivation for this model is based on our conceptualization that each patient on dialysis has an individual-specific probability of having a cardiovascular event; thus, a binomial probability model governs the binary outcome process of having no or positive cardiovascular events. Once the realization is positive - the ‘hurdle’ is crossed - the conditional distribution of positive counts of cardiovascular events is modeled as a zero-truncated distribution, such as a zero-truncated Poisson for $y_i > 0$ as in (1). This hurdle model conceptualization, in addition to allowing each individual to potentially have of one or more events, will also account for excess zeros. Excess zeros are common in count data,

as exhibited by our data (Vuong's test for zero-inflation [8]: test statistic −47.22, $p < 0.0001$, indicating significant zero-inflation).

However, when the observation period or follow-up time varies among individuals, ignoring the differential follow-up times will result in biased estimates for the probability of positive cardiovascular events (i.e., the binary outcome) in the standard hurdle model (1). The variable follow-up times for the positive count process is handled naturally through an offset term, as is typically done in log-linear Poisson regression. Our proposed weighted hurdle regression model addresses the variable follow-up time for the binary process by incorporating a weight function $w(t_i)$, where t_i is the follow-up duration for individual i, $w(t_i)$ is increasing in t_i and $w(t_i) = 1$ if an individual's follow-up period is complete (e.g., s/he was followed for five years from the start of dialysis). This weighting accounts for the simple fact that the likelihood of observing a first cardiovascular event after the start of the study is higher for individuals who are followed up for longer periods of time, all other things being equal. When the binary process of the model does not depend on covariates, then $w(t_i)$ can be estimated nonparametrically, based on the Kaplan-Meier estimate [9] of the time to the (first) cardiovascular event distribution. This weighting approach to adjust for the variable follow-up time without covariates is based on the works of [10, 11] who considered weighting in the zero-inflated Poisson (ZIP) model for the recurrence of adenomas, which were based on similar ideas from [12, 13].

When the binary process depends on covariates, in addition to variable individual follow-up times, we develop a weight function $w(t_i, \mathbf{z}_i)$ that depends on covariate values for individual i, denoted $\mathbf{z}_i$, to model the association between the likelihood of cardiovascular events and individual covariates, including demographics and comorbidities. The weight function $w(t_i, \mathbf{z}_i)$, incorporating both follow-up time and covariates, can be estimated semiparametrically based on the Cox regression model [14] or parametrically, for instance.

We note that the standard hurdle model (1) appears to have been developed independently by Mullahy [15] in economics applications, King [16] in political science applications for international relations, and Heilbron [17, 18] for the U.S. National AIDS Behavioral Study data. It accounts for excess zero counts, and in this sense, is similar to the ZIP model originally developed for modeling counts of defective components in manufacturing processes by Lambert [19]. However, the ZIP model formulation assumes that there are two subpopulations, one producing standard counts including zero counts (e.g., via a standard Poisson distribution) and a second subpopulation that produces only zeros. In the manufacturing process application, the second subpopulation can be conceptualized to represent a perfect manufacturing state that produces only perfect components. Similarly, in modeling counts of a behavior (e.g, smoking frequency or sexual behavior) a subpopulation of complete abstainers can be conceptualized that produces additional zero counts. For our application to cardiovascular events (outcome), we choose to develop a weighted hurdle model because it is more biologically plausible that the likelihood of a cardiovascular event depends on an individual's complex underlying genetic propensity, co-existing illnesses (e.g., diabetes, hypertension), habits (e.g., tobacco use), body composition, among other factors.

This paper is organized as follows. We develop the weighted hurdle regression model in Section 2, where we formulate the aforementioned weight functions, with and without covariates, to account for variable individual follow-up times. In Section 3, the proposed method is illustrated using data from the USRDS with an application to modeling counts of cardiovascular events. Simulation studies illustrating the efficacy of the proposed method are summarized in Section 4 and we conclude with a brief discussion in Section 5, where we outline implementation of the proposed method in standard statistical softwares.

2 Weighted Hurdle Regression Model

For simplicity, we first consider incorporating the duration of follow-up in estimating the probability of positive counts of cardiovascular events through a weight function of time-to-event that does not depend on covariates. As introduced earlier, let t_i denote the follow-up time for individual i and let the study period length of interest be τ. For example, we consider $\tau = 5$ years from the initiation of dialysis for our cohort, since the median patient survival is approximately 3 years. To incorporate variable follow-up time, we consider an increasing weight function $w(t_i)$, with $0 < w(t_i) \leq 1$, where $w(t_i) = 1$ indicates that the individual's follow-up period is complete, i.e., $t_i = \tau$. This weighting accounts for the increased likelihood of observing a first cardiovascular event after the start of the study for individuals who are followed up for longer periods of time. Further, let Y_i^* be the true number of cardiovascular events during the full follow-up period of interest (length τ) and Y_i be the observed number of cardiovascular events during the actual follow-up period. To fix notations, consider the simplified case where the probability of positive cardiovascular events during this period does not depend on individual covariate characteristics: $\pi_i = \Pr(Y_i^* > 0) = \pi$, for all i. The probability of no cardiovascular events at time t_i, $\Pr(Y_i = 0; t_i)$, is defined as $1 - F(t_i)$, where $F(t_i)$ is the distribution function for the time to the first cardiovascular event. We assume that the probability of observing positive cardiovascular events at time t_i is $F(t_i) = \Pr(Y_i > 0; t_i) \equiv w(t_i)\Pr(Y_i^* > 0)$. The weighted hurdle model at time t_i, accounting for variable follow-up t_i, is

$$\Pr(Y_i = y_i; t_i, \lambda_i, \pi) = \begin{cases} 1 - w(t_i)\pi, & y_i = 0 \\ w(t_i)\pi \frac{e^{\lambda_i t_i}(\lambda_i t_i)^{y_i}}{(1 - e^{-\lambda_i t_i})y_i!} & y_i > 0 \end{cases}, \quad (2)$$

where the Poisson rate $\lambda_i = \lambda(\mathbf{x}_i)$ depends on q_1 covariates $\mathbf{x}_i$. The weight function of time-to-event $w(t_i)$ is defined to be

$$w(t_i) = \frac{1 - S(t_i)}{1 - S(\tau)}, 0 < t_i \leq \tau, \quad (3)$$

where $S(t_i)$ denotes the survival distribution function. The weight function in (3) does not depend on covariates. Therefore, the weights can be evaluated using the nonparametric Kaplan-Meier survival curve estimate, $\hat{S}(t_i)$, based on the time to first cardiovascular event. That is, $\hat{w}(t_i) = (1 - \hat{S}(t_i))/(1 - \hat{S}(\tau))$.

In more realistic applications, the probability of positive cardiovascular events during the study period of interest will depend on individual covariate characteristics. Thus, we need to

model $\pi_i=\pi(\mathbf{z}_i)=\Pr(Y_i^*>0;\mathbf{z}_i)$, where the q_2 covariates $\mathbf{z}_i$ that affect the binary process may be chosen to be different from the q_1 covariates $\mathbf{x}_i$ that affect the Poisson rates (λ_i) generally. Therefore, we assume a more general model that incorporates variable follow-up time and covariate effects for the binary process; specifically, $F(t_i;\mathbf{z}_i)=\Pr(Y_i>0;t_i,\mathbf{z}_i)\equiv w(t_i,\mathbf{z}_i)\Pr(Y_i^*>0;\mathbf{z}_i)$. Consequently, the more general weighted hurdle model at time t_i, accounting for both variable follow-up t_i and covariates ($\mathbf{z}_i$ and $\mathbf{x}_i$), is

$$\Pr(Y_i=y_i;t_i,\lambda_i,\pi_i)=\begin{cases}1-w(t_i,\mathbf{z}_i)\pi_i, & y_i=0\\ w(t_i,\mathbf{z}_i)\pi_i\frac{e^{\lambda_i t_i}(\lambda_i t_i)^{y_i}}{(1-e^{-\lambda_i t_i})y_i!}, & y_i>0\end{cases}, \quad (4)$$

where the binomial probability π_i now depends on the q_2 covariates $\mathbf{z}_i$, $\pi_i = \pi(\mathbf{z}_i)$, and the Poisson rate is $\lambda_i = \lambda(\mathbf{x}_i)$. Choices for the link functions to relate the covariates $\mathbf{z}_i$ and $\mathbf{x}_i$ to the binomial probabilities (π_i) and Poisson rates (λ_i), respectively, are needed. For this we consider the common logit and log link functions for binary and count outcomes, respectively:

$$\mathrm{logit}(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right)=\mathbf{z}_i^{\mathrm{T}}\gamma\ \log(\lambda_i t_i)=\mathbf{x}_i^{\mathrm{T}}\beta, \quad (5)$$

although other choices of link functions are possible.

Naturally, the weight function (3) generalizes to accommodate the covariates $\mathbf{z}_i$ in the binary process as

$$\omega(t_i,\mathbf{z}_i)=\frac{1-S(t_i;\mathbf{z}_i)}{1-S(\tau;\mathbf{z}_i)}, 0<t_i\leq\tau, \quad (6)$$

where the survival distribution function $S(t_i; \mathbf{z}_i)$ now depends on individual covariates. Depending on the specific model for $S(t_i; \mathbf{z}_i)$ or equivalently $F(t_i; \mathbf{z}_i)$, traditional flexible parametric and semiparametric survival analysis techniques can be utilized to evaluate the weight function $w(t_i, \mathbf{z}_i)$. For this, denote the survival function estimate by $\hat{S}(t_i; \mathbf{z}_i)$; then the plug-in estimate of the weight function is $w(t_i, \mathbf{z}_i)$ with $\hat{w}(t_i, \mathbf{z}_i) = (1 — \hat{S}(t_i; \mathbf{z}_i))/(1 — \hat{S}(\tau; \mathbf{z}_i))$. To more concretely illustrate the computation, consider the popular semiparametric Cox regression for estimating the survival function which depends on individual covariates. In this case, $\hat{S}(t_i,\mathbf{z}_i)=\hat{S}_0(t_i)^{\exp(\mathbf{z}_i^{\mathrm{T}}\hat{\xi})}$, where $\hat{\xi}$ and $\hat{S}_0(t_i)$ are the parameter and baseline survival function estimates from the Cox regression model fit with covariates $\mathbf{z}_i$ and using the time to first cardiovascular event (or the time at the end of follow-up if censored), respectively. We note that generally the coefficients ξ, which are estimated by $\hat{\xi}$, are a function of the parameters of interest, namely 7 for the binomial part of hurdle model: $\mathrm{logit}(\pi_i)=\mathbf{z}_i^{\mathrm{T}}\gamma$. This is further discussed in the Appendix section. As mentioned above, other approaches to estimate survival, including parametric models, may be used to estimate the survival function; the proposed weighting framework for adjustment does not depend on the Cox model estimate of survival or its assumptions.

Thus, the weighted hurdle regression model, defined by (4)–(6), relates (a) the covariates $\mathbf{z}_i$ to the probability of positive cardiovascular events during the study period of length τ and (b) the covariates $\mathbf{x}_i$ to the cardiovascular event rate (given positive cardiovascular events), with both binary and count processes accounting for variable follow-up time t_i. Maximum likelihood estimates of the model parameters (γ, β) can be based on the weighted hurdle likelihood for n subjects from the distribution function (4). Thus, the weighted hurdle likelihood is

$$L(\gamma,\beta)=\prod_{i=1}^{n}[\Pr(Y_i=0;t_i,\lambda_i,\pi_i)]^{1-\delta_i}[\Pr(Y_i=y_i;t_i,\lambda_i,\pi_i)]^{1-\delta_i}, \quad (7)$$

where $\delta_i = \mathbb{I}(y_i > 0)$ and $\mathbb{I}(E)$ is the indicator function for event E. The weighted hurdle log-likelihood $l(\gamma,\beta) \equiv \log L(\gamma,\beta)$ factors: $l(\gamma,\beta) = l(\gamma)l(\beta)$, with

$$l(\gamma)=\sum_{i=1}^{n}\{\mathbb{I}(y_i>0)\log(\pi_i^*)+\mathbb{I}(y_i=0)\log(1-\pi_i^*)\}, \text{ and}$$
$$l(\beta)=\sum_{y_i>0}\left\{y_i\log(\lambda_i^*)-\lambda_i^*-\log(y_i!)-\log(1-\exp(-\lambda_j^*))\right\},$$

where $\pi_i^*=w(t_i,\mathbf{z}_i)(1/(1+\exp(-\mathbf{z}_i^{\mathrm{T}}\gamma))$ and $\lambda_i^*=\exp(\mathrm{x}_i^{\mathrm{T}}\beta+\log(t_i))$. Thus, the MLE for γ and β can be obtained by separately maximizing $l(\gamma)$ and $l(\beta)$, respectively. Additionally, it can be seen that the weights only impact the estimation of β. In (7) the weights, $w(t_i, \mathbf{z}_i)$, are replaced by estimates based on the data as described above. For this reason, and since the weights are estimated in practice, we estimate the weights based on a subset of the data and the remainder of the dataset is retained for estimation of γ in the binomial part of the model. For example, in the data analysis, data for 20% of the subjects were used to estimate the weight function, while data on the remaining 80% of the subjects were used for estimation of γ. We note that although the estimates $\hat{\gamma}$ will target the true parameters γ (without splitting the data), this approach will ensure that inference (e.g., confidence interval coverage) will be correct as well. Implementation using standard software, including SAS and R, are described in the Appendix.

We note that the ideas proposed here directly generalize to other link functions; more generally, $g_1(\pi_i)=\mathbf{z}_i^{\mathrm{T}}\gamma$ and $g_2(\pi_i)=\mathbf{x}_i^{\mathrm{T}}\beta$ for (5). Furthermore, one may also choose to model the positive count via another zero-truncated discrete distribution for (4). Conceptually, little will be gained from presenting the proposed weighted hurdle regression model in more general notations; therefore, we simply define the proposed weighted hurdle regression model through equation (4)–equation (6).

3 Application to Cardiovascular Events in Patients on Dialysis

To illustrate the proposed weighted hurdle regression method, we use data from the United States Renal Data System (USRDS), which collects data on nearly all (> 95%) patients with end-stage renal disease in the U.S. The USRDS is a national database that collects and maintains standard analytic files, including data on inpatient hospitalizations submitted to Medicare, patient demographics, dialysis modality, comorbidities and laboratory measures at the start of dialysis. We used USRDS data with follow-up through December 31, 2009.

The defined population included patients aged 18 or older who newly initiated dialysis between January 1, 2000 and December 31, 2007 without a prior history of renal transplant. Furthermore, patients were eligible for inclusion if (a) they survived the first 90 days of dialysis and did not recover renal function or receive a kidney transplant during this interval, (b) had Medicare as the primary payer on day 91 of dialysis, and (c) were receiving hemodialysis or peritoneal dialysis on day 91 of dialysis. Thus, the observation period began on day 91 of dialysis. The follow-up period for an individual ended at the time of kidney transplantation, renal function recovery, death, study end on December 31, 2009 or five years after the start of dialysis. For this analysis, the study period of interest was $\tau = 5$ years from the start of dialysis; this captured the follow-up lengths of most patients on dialysis. Furthermore, the median survival in this cohort is about 3 years.

The outcome variable was the count of the number of cardiovascular events, defined as myocardial infarction (MI), unstable angina, stroke, or transient ischemic attack (TIA), determined from primary discharge diagnosis and based on the International Classification of Disease, 9th Revision, Clinical Modification (ICD-9-CM) codes. The modeling objective, as described by equations (4)–(6), is to estimate the relationship/association of covariates with (a) the probability of positive cardiovascular events (binary/binomial process) during the five (τ) year-period after the start of dialysis and (b) the rate of cardiovascular events, conditioned on positive cardiovascular events. Covariates of interest include infection-related hospitalization rate during follow-up (per person-year), demographic variables (age, race, sex and ethnicity), comorbidities (congestive heart failure, coronary heart disease, cerebrovascular disease, peripheral vascular disease, hypertension, diabetes, chronic obstructive pulmonary disease and cancer), inability to ambulate or transfer, tobacco use, body mass index (BMI = kg/m^2) and MDRD eGFR (estimated glomerular filtration rate based on the Modification of Diet in Renal Disease (MDRD) equation from the National Kidney Foundation [20]). We restrict our analysis to cases where infection-related hospitalization rate is no more than 12 per person-year during follow-up, which represents 99.4% of the data. Similar to cardiovascular events, an infection-related hospitalization was ascertained based on the principal discharge diagnosis and based on CD-9-CM classification.

The final analysis cohort consists of $n = 434, 547$ subjects. Table 1 summarizes the covariates used for the weighted hurdle regression for the analysis cohort. The average follow-up time was 2.43 years (standard deviation [SD] 1.66), showing that most patients were not followed for the full $\tau = 5$ years from the start of dialysis and, hence, the need to account for variable follow-up time. With respect to infection-related hospitalization rate (mean 0.65 per person-year, SD 1.29) during the study period, 50.55% of individuals have no infection. The remaining 30.09, 10.72, 3.82, 1.83 and 2.98% of individuals have infection-related hospitalization rate in the range of (0, 1], (1, 2], (2, 3], (3, 4] and (4, 12] per person/year, respectively.

The result of the weighted hurdle regression model fit is summarized in Table 2 using the same covariates for the binomial and Poisson process (i.e., $\mathbf{z}_i = \mathbf{x}_i$). The estimates for the binomial component of the weighted hurdle model are provided in Table 2(A) along with 95% confidence intervals (CIs). Older patients and higher infection rate are associated with

an increased likelihood of positive cardiovascular events during the five-year study period, with 2.6% higher odds per year increase in age and 30.7% higher odds with each additional infection-related hospitalization per person-year. Male sex is associated with a lower odds (odds ratio [OR] 0.830, 95% CI 0.811–0.849) as is Hispanic ethnicity. Also, race 'black' and race 'other' have lower odds compared to race 'white'. We observe a slight negative BMI-positive cardiovascular events risk association (OR 0.987), similar to other studies examining BMI and mortality in the hemodialysis population (e.g., [21]). Patients with congestive heart failure, coronary heart disease, cerebrovascular disease, peripheral vascular disease, hypertension and diabetes have a higher likelihood of cardiovascular events. The largest increased odds of cardiovascular events were about 51%, 43% and 33% associated with diabetes, cerebrovascular disease and coronary heart disease, respectively.

Conditioned on observing positive cardiovascular events, estimates of the relative rates (RRs) of cardiovascular events for the zero-truncated Poisson model are presented in Table 2(B). Infection-related hospitalization rate is also positively associated with increased cardiovascular event rates (RR 1.146, 95% CI 1.124–1.169) and males have lower rates relative to females (RR 0.926, 95% CI 0.90 – 0.953). Diabetes, cerebrovascular disease and coronary heart disease are significantly associated with higher rates of cardiovascular events, similar to their effects on the probability of positive cardiovascular events from the binomial component of the model. However, conditioned on positive cardiovascular events, the relative rates of cardiovascular events are no longer associated with race, ethnicity, congestive heart failure, hypertension and tobacco use. (For the interested reader, results from an unweighted, biased, analysis are presented as supplemental materials available at http://www1.icts.uci.edu/dnguyen/suppl_wthurdle.html.)

We note that a general issue for consideration modeling count data is overdispersion - that is, the variance exceeds the mean. In such cases, a negative binomial (NB) model can be used to account for overdispersion, where the zero-truncated Poisson part in (4) by the zero-truncated NB distribution. For our data, we also fitted a NB hurdle model. The estimated dispersion parameter from the NB model fit is extremely small (log(dispersion) = −9.68, SE 13.25, $z = -0.73$, $p = 0.4654$) and does not indicate overdispersion. Not surprising, in this case, the conclusions/interpretations of parameter estimates and their significance from the Poisson hurdle and NB hurdle remain the same. (The estimates from the two model fits are nearly identical [results not shown]).

Finally, we note that, as described in Section 2, estimation of the weight function was based on a random selection of 20% of the subjects and using Cox regression. The remaining 80% of the data were used for estimation of model parameters of interest, γ. We considered sensitivity analyses that varied the amount of data used for estimating the weights; the results suggest that the conclusions summarized in this section above remain the same (results not shown). This issue is examined more systematically in Section 4.3, where we varied the amount of data used to estimate the weight function under small to moderate sample size settings.

4 Simulation Studies

In this section, we report on simulation studies, with simulated data similar to the USRDS data described above. The main objective is to assess the efficacy of the proposed weighted hurdle regression model, through the weighting function $w(t_i, \mathbf{z}_i)$, to target the true regression coefficients γ; hence, correctly estimating the true probabilities of positive cardiovascular events, $\pi_i = \pi(\mathbf{z}_i)$ in the binary process model, $\text{logit}(\pi_i) = \mathbf{z}_i^{\mathrm{T}}\gamma$. The Poisson model parameters β account for variable follow-up time t_i directly through the rate λ_i with offset $\log(t_i)$; therefore, we expect $\hat{\beta}$ to target β correctly. The weights are pertinent to the assessment of whether $\hat{\gamma}$ properly targets γ in the binary process. In what follows we also include comparisons to no weighting (i.e., ignoring differential follow-up time in the binomial model).

4.1 Estimation

In this Monte Carlo simulation study, we considered a combination of discrete and continuous covariates that mimic the corresponding covariates in the USRDS data. More specifically, we considered simulated data with covariates similar to the observed data in USRDS for sex (`Sex`), age at the start of dialysis (`Age`), body mass index (`BMI`) and MDRD eGFR (`eGFR`). For n subjects, we generated 54% males through `Sex` ~ Bin(n, 0.54), and gender-specific age at the start of dialysis using skewed normal (SN) distributions [22]: `Age|male` ~ $SN(\zeta_m, \omega_m, \alpha_m)$ and `Age|female` ~ SN(ζ_f, ω_f, α_f) with location, scale and shape parameters $\zeta_m = 82.34$, $\omega_m = 23.95$, $\alpha_m = -4.92$, respectively; similarly, for females $\zeta_f = 82.02$, $\omega_m = 22.65$, $\alpha_m = -4.39$. Figure 1 shows the SN fits to the observed skewed-left ages of patients. To induce a slight negative correlation of about −0.15 between `BMI` and `Age` in the observed data, `BMI` was generated from a Gamma regression model with identity link: $\texttt{BMI}_i \sim \text{Gamma}(s_i, a)$ with mean depending on $\texttt{Age}_i$, $s_i = (34 - 0.08\ \texttt{Age}_i)/a$, where $a = 14.33$ is the Gamma shape parameter. Finally, `eGFR` was generated from a Gamma(2.82, 3.74) distribution. The above parameters used to generate the covariate data (`Sex, BMI, Age, eGFR`) were based on the observed data. Figure 1 summarizes the simulated covariates along with the observed data. Denote the collection of the covariates for subject i by $\mathbf{z}_i$.

Next, the time-to-event ("first cardiovascular event") for each subject was obtained from a distribution function $F(t_i'; \mathbf{z}_i)$. The follow-up time (censoring) distribution is denoted by $G(t_i)$ and is independent of event times. We considered two time-to-event distributions, Weibull and exponential. For the Weibull distributed time-to-event, $F(t_i'; \mathbf{z}_i) = 1 - \exp(-\theta_i t_i'^{\nu_E})$, where $\theta_i = \theta_E \exp(\tilde{\gamma}^{\mathrm{T}}\mathbf{z}_i)$ depends on the covariates $\mathbf{z}_i$ with coefficient vector γ, $\hat{\theta}_E$ is a baseline parameter, and ν_E is the Weibull shape parameter. For the exponential time-to-event, $F(t_i'; \mathbf{z}_i) = 1 - \exp(-\theta_i t_i')$, where similarly $\theta_i = \theta_E \exp(\tilde{\gamma}^{\mathrm{T}}\mathbf{z}_i)$ and θ_E is the baseline rate parameter for the exponential distribution (i.e., Weibull with shape parameter $\nu_E = 1$). This setup allowed for examining the performance of the proposed weighted hurdle regression model under both constant (exponential) and non-constant (Weibull) hazards for time-to-event. For the follow-up time distribution $G(t_i)$, we also considered Weibull (with scale θ_C and shape ν_C) and exponential (with rate θ_C)

distributions. We selected the "baseline" parameters for $F(t_i^{'};\mathbf{z}_i)$ and $G(t_i)$ as $(\theta_E, \nu_E) = (0.1, 1.5)$ and $(\theta_C, \nu_C) = (0.3, 1.5)$, respectively, for the Weibull case. For the case of exponential distribution we take $\theta_E = 1/3$ and $\theta_C = 1/2$. These parameters were chosen so that the percentage of the observed times of first event $(t_i^{'})$ that occur prior to the follow-up time (t_i) is about 15%, i.e., $\#\left\{t_i^{'}<t_i\right\}/n \approx 0.15$, similar to the USRDS data analysis in Section 3.

Finally, the events count Y_i is simulated according to the hurdle distribution. That is, if the event time is greater than the follow-up time $(t_i^{'}>t_i)$ then $Y_i = 0$ since we would not be able to observe the event, where t_i is taken to be $\min\{t_i, \tau\}$ to ensure that data collection stops at the end of follow-up time or the end of the study period of length $\tau = 5$. Otherwise, if $t_i^{'}>t_i$ then the number of events, Y_i, is simulated from a zero-truncated Poisson distribution with mean $\lambda_i t_i$, where $\lambda_i=\exp(\mathrm{x}_i^{\mathrm{T}}\beta)$. For simplicity, and without loss of generality, we take $\mathbf{x}_i = \mathbf{z}_i$. Throughout, all simulation studies consist of 1,000 simulated datasets/replications.

The simulation study results are provided in Table 3, where a summary of the maximum likelihood estimates for the weighted hurdle regression model over 1,000 simulated datasets, with n = 6,000 8,000 and 10,000, are provided. The true coefficients of interest are $\gamma^{\mathrm{T}} = (-2.0, -0.25, -0.03, 0.05, -0.05)$ and $\beta^{\mathrm{T}} = (-4.0, -0.12, -0.05, 0.07, -0.10)$. Similar to the data analysis, the estimated weights used in the likelihood (7) are based on the fitted Cox regression using 20% of the data ($0.2 \times n$) for the binomial model and the remaining 80% are used for estimating γ. The results from Table 3 show that the estimates $\hat{\gamma}$ target the true parameters γ for the binomial/logit model, $\mathrm{logit}(\pi_i)=\mathbf{z}_i^{\mathrm{T}}\gamma$; and, as expected, the Poisson parameters are correctly targeted as well. These results hold similarly for both non-constant hazard (Weibull distribution; Table 3(A)) and constant hazard (exponential distribution; Table 3(B)). Also, as expected, variation in the parameter estimates decreases with increasing sample size.

As a baseline comparison, we provide in Table 4 the corresponding simulation studies ignoring the differential follow-up time in the binomial fit (i.e., without weighting). These results show that although the variance is decreasing with sample size, the estimation bias of $\hat{\gamma}$ remains without weighting. Clearly, ignoring differential follow-up time in the binomial model is not an option since the estimated probability of events during the study period will be biased.

4.2 Confidence Interval Coverage

As described in Section 2 and implemented in the data analysis, we split data for estimation of the weights and γ separately for the binomial fit. For instance, in the data analysis we used 20% of the subjects for estimating the weights, while data on the remaining 80% of the subjects were used for estimation of γ. As noted in Section 2, although the estimates $\hat{\gamma}$ will target the true parameters γ without splitting the data, this approach will ensure that inference procedures, such as confidence interval coverage will be valid. Thus, we performed simulation studies in this section to examine 95% confidence interval coverage; the results are summarized in Table 5. First, we note the coverages for the Poisson model of the positive counts for the (A) proposed weighted hurdle model accounting variable follow-

up time, and (B) hurdle model without weighting (ignoring variable follow-up time) are near the target of 95%. However, for the binomial part of the model, only the proposed weighted hurdle model accounting for variable follow-up time provide adequate coverage, as expected due to the severe bias.

4.3 Small-Moderate Sample Sizes and Sensitivity to Amount of Data Used to Estimate the Weight Function

In this section we detail simulation studies to further examine the following issues: (a) the performance of the proposed method under small to moderate sample size and (b) the sensitivity to the amount of data used to estimate the weight function. To properly examine these issues, one must keep in mind the *effective sample size* in the context of censored data, such as for estimation of the Cox regression model; this is the number of events (uncensored observations). Under similar simulation design settings as described above, we consider a nominal sample size of $n = 600$, 800 and 1000 and the overall rate of censored observations is 75%. We varied the percentage of data used for estimation from 20%, 30%, 40% to 50% of the nominal sample size n. Thus, with respect to the estimation of the weight function (using Cox regression), the corresponding effective sample sizes, denoted n_e, are $n_e = 30$, 45, 60, and 75 for $n = 600$. Similarly, the effective sample sizes are $n_e = (40, 60, 80, 100)$ and $n_e = (50, 75, 100, 125)$ corresponding to $n = 800$ and $n = 1000$, respectively. These settings were designed to push our proposed method to the breaking point in order to provide practical guidance on the aforementioned issues (a) and (b). The results are summarized in Table 6. First, not surprisingly, the proposed method still performs well (still targets the true parameters well) for small to moderate sample sizes and for a wide range of the amount of data used to estimate the weight function (20% to 50%) compared to no weighting. However, there is a small tradeoff with respect to bias reduction (and variance) when one allocates too much data (e.g., 40%, 50%) for estimating the weight function. For example, at 20% ($n = 800$, $n_e = 40$) compared to 50% ($n = 800$, $n_e = 100$) of data allocated for estimating the weight function, the resulting parameter estimates are $\hat{\gamma} = (-2.147, -0.247, -0.031, 0.053, -0.052)$ compared to $\hat{\gamma} = (-2.192, -0.248, -0.032, 0.054, -0.054)$; $\gamma = (-2.0, -0.25, -0.03, 0.05, -0.05)$. Thus, there is less reduction in bias, albeit very small, when too much data is allocated to weight estimation (e.g., 50%). Also, the variation in $\hat{\gamma}$ is slightly higher for 50% allocation, as expected, since the overall sample size available for estimation in the hurdle model is reduced accordingly (see Table 6). Thus, overall, for small to moderate sample size settings, it is preferable to allocate 20%–30% of the data for estimating the weight function (as compared to allocating excessively large amount of data, e.g., 40% – 50%). This strategy, of course, must be balanced with the basic requirement in terms of the "minimal" number of events (n_e) needed to fit a Cox regression model. For example, consider the case of 20% of $n = 600$ samples allocated for weight function estimation. This is inadequate, because n_e is only 30 events (on average) which si extremely small for fitting a Cox model with 4 covariates (i.e., only 7–8 events per covariates); therefore, estimation of the weight function is not feasible (not stable). Thus, one must necessarily increase the amount of data for this purpose; in this case, 30% of the data provides adequate estimation of the weight function (see Table 6, $n = 600$ case). This suggests a simple practical strategy for small to moderate sample size settings: start with 20% – 30% of the data allocated for weight function estimation and ensure that this also is

adequate data for the fitting the Cox model, which one can use a (minimum) rule of thumb of 10 events per covariates [23, 24]. For example, at 30% data allocation when $n = 600$ with $n_e = 45$ and 4 covariates in the simulation study, this is about 11 events per covariates on average.

5 Discussion

In this work, we proposed a simple and effective weighting method to handle variable follow-up time when simultaneously modeling the association between covariates and (a) the binomial probability of positive events ($y > 0$) during a fixed study period and (b) the rate of events conditioned on observing positive events, using a zero-truncated distribution (such as a zero-truncated Poisson distribution). The proposed weight functions incorporate individual follow-up time and covariate effects. They can be estimated using standard censored regression analysis, such as Cox regression analysis. The weights can then be used to fit the weighted hurdle regression model. As described in more detail in the Appendix, coefficient estimates obtained via maximizing the weighted hurdle likelihood can be implemented in optimization routines using readily available softwares, such as SAS `PROC NLMIXED` and the R function `optim()`.

We developed and implemented several simulation studies with the dependency structure among the covariates that were similar to the real USRDS data. The results show that the weighted hurdle regression estimates target the true covariate effects under variable follow-up time. In illustrating the proposed weighted hurdle regression method with the USRDS data, we identified factors (such as infection and diabetes) associated jointly with an increase in the likelihood of positive cardiovascular events as well as the rate of cardiovascular events, conditioned on having positive cardiovascular events. Likewise, the method also allowed for modeling simultaneously factors that *differentially* affect the binomial and Poisson processes. For example, congestive heart failure and tobacco use were strongly associated with the binomial probability of positive cardiovascular events, but these same factors were no longer associated with the subsequent rate of cardiovascular events from the Poisson process. Thus, conditioned on positive cardiovascular events, these two factors were not associated with the rate of cardiovascular events. In conclusion, the proposed regression method is relatively straight-forward to implement with existing software, accommodates variable follow-up time in modeling count data via joint models for binary and positive count processes, and can flexibly model covariate effects in both processes.

Finally, as we discussed at the end of sections 2 and 3, other discrete distributions may be used instead of the Poisson model. For example, a weighted NB hurdle may be used for data with overdispersion. However, there is a need to develop and study formal testing procedures to compare model fits, particularly, for weighted hurdle models. This is currently an open problem.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Center for Advancing Translational Sciences, National Institute of Health, through grant #UL1 TR000153, by the National Institute of Diabetes and Digestive and Kidney Diseases grant #R01 DK092232 and grant #K23 DK093584, and a grant from Dialysis Clinics, Inc. The interpretation and reporting of the data presented here are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the United States government. We are grateful to two reviewers for their constructive reviews which helped improve the paper.

Appendix

Weighted Hurdle Regression Implementation

Most standard software, including SAS and R, can be used to obtain the estimates $\hat{\gamma}$ and $\hat{\beta}$. The simple steps are as follows. First, survival regression analysis can be used to obtain $\hat{w}(t_i, \mathbf{z}_i)$ as described in Section 2. Depending on the model choice (e.g., parametric, semi-parametric etc.), standard software can be used to obtain the weights; e.g, using SAS `PROC PHREG` or `PROC LIFEREG` or R functions `coxph ()` and `survfit()`). Then the above log-likelihoods following equation (7) can be maximized in standard optimization routines. For example, we have used and tested SAS `PROC NLMIXED` and R general-purpose optimization function `optim()`. Our R codes for the proposed weighted hurdle models are available at http://www1.icts.uci.edu/dnguyen/suppl_wthurdle.html.

Relationship Between γ and ξ

As described in Section 2, the proposed weighted hurdle regression method accounts for variable follow-up time that depends on covariates through the model $F(t_i;\mathbf{z}_i)=\Pr(Y_i>0;t_i,\mathbf{z}_i) \equiv w(t_i,\mathbf{z}_i)\Pr(Y_i^*>0;\mathbf{z}_i)$. Thus, the probability of positive counts during the study period is related to the distribution of time-to-event. More precisely, since the weights $w(t_i, \mathbf{z}_i)$, given by equation (6), depend on ξ and $\pi_i=\Pr(Y_i^*>0;\mathbf{z}_i)$ depends on γ, their relationship can be determined. For example, with Weibull(θ_i, ν_E) distributed time-to-event (see Section 4.1), $F(t_i;\mathbf{z}_i)=1-\exp(-\theta_i t_i^{\nu_E})$, where $\theta_i=\theta_E\exp(\xi^{\mathrm{T}}\mathbf{z}_i)$. Thus, from the logistic model $\mathrm{logit}(\pi_i)=\mathbf{z}_i^{\mathrm{T}}\gamma$, we have that $\xi = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{a}$, where $\mathbf{a}^{\mathrm{T}} = (a_1,\ldots,a_n)$, $a_i=\log\left\{\log(1+\exp(\mathbf{z}_i^{\mathrm{T}}\gamma))/(\tau^{\nu_E}\theta_E)\right\}$ and $\mathbf{Z}$ is the $n \times (q_2 + 1)$ matrix of covariate data. Similar calculations can be made for a given distribution function of time-to-event $F(\cdot)$. We note that the choice of the link function does not influence the estimate of ξ in practice.

References

1. US Renal Data System. USRDS 2011 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States. Bethesda, MD: National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; 2011.
2. Smeeth L, Thomas SL, Hall AJ, Hubbard R, Farrington P, Vallance P. Risk of myocardial infarction and stroke after acute infection or vaccination. New England Journal of Medicine. 2004; 351:2611–2618. [PubMed: 15602021]
3. Walley T, Mantgani A. The UK General Practice Research Database. Lancet. 1997; 350:1097–1099. [PubMed: 10213569]
4. Dalrymple LS, Mohammed SM, Mu Y, Johansen KL, Chertow GM, Grimes B, Kaysen GA, Nguyen DV. The risk of cardiovascular-related events following infection-related hospitalizations

in older patients on dialysis. Clinical Journal of the American Society of Nephrology. 2011; 6:1708–1713. [PubMed: 21566109]

5. Mohammed SM, Senturk D, Dalrymple DS, Nguyen DV. Measurement error case series models with application to infection-cardiovascular risk in older patients on dialysis. Journal of the American Statistical Association. 2012; 107:1310–1323. [PubMed: 23650442]
6. Ishani A, Collins AJ, Herzog CA, Foley RN. Septicemia, access and cardiovascular disease in dialysis patients: the USRDS Wave 2 study. Kidney International. 2005; 68:311–318. [PubMed: 15954922]
7. Foley RN, Guo H, Snyder JJ, Gilbertson DT, Collins AJ. Septicemia in the United States dialysis population, 1991 to 1999. Journal of the American Society of Nephrology. 2005; 15:1038–1045. [PubMed: 15034107]
8. Vuong QH. Likelihood ratio tests for model selection and non-nested hypothesis. Econometrica. 1989; 57:307–333.
9. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958; 53:457–481.
10. Hsu C-H. Joint modelling of recurrence and progression of adenomas: a latent variable approach. Statistical Modelling. 2005; 5:201–215.
11. Hsu C-H. A weighted zero-inflated Poisson model for estimation of recurrence of adenomas. Statistical Methods in Medical Research. 2007; 16:155–166. [PubMed: 17484298]
12. Emerson SS, McGee DL, Fennerty B, Hixson L, Garewal H, Alberts D. Design and analysis of studies to reduce the incidence of colon polyps. Statistics in Medicine. 1993; 12:339–351. [PubMed: 8456216]
13. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. Biometrics. 2000; 56:1177–1182. [PubMed: 11129476]
14. Cox DR. Regression models and life tables. Journal of the Royal Statistical Society B. 1972; 34:187–202.
15. Mullahy J. Specification and testing of some modified count data models. Journal of Econometrics. 1986; 33:341–365.
16. King G. Event count models for international relations: Generalizations and applications. International Studies Quarterly. 1989; 33:123–147.
17. Heilbron DC. Generalized linear models for altered zero probabilities and overdispersion in count data. Unpublished technical report, University of California, San Francisco, Dept. of Epidemiology and Biostatistics. 1989
18. Heilbron DC. Zero-altered and other regression models for count data with added zeros. Biometrical Journal. 1994; 36:531–547.
19. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992; 34:1–14.
20. National Kidney Foundation, Inc. Part 5. Evaluation of laboratory measurements for clinical assessment of kidney disease. American Journal of Kidney Diseases. 2002; 39:S76–S110.
21. Stack AG, Murthy BV, Molony DA. Survival differences between peritoneal dialysis and hemodialysis among “large” ESRD patients in the United States. Kidney International. 2004; 65:2398–2408. [PubMed: 15149353]
22. Azzalini A. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics. 1985; 12:171–178.
23. Concato J, Peduzzi P, Holfold TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. Journal of Clinical Epidemiology. 1995; 48:1495–1501. [PubMed: 8543963]
24. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of Clinical Epidemiology. 1995; 48:1503–1510. [PubMed: 8543964]

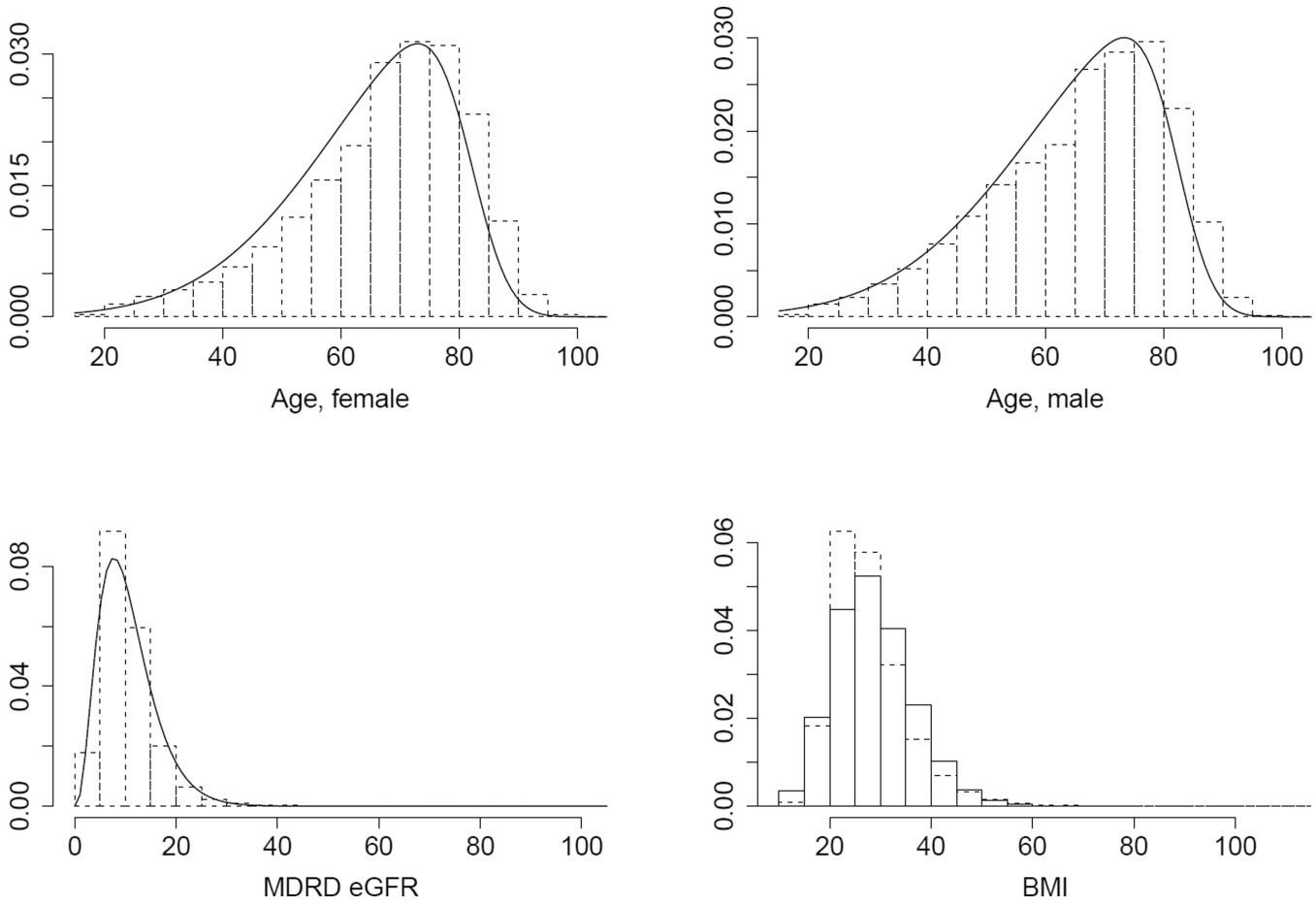

Figure 1.
Simulated data (solid) and real data (dotted) for simulation study.

Table 1

Summary of follow-up time and covariates in the weighted hurdle regression model for n = 434,547 subjects. Provided are mean and standard deviation [SD] for continuous variables (Age, BMI, eGFR, Infection rate) and count and percent for categorical variables.

Variable	Group	Mean or Count	SD or Percent
Follow–up time (years)	-	2.43	1.66
Age (at dialysis)	-	65.75	14.81
BMI	-	27.91	7.52
eGFR	-	10.32	5.40
Infection rate (per person/year)	-	0.64	1.28
Sex	Male	234125	53.88
Race	Black	125756	28.94
Race	Other	28321	6.52
Ethnicity	Non-Hispanic	380669	87.60
Congestive heart failure	Yes	152822	35.17
Coronary heart disease	Yes	122552	28.20
Cerebrovascular disease	Yes	45567	10.49
Peripheral vascular disease	Yes	69113	15.90
Hypertension	Yes	374089	86.09
Diabetes	Yes	252578	58.12
Chronic obstructive pulmonary disea	se Yes	39307	9.05
Tobacco use	Yes	24944	5.74
Cancer	Yes	28875	6.64
Inability to ambulate or transfer	Yes	22088	5.08

Table 2

Weighted hurdle regression fit for the analysis cohort (n = 434, 547) from the United Sates Renal Data System: (A) binomial process and (B) Poisson process.

(A) Binomial model fit

					95% CI[a]	
Variable	**Group**	**Est.**	**Std Error**	**OR[b]**	**Lower**	**Upper**
Intercept		−2.758	0.047	-	-	-
Age (at dialysis)		0.026	0.001	1.026*	1.025	1.027
BMI		−0.014	0.001	0.986*	0.985	0.988
eGFR		−0.001	0.001	0.999	0.997	1.002
Infection rate		0.268	0.009	1.307*	1.283	1.331
Sex	Male	−0.187	0.012	0.830*	0.811	0.849
Race[d]	Black	−0.074	0.014	0.928*	0.904	0.953
Race[d]	Other	−0.087	0.023	0.917*	0.877	0.958
Ethnicity	Non-Hispanic	0.081	0.018	1.085*	1.048	1.123
Congestive heart failure	Yes	0.090	0.013	1.094*	1.067	1.122
Coronary heart disease	Yes	0.283	0.014	1.327*	1.292	1.363
Cerebrovascular disease	Yes	0.355	0.019	1.426*	1.374	1.478
Peripheral vascular disease	Yes	0.134	0.017	1.143*	1.107	1.181
Hypertension	Yes	0.057	0.018	1.059*	1.023	1.096
Diabetes	Yes	0.414	0.013	1.513*	1.476	1.550
Chronic obstructive pulmonary disease	Yes	0.018	0.022	1.018	0.975	1.063
Tobacco use	Yes	0.136	0.026	1.146*	1.090	1.205
Cancer	Yes	−0.144	0.026	0.866*	0.824	0.910
Inability to ambulate or transfer	Yes	−0.109	0.031	0.897*	0.845	0.953

(B) Zero-truncated Poisson model fit

					95%CI	
Variable	Group	Est.	Std Error	RR[c]	Lower	Upper

(A) Binomial model fit						
					95% CI[a]	
Variable	**Group**	**Est.**	**Std Error**	**OR[b]**	**Lower**	**Upper**
Intercept		−1.963	0.064	-	-	-
Age (at dialysis)		0.001	0.001	1.001*	1.000	1.002
BMI		−0.003	0.001	0.997*	0.995	0.999
eGFR		0.004	0.002	1.004*	1.001	1.007
Infection rate		0.137	0.010	1.146*	1.124	1.169
Sex	Male	−0.077	0.015	0.926*	0.900	0.953
Race	Black	−0.018	0.017	0.982	0.950	1.016
Race	Other	0.050	0.028	1.051	0.996	1.109
Ethnicity	Non-Hispanic	−0.027	0.022	0.973	0.932	1.016
Congestive heart failure	Yes	0.009	0.016	1.009	0.978	1.041
Coronary heart disease	Yes	0.247	0.016	1.280*	1.240	1.321
Cerebrovascular disease	Yes	0.150	0.020	1.162*	1.117	1.209
Peripheral vascular disease	Yes	0.052	0.019	1.054*	1.015	1.094
Hypertension	Yes	−0.025	0.023	0.976	0.934	1.020
Diabetes	Yes	0.133	0.017	1.142*	1.105	1.180
Chronic obstructive pulmonary disease	Yes	0.045	0.026	1.046	0.993	1.101
Tobacco use	Yes	−0.003	0.033	0.997	0.935	1.063
Cancer	Yes	0.008	0.033	1.008	0.945	1.077
Inability to ambulate or transfer	Yes	−0.101	0.039	0.904*	0.837	0.977

[a] Confidence interval;

[b] Odds ratio;

[c] Relative rate (or rate ratio);

[d] The reference group for race is ‘White’.

* 95% CI does not contain 1 (significant).

Table 3

Weighted estimation: Average of parameter estimates and standard deviation (SD) over 1,000 Monte Carlo datasets of size n = 6,000, 8,000 and 10,000 for (A) Weibull and (B) Exponential distributed time-to-event. Given are averages over 1,000 datasets.

		n = 6000		n = 8000		n = 10000	
Variable	**Parameter**	**Estimate**	**SD**	**Estimate**	**SD**	**Estimate**	**SD**
(A) Weibull							
Binomial (γ)							
Interc.	−2.00	−2.095	0.461	−2.125	0.407	−2.078	0.375
Male	−0.25	−0.262	0.131	−0.258	0.115	−0.253	0.102
BMI	−0.03	−0.031	0.009	−0.031	0.008	−0.031	0.007
Age	0.05	0.052	0.005	0.052	0.005	0.052	0.004
eGFR	−0.05	−0.052	0.012	−0.051	0.011	−0.052	0.010
Poisson (β)							
Interc.	−4.00	−3.912	0.417	−3.887	0.346	−3.901	0.324
Male	−0.12	−0.116	0.076	−0.119	0.066	−0.123	0.057
BMI	−0.05	−0.050	0.006	−0.051	0.005	−0.050	0.004
Age	0.07	0.069	0.005	0.069	0.004	0.069	0.004
eGFR	−0.10	−0.100	0.010	−0.100	0.008	−0.100	0.007
(B) Exponential							
Binomial (γ)							
Interc.	−2.00	−2.133	0.407	−2.083	0.348	−2.113	0.300
Male	−0.25	−0.258	0.111	−0.256	0.097	−0.253	0.086
BMI	−0.03	−0.031	0.008	−0.031	0.006	−0.031	0.006
Age	0.05	0.052	0.005	0.052	0.004	0.052	0.004
eGFR	−0.05	−0.052	0.011	−0.052	0.009	−0.051	0.008
Poisson (β)							
Interc.	−4.00	−3.890	0.368	−3.896	0.323	−3.884	0.280
Male	−0.12	−0.121	0.070	−0.122	0.060	−0.121	0.053
BMI	−0.05	−0.050	0.005	−0.050	0.005	−0.051	0.004
Age	0.07	0.069	0.004	0.069	0.004	0.069	0.003
eGFR	−0.10	−0.100	0.009	−0.100	0.008	−0.100	0.007

Table 4

Estimation ignoring differential follow-up (no weighting). Given are averages over 1,000 datasets.

		n = 6000		n = 8000		n = 10000	
Variable	**Parameter**	**Estimate**	**SD**	**Estimate**	**SD**	**Estimate**	**SD**
(A) No weighting							
Binomial (γ) - Weibull							
Interc.	−2.00	−3.292	0.279	−3.292	0.245	−3.282	0.221
Male	−0.25	−0.191	0.075	−0.191	0.066	−0.188	0.059
BMI	−0.03	−0.023	0.005	−0.023	0.004	−0.023	0.004
Age	0.05	0.040	0.003	0.040	0.003	0.040	0.002
eGFR	−0.05	−0.039	0.007	−0.039	0.006	−0.039	0.006
Binomial (γ) - Exponential							
Interc.	−4.00	−3.033	0.262	−3.055	0.223	−3.039	0.205
Male	−0.12	−0.196	0.068	−0.194	0.059	−0.194	0.053
BMI	−0.05	−0.024	0.005	−0.024	0.004	−0.024	0.004
Age	0.07	0.041	0.003	0.041	0.002	0.041	0.002
eGFR	−0.10	−0.040	0.007	−0.040	0.006	−0.040	0.005

Stat Med. Author manuscript; available in PMC 2014 November 09.

Table 5

95% confidence interval coverage for the (A) proposed weighted hurdle model accounting for variable follow-up time, and (B) hurdle model without weighting (ignoring variable follow-up time).

	(A)			(B)		
	Sample size *n*					
	6000	8000	10000	6000	8000	10000
Weibull						
Binomial γ						
Interc.	94.1	92.7	92.0	0.4	0.1	0.1
Male	95.0	93.1	94.1	88.5	84.4	81.1
BMI	94.6	94.1	93.7	74.1	65.1	61.6
Age	93.1	90.6	91.7	14.7	5.7	2.4
eGFR	94.0	94.3	93.4	68.3	61.7	50.4
Poisson β						
Interc.	94.9	94.5	93.1	94.2	94.2	94.1
Male	94.1	95.1	95.3	95.9	95.7	96.3
BMI	96.5	96.0	94.7	95.6	94.3	95.4
Age	94.0	95.8	93.1	93.7	93.0	93.4
eGFR	95.9	94.9	95.1	94.5	94.9	94.9
Exponential						
Binomial γ						
Interc.	92.4	93.7	93.8	2.5	0.2	0.0
Male	95.3	94.1	96.3	88.1	83.7	82.2
BMI	94.1	94.8	94.4	73.0	65.7	59.0
Age	90.5	92.2	89.6	10.6	5.8	2.5
eGFR	94.5	93.8	93.5	66.2	58.4	49.1
Poisson β						
Interc.	93.3	93.6	93.7	93.4	95.0	93.3
Male	94.0	94.6	95.2	95.4	94.3	96.1
BMI	95.2	94.2	95.7	95.1	95.0	94.4
Age	92.9	93.7	93.9	94.4	94.2	92.7

(A)	(B)					
	Sample size *n*					
	6000	**8000**	**10000**	**6000**	**8000**	**10000**
eGFR	94.5	95.5	95.4	95.0	94.7	94.2

Table 6

Estimation for small to moderate sample sizes and sensitivity to the amount of data used to estimate the weight function. Results presented are averages over 1,000 datasets for each nominal sample size n and effective sample size n_e (to estimate the weight function) with overall censoring rate of 75% under the exponential model described in Section 4.1. The percentage of data used to estimate the weight function ranges from 20% to 50% and the case of no weighting (0%) is also provided.

		Percent of Data Used to Estimate the Weight Function									
		20%		30%		40%		50%		0% (No wt.)	
Variable	Parameter	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
n=600		(n_e = 30)		(n_e = 45)		(n_e = 60)		(n_e = 75)			
Interc.	−2.00	−	−	−2.179	1.155	−2.133	1.243	−2.157	1.329	−2.681	0.749
Male	−0.25	−	−	−0.263	0.329	−0.258	0.335	−0.267	0.369	−0.199	0.192
BMI	−0.03	−	−	−0.032	0.022	−0.033	0.024	−0.032	0.026	−0.026	0.014
Age	0.05	−	−	0.054	0.014	0.054	0.014	0.055	0.015	0.043	0.008
eGFR	−0.05	−	−	−0.054	0.031	−0.053	0.034	−0.056	0.036	−0.042	0.019
n=800		(n_e = 40)		(n_e = 60)		(n_e = 80)		(n_e = 100)			
Interc.	−2.00	−2.147	0.922	−2.105	0.975	−2.183	1.042	−2.192	1.231	−2.678	0.636
Male	−0.25	−0.247	0.249	−0.253	0.263	−0.259	0.290	−0.248	0.327	−0.199	0.175
BMI	−0.03	−0.031	0.017	−0.032	0.018	−0.031	0.020	−0.032	0.023	−0.026	0.012
Age	0.05	0.053	0.010	0.053	0.011	0.054	0.012	0.054	0.014	0.043	0.007
eGFR	−0.05	−0.052	0.025	−0.053	0.026	−0.053	0.028	−0.054	0.031	−0.043	0.016
n=1000		(n_e = 50)		(n_e = 75)		(n_e = 100)		(n_e = 125)			
Interc.	−2.00	−2.141	0.828	−2.104	0.865	−2.136	0.953	−2.123	1.041	−2.690	0.548
Male	−0.25	−0.244	0.231	−0.267	0.246	−0.261	0.268	−0.257	0.289	−0.202	0.149
BMI	−0.03	−0.031	0.014	−0.031	0.016	−0.032	0.018	−0.032	0.020	−0.025	0.010
Age	0.05	0.053	0.009	0.052	0.010	0.054	0.011	0.053	0.012	0.043	0.006
eGFR	−0.05	−0.052	0.022	−0.052	0.023	−0.055	0.026	−0.053	0.028	−0.042	0.015