Statistics
in Medicine

Research Article

Received 15 April 2016, Accepted 15 January 2017 Published online 13 February 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7237

Semiparametric profile likelihood estimation for continuous outcomes with excess zeros in a random-threshold damage-resistance model

John D. Rice[a*†] and Alex Tsodikov[b]

Continuous outcome data with a proportion of observations equal to zero (often referred to as semicontinuous data) arise frequently in biomedical studies. Typical approaches involve two-part models, with one part a logistic model for the probability of observing a zero and some parametric continuous distribution for modeling the positive part of the data. We propose a semiparametric model based on a biological system with competing damage manifestation and resistance processes. This allows us to derive a closed-form profile likelihood based on the retro-hazard function, leading to a flexible procedure for modeling continuous data with a point mass at zero. A simulation study is presented to examine the properties of the method in finite samples. We apply the method to a data set consisting of pulmonary capillary hemorrhage area in lab rats subjected to diagnostic ultrasound. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: damage threshold modeling; profile likelihood; retro-hazard; semicontinuous data; semiparametric methods

1. Introduction

There is often scientific interest in quantifying the effects of some external stress on a biological system. An area of application of particular relevance is carcinogenesis, for which dose–response models have been used extensively [1]. Often, the statistical question addressed is that of estimation of a threshold below which the probability of toxicity is zero [2]. In this paradigm, the outcome is typically binary: presence or absence of some adverse effect. Another perspective involves analysis of time to failure of a system, as in cumulative damage/shock models [3, 4]. These models, which are also germane to the multi-stage models of carcinogenesis [1], suppose that after some number of insults, the system breaks down.

In both cases, the outcome data take a particular form, binary for the dose–response models and time-to-event for the cumulative damage models. If, however, we are confronted with data containing outcomes that may be either exactly zero or positive (but not necessarily discrete), then another approach is needed. This is known as semicontinuous data, and a large body of research exists on modeling this sort of outcome [5–8]. Such data may occur in experimental setups in which test animals are subjected to external stress, and a measure of the damage caused by such pressures is obtained as the outcome. The motivation for this work is a data set consisting of 109 rats subjected to diagnostic ultrasound (DUS) [9]. From previous studies, it is known that DUS can induce pulmonary capillary hemorrhage (PCH) in rats. This is of clinical relevance for human patients because it demonstrates the potential for pulmonary injury following ultrasound examinations (e.g., examinations to diagnose conditions such as pulmonary edema, effusion, and embolism).

[a] *Department of Biostatistics and Computational Biology, University of Rochester, 265 Crittenden Blvd., Rochester, NY 14642, U.S.A.*
[b] *Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48104, U.S.A.*
[*] *Correspondence to: John D. Rice, Department of Biostatistics and Computational Biology, University of Rochester, 265 Crittenden Blvd., Rochester, NY 14642, U.S.A.*
[†] *E-mail: john_rice@urmc.rochester.edu*

The outcome of interest in Miller [9] was the measured area of PCH for each rat, in mm^2, obtained using photographs from a stereomicroscope with digital camera. The marginal mean of the outcome for all rats, including those with no damage, was 17.63 mm^2. When restricted to those rats with positive damage, the mean area was 26.69 mm^2. As 33.9% of rats exhibited no hemorrhagic damage, there is a definite point mass at zero.

The usual approach in analysis of semicontinuous data is based on the two-part model of Aitchison [5], which models the probability of an outcome exactly equal to zero and the distribution of the outcome given that it is greater than zero separately. Over the years, this model has been extended in various ways. Siegel [6] uses what amounts to a profile likelihood method to obtain maximum likelihood estimates for the parameters of a noncentral chi-squared distribution with zero degrees of freedom (a distribution that contains a point mass at zero). Foster and Bravington [8] propose a model based on an extension of the Tweedie generalized linear model, specifically the compound Poisson formulation, in which the outcome $y_i = \sum_{j=1}^{N_i} w_{ij}$, where N_i is Poisson and w_{ij} are gamma random variables. Polansky [10] provides a nonparametric method for estimation of the distribution function associated with a 'nonstandard mixture' model (meaning a model with probability mass at known discrete points) using a combination of an empirical distribution function and a kernel estimate of a distribution function, but does not address regression modeling. Zhou and Liang [11] present a method for the analysis of skewed data with excess zeros based on a two-part model, with the probability of a zero outcome being observed following a logistic model and the continuous positive outcome's conditional mean being modeled using a nonparametrically estimated smooth link function.

Our goal in this paper is to semiparametrically model data where the outcome represents some measure of damage to a biological system, in which two competing processes are at work. On the one hand, we have the damage manifestation process, which leads to expression of the damage in some observable form; on the other, we have the damage resistance process, which, up to a random, subject-specific threshold, may prevent the expression of the damage entirely, leading to an observed outcome of zero.

Figure 1 depicts schematically the relationship between applied stress and observed damage (left panel) and potential damage and observed damage (right panel) in two alternative models. The left panel represents a model for which the threshold is on the scale of some variable associated with the applied stress. The dose–response model of Crump [1] would be a special case of this: We would observe a binary outcome corresponding to whether or not the dose threshold was exceeded. The right panel shows the model corresponding to our proposed method, for which observed damage is equal to zero up to the threshold, from which point observed damage equals potential damage; in this case, the threshold is measured on the scale of damage itself.

The remainder of this paper is structured as follows: in Section 2, we lay out the details of our model for the competing damage and resistance processes; in Section 3, we propose an estimator for the parametric part of the model based on a profile likelihood defined using a function analogous to the hazard in time-to-event models; Section 4 presents simulation results; and Section 5 describes an application of the proposed method to a study of PCH in rats exposed to DUS.

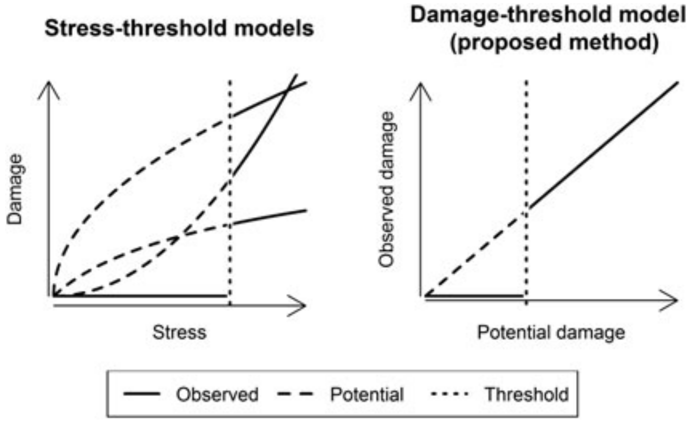

Figure 1. Schematic relationship between applied stress and observed damage (left panel) and potential damage and observed damage (right panel) in two alternative models.

2. The competitive damage/resistance model

Consider a regression model based on the Lehmann [12] family of alternatives, proposed in the context of nonparametric testing of the equality of distribution functions: For some outcome $X \in [0, \infty)$ and baseline CDF F, we have the model as written below in Equations (4) and (5)

$$P(X \leq x|\mathbf{z}) \equiv F_{\mathbf{z}}(x) = [F(x)]^{\exp\{\mathbf{z}'\boldsymbol{\beta}\}}. \tag{1}$$

We may also write this as a linear transformation model (LTM; see, e.g., [13]): Rearranging (1) yields $-\log\left[-\log F_{\mathbf{z}}(x)\right] = -\mathbf{z}'\boldsymbol{\beta} - \log\left[-\log F(x)\right]$. Because $X|\mathbf{z}$ has distribution function $F_{\mathbf{z}}$, simple calculations show that the random variable $-\log\left[-\log F_{\mathbf{z}}(X)\right]$ has density $\exp\{-x - e^{-x}\}$. Therefore,

$$g(X) = \mathbf{z}'\boldsymbol{\beta} + \epsilon, \tag{2}$$

is equivalent to (1), where $g(x) = -\log\left[-\log F(x)\right]$ is an unspecified increasing function and ϵ has density $\exp\{-x - e^{-x}\}$.

Our two-part model is based on an unknown baseline CDF F: if D_i is the random variable representing the damage expression and R_i the damage resistance capacity, then our observed data are

$$X_i = D_i\mathbf{1}(D_i > R_i), \tag{3}$$

where $\mathbf{1}(\cdot)$ is the indicator function, equal to 1 if $\cdot$ is true and zero otherwise. This model implies that we observe the damage D_i if and only if it exceeds the resistance capacity of the organism R_i; otherwise, we observe zero for the outcome. The probability model for D_i and R_i is

$$P(R_i \leq r) = [F(r)]^{\mu_i} \tag{4}$$

$$P(D_i \leq d) = [F(d)]^{\eta_i}. \tag{5}$$

We refer to this as the competitive damage/resistance (CDR) model.

Although not explicitly a dose–response model, our approach is similar to that of, for example, Cox [2] or Crump [1]. These authors, however, focused on estimation of the threshold, in contrast to our situation, where the threshold is random and dependent on the subject. One reference in which the threshold is random is Brockhoff and Muller [14], in which the authors make use of quasi-likelihood estimation in the analysis of repeated measures data. Dabrowska and Doksum [15] give an example of a linear transformation model for dose–response studies that has the form of (2), but where $\epsilon \sim N(0, 1)$ and $g(x) = \Phi^{-1}[F(x)]$, with $\Phi(\cdot)$ the standard normal CDF. It is not obvious for this model, however, how the regression coefficients should be interpreted. An advantage to our model (as will be discussed in Section 2.2) is a natural interpretation of the regression coefficients in terms of the probability of damage being greater in one group than another.

The model given by Equations (4) and (5) induces dependence between the observed damage X and the resistance capacity R by the shared baseline CDF. That is to say, the fact that the baseline CDF is common to both the damage and resistance variables is a natural way to incorporate the inherent association of the two processes within an organism. In effect, the baseline CDF is analogous to the variance components in a mixed effects model: These are common to all subjects, but also give rise to the random effect that induces the correlation within subjects.

2.1. *Biological motivation for the model*

The biological motivation for this model derives from the concept in cancer etiology of growth-promoting and growth-inhibitory signals [16]. On the one hand, proto-oncogenes encourage cell proliferation, while on the other, tumor suppressor genes actively inhibit such proliferation. The failure of these tumor suppressor genes can lead to uncontrolled growth and ultimately to the development of a cancerous tumor, but in the normal course of cell functioning, these genes prevent any cancer from manifesting. In the context of our model, we may view the unobserved R_i as representative of the action of growth-inhibitory signals; D_i, by contrast, corresponds to the action of growth-promoting signals. The event $D_i > R_i$ would then correspond to the point at which the tumor suppressor genes have failed and allowed a tumor to develop because of runaway cell proliferation.

Many studies in the field of cancer epidemiology simply address the question of what factors lead to the development of a tumor, which is analyzed using some binary response model such as logistic regression [17]. This approach produces estimates in the form of odds or risk ratios, but is unable to account for severity of disease in cases that do develop cancer. In contrast to studies seeking to determine the effect of various covariates on risk of developing cancer, other studies collect data on tumor size. However, these data are typically used to stratify an analysis of temporal trends in incidence [18, 19]. Essentially, this becomes an analysis of tumor size conditional on subjects having developed cancer. The model we propose here provides a unifying framework for the analysis of studies investigating the etiology of cancer, whereby both incidence and severity of disease may be assessed jointly.

2.2. *Specification of the model*

Using Equations (4) and (5), it may be shown that the CDF of the observed outcome $X = D\mathbf{1}(D > R)$ will be

$$\begin{aligned} P(X \le x) &= P\left[D\mathbf{1}(D > R) \le x\right] \\ &= \frac{\mu + \eta[F(x)]^{\eta+\mu}}{\eta + \mu}. \end{aligned} \tag{6}$$

Note that for $x = 0$, the marginal CDF is equal to $\mu/(\eta + \mu)$. This corresponds to a point mass at zero in the marginal distribution of damage. The intuition behind this is in the relative magnitudes of η and μ: The larger μ is relative to η, the greater the probability that no damage will be observed because of an increased resistance capacity.

The parameters η and μ will incorporate covariates $\mathbf{z}_i$ as follows:

$$\eta_i = e^{\mathbf{z}_i'\boldsymbol{\beta}_\eta}, \quad \mu_i = \frac{\theta_i}{1-\theta_i}\eta_i, \quad \theta_i = \frac{e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}}{1+e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}}, \tag{7}$$

where $\mathbf{z}_i$ is a $p \times 1$ vector. The parameter vectors $\boldsymbol{\beta}_\eta$ and $\boldsymbol{\beta}_\theta$ are also each $p \times 1$ vectors, but may have elements constrained to be zero if the corresponding covariate is not wanted in that part of the model. Model identifiability is possible because of the shared baseline CDF between the damage and resistance processes and the exclusion of an intercept term in η_i.

This parameterization follows by defining $\theta_i = \mu_i/(\eta_i + \mu_i)$ and then using a logistic link function to model θ_i. This allows for the interpretation of the intercept parameter β_0 as $\log P(D \le R)/P(D > R)$ for a subject with covariate vector of $\mathbf{0}$, while the remaining regression coefficients in this part of the model have the usual interpretation as log odds ratios for the event $\{D \le R\}$.

The regression coefficients in the continuous part of the model (i.e., $\boldsymbol{\beta}_\eta$) have a similar interpretation. An example will serve to illustrate this point: Suppose we have a single binary covariate Z, equal to 1 for treatment and zero for control, and associated regression coefficient β. Then $P(D \le d|Z = 1) = [F(d)]^{e^\beta}, P(D \le d|Z = 0) = F(d)$. Therefore, as in Lehmann [12], we have

$$P(D_{Z=0} < D_{Z=1}) = \int F\, d\left(F^{e^\beta}\right) = \int F e^\beta F^{e^\beta - 1}\, dF = \frac{e^\beta}{1+e^\beta}.$$

That is, β here is the logit of the probability that damage in a treated subject exceeds damage in a control subject. Thus, $\beta < 0$ implies a protective effect of treatment, while $\beta > 0$ implies a harmful effect.

The derivation of the profile likelihood that follows in Section 3 retains the original parameterization using only η and μ. This allows for simpler expressions throughout, but for implementation of the method, we will use the parameterization with η and θ.

3. Semiparametric estimation based on profile likelihood

3.1. *Left censoring and the retro-hazard function*

In order to obtain profile likelihood estimates for the regression parameters, by which we avoid having to specify the baseline CDF, we introduce the retro-hazard function, analogous to the hazard function in survival analysis. Specifically, consider a random variable T taking values on the interval $(0, \infty)$. The baseline cumulative hazard, $H(t)$, is defined as $H(t) = -\log S(t)$, where $S(t) = P(T > t)$ is the survival function [20]. This works well for right-censored data, but is an inconvenient way to formulate the model for left-censored data. This arises, in our model, when resistance capacity exceeds damage: the observed outcome will be censored on the left, because we will know only that damage was less than the resistance capacity.

Instead, we define

$$H^*(x) \equiv -\log F(x), \tag{8}$$

where $F(x) = P(X \leq x)$ is the CDF. Lagakos *et al*. [21] introduced a similar function in the context of the analysis of right-truncated survival data, which they refer to as a 'reverse-time hazard function'. Gross and Huber-Carol [22] further develop the ideas of the 'retro-hazard', but are also primarily interested in dealing with right-truncated data.

3.2. *Counting process formulation*

In the setting of left-censored data (Appendix S1), recall the counting process notation of survival analysis, where $N(t)$ denotes the counting process that takes value zero until the event occurs, then jumps to 1 (right continuous); and $Y(t)$, which takes value 1 while the subject is at risk of the event, and zero otherwise (left continuous by convention [20,p. 25]).

For our purposes, we will imagine a reversal of the time scale (similar to the approach of Lagakos *et al.* [21]) and define new processes

$$N^*(t) = 1 - N(t^-) \tag{9}$$

$$Y^*(t) = 1 - Y(t^+). \tag{10}$$

The process defined by (9) will be left continuous, while the process defined by (10) will be right continuous (somewhat different from the definitions given by Gross and Huber-Carol [22, sections 4.1–4.2]). Appendix A of the supplementary material presents our derivation of the nonparametric maximum likelihood estimator of the retro-hazard function in the general case for independent left-censoring. Recall that for our model, the censoring process, while not independent, results in all censored observations being equal to zero: thus, censored observations contain no information about the retro-hazard function.

3.3. *Profile likelihood for the competitive damage/resistance model*

Based on the marginal CDF (6) and defining $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}'_\theta, \boldsymbol{\beta}'_\eta)'$, we may now write the marginal likelihood for these data:

$$L(\boldsymbol{\beta}; H^*) = e^{\ell_1(\boldsymbol{\beta}) + \ell_2(\boldsymbol{\beta}; H^*)} \tag{11}$$

where

$$\ell_1(\boldsymbol{\beta}) = \sum_{i:X_i=0} \log \frac{\mu_i}{\eta_i + \mu_i}$$
$$\ell_2(\boldsymbol{\beta}; H^*) = \sum_{i:X_i>0} \log \left[-\eta_i e^{-(\eta_i+\mu_i)H^*(X_i)}\, dH^*(X_i) \right].$$

We show in Appendix B of the supplementary material that substitution of the nonparametric maximum likelihood estimator of H^* into (11) leads to a profile likelihood (over the infinite-dimensional H^*)

$$L(\boldsymbol{\beta}; \widehat{H^*}) \propto \prod_{i:X_i=0} \frac{\mu_i}{\eta_i + \mu_i} \prod_{i:X_i>0} \frac{\eta_i}{\sum_{j:0<X_j \leq X_i} (\eta_j + \mu_j)}. \tag{12}$$

This is analogous to a partial likelihood for $\boldsymbol{\beta}$, in that the right-hand side of (12) is proportional to the profile likelihood over H^* (see Breslow's contribution to the discussion of Cox [23, pp. 216–217]). This implies that we may base our inferences about these parameters on

$$\ell_{\text{pr}}(\boldsymbol{\beta}) = \sum_{i:X_i=0} \left[\log \mu_i - \log(\eta_i + \mu_i) \right] + \sum_{i:X_i>0} \left[\log \eta_i - \log \sum_{j:0<X_j \leq X_i} (\eta_j + \mu_j) \right]. \tag{13}$$

Using the parameterization given by (7), we may rewrite (13) in the form we will be using for estimation:

$$\ell_{\mathrm{pr}}(\boldsymbol{\beta}) = \sum_{i:X_i=0} \left[\beta_0 + \mathbf{z}_i'\boldsymbol{\beta}_\theta - \log\left(1 + e^{\beta_0+\mathbf{z}_i'\boldsymbol{\beta}_\theta}\right)\right] + \sum_{i:X_i>0} \left[\mathbf{z}_i'\boldsymbol{\beta}_\eta - \log \sum_{j:0<X_j\le X_i} e^{\mathbf{z}_j'\boldsymbol{\beta}_\eta}\left(1 + e^{\beta_0+\mathbf{z}_j'\boldsymbol{\beta}_\theta}\right)\right]. \quad (14)$$

The variance-covariance matrix of the parameter estimates may be estimated consistently by $\mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}})$; see Appendix C of the supplementary materials for the derivation of $\mathcal{I}(\boldsymbol{\beta})$. A proof of the asymptotic normality of a similar estimator is given by Gross and Huber-Carol [22]; only slight modifications of their proof are necessary for our estimator.

Cook and Farewell [24] give a similar example of the use of a partial likelihood in the analysis of left-censored data, but it is based on the conventional hazard rather than the retro-hazard. Our method can in fact be viewed as a generalization of theirs for a random left-censoring point that varies by subject and is related in a specific way to the outcome (in our case, by the proportionality of the retro-hazard functions). The authors do not, however, provide much in the way of interpretation of the model parameters. This is further elaborated on in Farewell [25], although the author simply changes the sign of the original outcomes in order to make use of Kaplan–Meier methodology for estimation of the CDF; the Lehmann family of alternatives is also mentioned [25, pp. 288–289].

4. Simulation studies

This section presents some simulation studies to examine the finite-sample properties of the proposed method. We simulated data under three scenarios: first, with the model correctly specified; second, with misspecification in the form of non-proportionality of the baseline retro-hazards; and third, with misspecification in the form of complete separation between the model for the probability of observing damage and the model for positive damage. The second and third scenarios were designed with the goal of checking the robustness of our proposed method to violations of the model assumptions.

In the second and third scenarios, we compare our proposed method with a standard, flexible semiparametric method of addressing the problem, similar to the model proposed by Zhou and Liang [11]. Specifically, for each data set, we fit a standard logistic model, with the outcome being $\mathbf{1}(X_i = 0)$; this fit corresponds to the θ part of our model. For the subset of observations greater than zero, we fit a semiparametric single-index model [26], implemented in the R package `np` [27] via the `npindex()` function. This method fits a model of the form $X = g(\mathbf{z}'\boldsymbol{\gamma}) + \epsilon$, where $\boldsymbol{\gamma}$ is a vector of parameters and g is an unknown function. Estimates $\tilde{g}$ and $\tilde{\boldsymbol{\gamma}}$ are obtained by minimizing a least-squares criterion (the bandwidth for estimation of g is chosen by cross validation).

Because this model is for the conditional mean, the parameter $\boldsymbol{\gamma}$ has no direct relationship to $\boldsymbol{\beta}_\eta$ in our model. This means that we are not able to compare our proposed method directly with this competitor on the basis of mean-square error of parameter estimates, for example. However, we may compare the methods indirectly using the fitted values.

To obtain fitted values for our method, we used the parameter estimates produced by our method and the Breslow-type estimator of the retro-hazard (derived in Appendix S2). Then Equation (6) gives the predicted CDF, which will be a step function; the jump sizes in this estimated CDF will correspond to an estimate of the density. If we denote this estimate as $\widehat{f}_i$, then the fitted value (i.e., expected damage conditional on covariates) for subject i will be $\sum_{j:X_j>0} X_j\widehat{f}_i(X_j)$. A similar method could be used to obtain estimates of the mean of the resistance variable for each subject, by replacing $\widehat{f}_i$ with an estimate of the conditional density of R_i given the observed value of X_i.

For the comparison method, we predict the expected outcome conditional on covariates as $\left[1 + e^{\mathbf{z}'\tilde{\boldsymbol{\beta}}_\theta}\right]^{-1} \tilde{g}\left(\mathbf{z}'\tilde{\boldsymbol{\gamma}}\right)$. We refer to this method as the logistic/semiparametric single-index model.

4.1. Correct specification

We simulated 1000 data sets for each of three sample sizes and three intercept values; the intercept was varied in order to produce different proportions of zero observations in the response. A baseline retro-hazard of $H^*(t) = -\log\left(1 - e^{-t/10}\right)$ was used, corresponding to an exponential model. We included two covariates, both of which were included in each part of the model: $Z_1 \sim N(0,1)$ and $Z_2 \sim B(1/2)$. Then we have for each subject

$$\theta = \frac{e^{\beta_0 + 2Z_1 - Z_2}}{1 + e^{\beta_0 + 2Z_1 - Z_2}} \tag{15}$$

$$\eta = e^{-Z_1 + 2Z_2} \tag{16}$$

$$\mu = \frac{\theta}{1-\theta}\eta. \tag{17}$$

Thus, for the true parameters, we have $\beta_{\theta 1} = 2$, $\beta_{\theta 2} = -1$, $\beta_{\eta 1} = -1$, and $\beta_{\eta 2} = 2$. For these simulations, the intercept β_0 in the θ part of the model was allowed to take values -2, 0, and 2, corresponding to, respectively, approximately 18%, 43%, and 71% of observations equal to zero.

The full results of the simulation study for the scenario without misspecification are given in Appendix D of the supplementary materials, but broadly, the method shows quick reduction in bias of parameter estimates with increases in sample size as well as good agreement between empirical and estimated standard errors.

4.2. *Non-proportional retro-hazards*

In order to address the issue of robustness, we generated data according to the following model:

$$P(R_i \leq r) = e^{-\mu_i H^*(r)} \tag{18}$$

$$P(D_i \leq d) = e^{-\eta_i [H^*(d)]^{\alpha}}. \tag{19}$$

The function H^* was the same as for the simulations in the first scenario, while we varied α between 0.7 and 1.3 to determine the effect of varying degrees of model misspecification (Figure 4 in Appendix D of the supplementary materials). Equations 15, (16), and (17) were used as in the first scenario to generate the true values of μ and η; covariate distributions were likewise the same.

Full results for this simulation setting are given in Appendix D of the supplementary materials. Overall, however, the effect of this kind of misspecification seems to be quite limited, both on our proposed method as well as the standard method. The results do indicate that our method outperforms the standard method uniformly and by a large margin, generally 40–50% regardless of other model parameters.

4.3. *Unlinked models*

As another check on the robustness of our method, we generated data assuming that the probability of observing damage is not linked with the distribution of positive damage. Specifically, under this scenario, the model for the probability of observing any damage was unchanged and is given by Equation (15). For the distribution of the outcome given it is positive, we used a single-index model:

$$X_i | X_i > 0 = g\left(\gamma_0 + Z_1 - Z_2\right) + \epsilon, \tag{20}$$

where $\epsilon \sim N(0, 5^2)$ for all settings in this scenario; covariate distributions were the same as in the previous scenarios. Because for some possible g functions the outcome could be negative after adding the random error term, the continuous outcome generated was truncated at zero. This means that the LSSI model is not precisely correctly specified, as there are some observations that should have been positive according to the logistic part of the model, but appear as zero because of this truncation.

Two general families were used for the unknown function g, in an attempt to obtain an idea of the behavior of the estimators over a range of possible shapes for this function. In the first setting, $g(u) = 6|u|^{\omega}$; ω was varied between 1 and 3. We set $\gamma_0 = g^{-1}(100) = (100/6)^{1/\omega}$, which minimizes the effect of ω on the mean outcome, kept approximately constant. This allows us to focus on the effect of the shape of the unknown function g on the prediction errors produced by our proposed method and the competing method. In the second setting, $g(u) = 100 + \frac{80}{\pi}\tan^{-1}\left(\frac{u}{\sigma}\right)$ and $\gamma_0 = 0$; σ was varied between 1 and 4. No intercept is necessary here because the location of the function is fixed as we vary σ. See Figure 2 for plots of these functions.

We are now less interested in comparisons of predictive error across different values of ω or σ, as there is no 'reference' level corresponding to correct model specification. Therefore, we use the usual

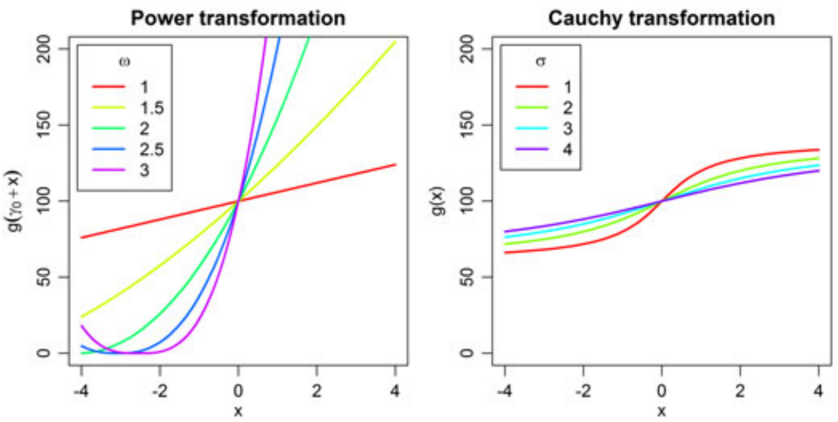

Figure 2. Plots of the unknown function g for the simulations in which the logistic/semiparametric single-index model is the data-generating mechanism. The x-axis of each plot shows the value of the linear predictor that corresponds to the indicated conditional mean of the response variable (given that the response exceeds zero).

Table I. Simulation results under misspecified model, power transformation as unknown function of the index: predictive errors, $n = 500$.

β_0	ω	$\mathbb{E}X$	LSSIM	CDRM	Ratio
−2	1	78.5	2.960	3.312	1.119
	1.5	69.1	3.326	3.407	1.024
	2	59.3	4.647	4.304	0.926
	2.5	54.1	17.147	13.934	0.813
	3	54.0	38.750	34.723	0.896
−1	1	67.3	3.005	3.550	1.181
	1.5	57.6	3.046	3.181	1.044
	2	47.0	3.836	3.680	0.959
	2.5	40.4	11.838	8.897	0.752
	3	38.0	25.565	20.240	0.792
0	1	53.7	2.930	3.379	1.153
	1.5	44.3	2.898	3.012	1.039
	2	34.0	3.151	3.056	0.970
	2.5	27.2	7.536	5.276	0.700
	3	23.7	14.612	9.969	0.682

This table shows the $\sqrt{\text{MSEP}_2}$ for both the standard method (LSSIM) and our proposed method (CDRM); the final column is a measure of relative efficiency, calculated as the ratio of the $\sqrt{\text{MSEP}_2}$ of the CDRM method to that of the LSSIM method. This is averaged over 1000 simulated data sets at each distinct combination of intercept value β_0 and misspecification parameter ω. Also, displayed in this table is the average outcome across all subjects and simulated data sets, intended to give an idea of the relative size of the $\sqrt{\text{MSEP}_2}$ values (which are not normalized as they are for MSEP_1). The intercept parameter β_0 was allowed to take values −2, −1, and 0 (shown in the first column), corresponding to, respectively, approximately 18%, 29%, and 43% of observations equal to zero. $\sqrt{\text{MSEP}_2}$, root mean-square error of the predictions; CDRM, competitive damage/resistance model; LSSIM, logistic/semiparametric single-index model.

mean-square error criterion to compare our proposed method and the standard method within each value of ω:

$$\text{MSEP}_2 = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{X}_i - \mathbb{E}(X_i|\mathbf{z}_i)\right]^2 . \tag{21}$$

Average values of this quantity across 1000 simulated data sets are shown Tables I and 5.

Table I shows that the standard method slightly outperforms our proposed method for smaller values of the exponent ω, but becomes relatively less efficient as ω increases further. However, we must keep in mind that there is a dramatic increase in predictive error for both the proposed method and the standard method as ω increases, likely due to increasing potential for extremely large values of the outcome (see the left panel of Figure 2). This is even more pronounced when we consider the fact that $\mathbb{E}X$ is decreasing with increasing ω. We may conclude that for shallower functions g, the standard method performs better than our proposed method, albeit only by approximately 10%. For steeper g functions, by contrast, neither method performs well, but our proposed method is relatively more robust than the standard method, with possibly 20–30% increases in efficiency depending on ω.

Results for the Cauchy transformation are given in Appendix D of the supplementary materials. They are generally more favorable for the standard method than our proposed method, but as the curve becomes increasingly linear (i.e., with increasing σ), our proposed method becomes competitive.

5. Rat pulmonary capillary hemorrhage data analysis

To evaluate the CDR model in practice, we applied it to the data of Miller [9]. The rats in this study were evaluated at various combinations of ultrasonic frequencies (1.5, 4.5, 7.6, and 12 MHz) and peak rarefactional pressure amplitude (referred to hereafter simply as amplitude). There was especial interest in thresholds for PCH expressed in terms of the amplitude, which makes this data particularly suitable for our method, as it explicitly models the probability of exceeding subject-specific damage thresholds as a function of covariates.

The results of applying our method to these data are displayed in Table II and Figure 3. Two covariates (along with possible interactions) were considered in this analysis: frequency, which takes only four possible values in this data set, and amplitude. It was found that treating frequency as a categorical rather than a continuous variable in the η part of the model provided a substantial improvement in fit to the data without sacrificing too much in terms of efficiency (as measured by AIC).

In Table II, we see that the coefficient estimates for amplitude are large in magnitude but opposite in sign in the two parts of the model: This is sensible, recalling that we are modeling the probability of damage not being manifested with the θ part of the model and that the η part of the model essentially scales the CDF of observed positive damage, so that more positive coefficient estimates indicate increased damage. Specifically, with a coefficient estimate of 8.290, the probability that damage in a rat exposed to an additional unit of amplitude exceeds damage in a 'control' rat is essentially 1. The overall interpretation is that larger amplitudes lead both to increased probability of exceeding the resistance threshold as well as to increased damage once the threshold has been exceeded.

The interpretation of the effect of frequency on PCH area is somewhat more complicated, both because it is treated as continuous in the logistic (θ) part of the model and categorical in the positive (η) part, as well as because of the inclusion of an interaction term in the positive part. However, we can say that increasing frequency leads to decreasing probability of exceeding the resistance threshold, because the coefficient estimate for this covariate in the θ part of the model is positive. Although the coefficient

Table II. Parameter estimates for the rat PCH data.

Model	Covariate	Estimates	SE	*p*-value
Logistic	(Intercept)	6.064	1.621	0.0002
	Amplitude	−7.696	1.658	0.0000
	Frequency	0.356	0.101	0.0004
Continuous	Amplitude	8.290	1.009	0.0000
	Frequency (1.5 MHz: ref.)	—	—	—
	Frequency (4.5 MHz)	2.632	1.271	0.0384
	Frequency (7.6 MHz)	3.143	1.461	0.0314
	Frequency (12 MHz)	2.230	3.209	0.4871
	Amplitude × frequency (4.5 MHz)	−3.907	0.836	0.0000
	Amplitude × frequency (7.6 MHz)	−4.420	1.020	0.0000
	Amplitude × frequency (12 MHz)	−4.257	2.096	0.0423

The final model was chosen on the basis of visual fit to the observed data (Figure 3). The column labeled 'model' denotes the part of the model to which the covariates refer: either the logistic model for the probability of not exceeding the resistance threshold or the continuous model for the positive responses (i.e., observed damage > 0).
PCH, pulmonary capillary hemorrhage; SE, standard error.

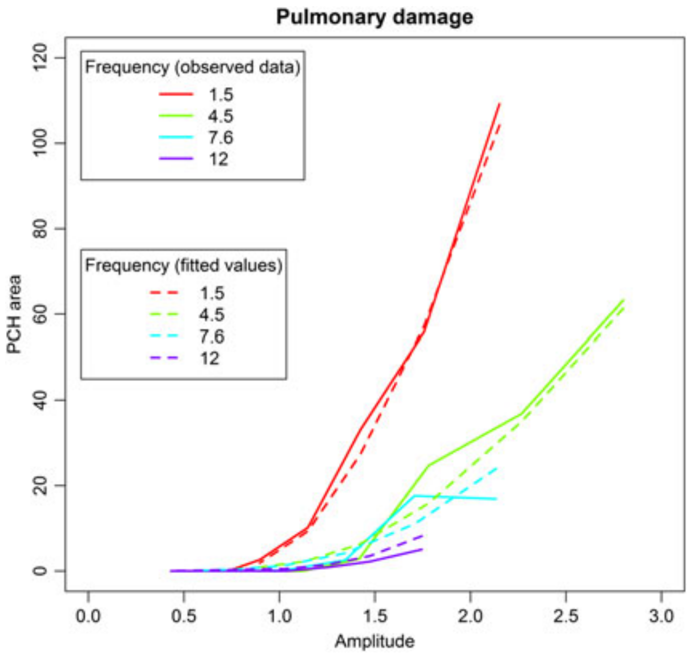

Figure 3. Observed and fitted values for the rat pulmonary capillary hemorrhage (PCH) data. Curves labeled 'observed data' are conditional means for the amplitude and frequency values depicted. Curves labeled 'fitted values' were obtained by fitting the competitive damage/resistance model using the partial likelihood technique outlined in Section 3; the retro-hazard was then obtained using the estimation procedure given in Appendix A of the supplementary materials; finally, these elements were combined to give an estimate of the conditional density function, which was then used along with the observed damage values to obtain expectations numerically.

estimates for the frequency terms alone are all positive in the η part of the model, which would indicate an association of increasing frequency with increasing damage (given exceedance of the threshold), note that the interaction terms all have greater magnitude and negative sign. Therefore, as long as amplitude is greater than zero, the net effect of frequency will be negative, which coincides with what intuition suggests given the positive sign of this coefficient in the logistic part of the model.

Turning now to Figure 3, we may observe the visual fit of the model to the data, obtained using the procedure outlined earlier in Section 4. It is clear from this figure that the model provides a good fit to the data for each frequency and across amplitudes. There may be slight overestimation in the fitted values for the highest frequency, but overall we see precisely the patterns in the observed data, with smooth curves rising from zero (no damage observed) at the lowest amplitudes.

6. Discussion

In this paper, we have proposed a model for competitive damage and resistance processes in a biological system, motivated by a data set consisting of test animals subjected to an external force expected to lead to injury. Our model, using the retro-hazard function first proposed by Lagakos *et al*. [21] and later elaborated upon by Gross and Huber-Carol [22], leads to an estimation procedure based on a closed-form profile likelihood. This procedure is fast, efficient, and does not require any distributional assumptions on the observed damage outcome. Parameter interpretation is provided with reference to the probability of damage exceeding repair capacity (for the logistic part of the model) and to the probability of damage in one group exceeding damage in another group (for the continuous part of the model).

The assumption of a common baseline retro-hazard for both the damage and resistance systems could be questioned in a particular application. However, the inclusion of covariates in each part of the model, which may of course take the same or opposite signs, seems to allow sufficient flexibility in terms of the effect of a particular factor on the observed outcome. There are always trade-offs between fidelity to biological reality on the one hand and statistical or mathematical convenience on the other. Our modeling approach is motivated by the former, but makes the necessary concessions to the latter in order for the model to be identifiable. In our model, the rationale for the damage and resistance variables sharing the same baseline CDF is that the stressor should provoke similar but opposite reactions from these systems.

In the context of the rat PCH data, this stressor is the DUS: This is applied to each organism and triggers two biological reactions, damage, and resistance. Although diametrically opposed to one another, both are responding to the same stressor. We may also imagine, in the more general setting, that such a stressor could be an environmental exposure in a study of the etiology of cancer, for example.

Future research may examine the possibility of relaxing this assumption via inclusion of shared variables, similar to frailties in survival analysis, between the two parts of the model. Another possible direction for further study is explicit incorporation of a dose–response relationship in the model, as is depicted in the left panel of Figure 1 (with dose corresponding to stress). Currently, our approach implicitly assumes that the outcome is the response to some applied dose; however, a dynamic model for variable dose over time could be quite interesting.

Acknowledgements

The authors thank Douglas Miller for providing the rat PCH data used in this paper. This research was partially supported by the grant 1U01CA199338 (CISNET) from the National Cancer Institute; partial support was also provided by the grant 5-R01-HL-116434-02 from the NIH.

References

1. Crump KS. Dose response problems in carcinogenesis. *Biometrics* 1979; **35**:157–167.
2. Cox C. Threshold dose-response models in toxicology. *Biometrics* 1987; **43**:511–523.
3. Ebrahimi N. Stochastic properties of a cumulative damage threshold crossing model. *Journal of Applied Probability* 1999; **36**:720–732.
4. Esary JD, Marshall AW. Shock models and wear processes. *The Annals of Probability* 1973; **1**:627–649.
5. Aitchison J. On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association* 1955; **50**:901–908.
6. Siegel AF. Modelling data containing exact zeroes using zero degrees of freedom. *Journal of the Royal Statistical Society, Series B (Methodological)* 1985; **47**:267–271.
7. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**: 1–14.
8. Foster SD, Bravington MV. A Poisson-gamma model for analysis of ecological non-negative continuous data. *Environmental and Ecological Statistics* 2013; **20**:533–552.
9. Miller DL. Induction of pulmonary hemorrhage in rats during diagnostic ultrasound. *Ultrasound in Medicine and Biology* 2012; **38**:1476–1482.
10. Polansky AM. Nonparametric estimation of distribution functions of nonstandard mixtures. *Communications in Statistics—Theory and Methods* 2005; **34**:1711–1724.
11. Zhou XH, Liang H. Semi-parametric single-index two-part regression models. *Computational Statistics and Data Analysis* 2006; **50**:1378–1390.
12. Lehmann EL. The power of rank tests. *The Annals of Mathematical Statistics* 1953; **24**:23–43.
13. Cheng S, Wei L, Ying Z. Analysis of transformation models with censored data. *Biometrika* 1995; **82**:835–845.
14. Brockhoff PM, Muller HG. Random effect threshold models for dose–response relationships with repeated measurements. *Journal of the Royal Statistical Society, Series B (Methodological)* 1997; **59**:431–446.
15. Dabrowska DM, Doksum KA. Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics* 1988; **15**:1–23.
16. Weinberg RA. Tumor suppressor genes. *Science* 1991; **254**:1138–1146.
17. Boffetta P, Nyberg F. Contribution of environmental factors to cancer risk. *British Medical Bulletin* 2003; **68**:71–94.
18. Enewold L, Zhu K, Ron E, Marrogi AJ, Stojadinovic A, Peoples GE, Devesa SS. Rising thyroid cancer incidence in the United States by demographic and tumor characteristics, 1980–2005. *Cancer Epidemiology, Biomarkers & Prevention* 2009; **18**:784–791.
19. Chow WH, Dong LM, Devesa SS. Epidemiology and risk factors for kidney cancer. *Nature Reviews Urology* 2010; **7**: 245–257.
20. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data* Second edn. Wiley: Hoboken, New Jersey, 2002.
21. Lagakos SW, Barraj LM, De Gruttola V. Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika* 1988; **75**:515–523.
22. Gross ST, Huber-Carol C. Regression models for truncated survival data. *Scandinavian Journal of Statistics* 1992; **19**: 193–213.
23. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)* 1972; **34**:187–220.
24. Cook RJ, Farewell VT. The utility of mixed-form likelihoods. *Biometrics* 1999; **55**:284–288.
25. Farewell VT. Some comments on analysis techniques for censored water quality data. *Environmental Monitoring and Assessment* 1989; **12**:285–294.
26. Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 1993; **58**:71–120.
27. Hayfield T, Racine JS. Nonparametric econometrics: the np package. *Journal of Statistical Software* 2008; **27**.

28. Hu C, Tsodikov A. Joint modeling approach for semicompeting risks data with missing nonterminal event status. *Lifetime Data Analysis* 2014; **20**:563–583.

Supporting Information

Additional supporting information (Appendices A, B, C, and D) may be found online in the supporting information tab for this article.