Published in final edited form as:

Stat Med. 2015 July 30; 34(17): 2514–2527. doi:10.1002/sim.6503.

Computationally Simple Analysis of Matched, Outcome-Based Studies of Ordinal Disease States

Rebecca A. Betensky^{1,2,*}, Jackie Szymonifka², Eudocia Q. Lee³, Catherine L. Nutt⁴, and Tracy T. Batchelor⁵

¹Department of Biostatistics, Harvard School of Public Health, USA

²Biostatistics Center, Massachusetts General Hospital, USA

³Center for Neuro-Oncology, Dana-Farber/Brigham and Women's Cancer Center, USA

⁴Pathology Service, Massachusetts General Hospital, USA

⁵Brain Tumor Center, Massachusetts General Hospital, USA

Abstract

Outcome-based sampling is an efficient study design for rare conditions, such as glioblastoma. It is often used in conjunction with matching, for increased efficiency and to potentially avoid bias due to confounding. A study was conducted at Massachusetts General Hospital (MGH) that involved retrospective sampling of glioblastoma patients with respect to multiple ordered disease states, as defined by three categories of overall survival time. To analyze such studies, we posit an adjacent categories logit model and exploit its allowance for prospective analysis of a retrospectively sampled study and its advantageous removal of set and level specific nuisance parameters through conditioning on sufficient statistics. This framework allows for any sampling design and is not limited to one level of disease within each set, as is the development of Mukherjee *et al.* [1]. We describe how this ordinal conditional model can be fit using standard conditional logistic regression procedures. We consider an alternative pseudo-likelihood approach that potentially offers robustness under partial model misspecification at the expense of slight loss of efficiency under correct model specification for small sample sizes. We apply our methods to the MGH glioblastoma study.

Keywords

Adjacent categories model; Conditional logistic regression; Glioblastoma; Outcome-based sampling; Pseudo-likelihood

1. Introduction

Glioblastoma (GBM) is the most common malignant primary brain tumor and accounts for 17% of all brain tumors diagnosed in the United States [2]. Despite standard treatment with surgery, radiation (RT), and temozolomide (TMZ), median overall survival is only 14.6

*Correspondence to: Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115. betensky@hsph.harvard.edu.

months [3]. Nonetheless, there is a small subset of GBM patients (3–5%) who live longer than 3 years and are described in the literature as long-term survivors (LTS) ([4], [5], [6]). Young age, good Karnofsky Performance Status (KPS), multimodality therapy, and O⁶-methylguanine-DNA methyltransferase (MGMT) promoter methylation status are associated with long-term survival ([4], [5], [6], [7], [8], [9]). Few studies have examined GBM patients living 6 months or less, termed short-term survivors (STS). In a study of 12 short-term survivors, Barbus and colleagues [8] found no isocitrate dehydrogenase 1 (IDH1) mutations.

Hypoxia-induced necrosis and microvascular proliferation are pathologic hallmarks in GBM [10]. Increased levels of tumor hypoxia are associated with worse outcomes. In addition, hypoxia helps maintain glioma stem cells, which are inherently resistant to treatment with radiotherapy and chemotherapy [11]. Anti-VEGF agents, which are now frequently used to treat patients with recurrent GBM, may transiently normalize abnormal tumor vasculature enabling more efficient oxygen and drug delivery ([12], [13]). This led us to study the roles of hypoxia and vasculature in long-term and short-term survival in glioblastoma [14]. We hypothesized that increased hypoxia, increased microvascular proliferation, and increased microvessel density are associated with shorter survival. We also examined the roles of clinical factors, MGMT methylation status, and IDH1 mutation status.

Outcome-based sampling is an efficient study design for rare conditions [15]. It is often used in conjunction with matching, for increased efficiency and to potentially avoid bias due to confounding. The one-to-one matched case-control design in which a subject with “disease” (i.e., the case) is matched to a subject without “disease” (i.e., the control) is widely used. The analysis of such studies should account for the matched design; otherwise, invalid results may arise [16]. With two outcomes, conditional logistic regression is typically used, and is implemented in statistical software packages. The logistic model permits prospective modeling of the disease outcome in spite of the retrospective, outcome-based sampling for the purpose of inference about the association parameters [17].

A more general study design involves retrospective sampling with respect to multiple ordered disease states. For example, the study of genetic and histopathological correlates of glioblastoma survival conducted at MGH involved retrospective sampling of the tumor registry and tumor bank, and the investigators constructed matched sets, which ideally included one short-term survivor (STS), three standard-term survivors (survival between six months and three years), and one long-term survivor (LTS). The subjects were matched on the basis of age and Karnofsky Performance Score (KPS). There were 145 subjects who comprised 30 matched sets; 22 of these were complete, while eight deviated from the targeted numbers.

One obvious approach for analysis of these data is to fit two separate conditional logistic regression models [18]: one with short and standard term survivors and one with long and standard term survivors. While simple to implement, this approach is less efficient than fitting a joint model when there is a common association among adjacent levels of disease [19]. Others have extended the conditional logistic regression methodology to polychotomous disease categories (e.g., [20], [21], [22], [23],[24], [25]). In the case of a

single binary covariate, Liang and Stewart [20] demonstrated that this joint modeling approach is more efficient than conducting standard conditional logistic regression on subgroups. A drawback of this approach is that it does not acknowledge the ordering among disease states. Mukherjee *et al.* [1] considered a matched study design based on ordinal disease states and evaluated various modeling approaches that acknowledged the ordering. They noted that the commonly used proportional odds model for ordinal logistic regression does not allow for elimination of the nuisance parameters due to the matched sets through conditioning. Instead, they showed that an adjacent categories logit model for these data does facilitate this, and furthermore, can be translated into a multinomial logit model, through which prospective analysis of the ordinal disease outcome is justified. Importantly, they also showed that with proper construction of the data, standard conditional logistic regression software could be used to fit the ordinal adjacent categories model. A popular competing model to the adjacent categories model is the cumulative logits model with proportional odds structure. It has been noted that these models actually provide similar results, as both imply orderings of the conditional distribution of the outcome across predictor values [26]. However, as we are interested in the three categories of survival for brain tumors, and not in the underlying survival times, the adjacent categories model is more appropriate than the cumulative logits model, which implies an underlying continuous scale. Thus, for reasons of efficiency and computational simplicity in the presence of matching and retrospective sampling, and interpretation, we pursue the extension of the adjacent categories model to our setting.

While the adjacent categories modeling approach [1] is appealing due to its natural accommodation of retrospective, matched data, it is limited to matched sets that contain a single “case” who resides in one of the ordinal disease states, and multiple controls (considered to reside in a single, fixed state). Across the matched sets, the cases may vary in their ordinal levels of disease state. Thus, this approach is not directly applicable to the glioblastoma study, in which *each* matched set contains three severity levels of disease. In this paper, we extend the results of Mukherjee *et al.* [1] to the case of adjacent categories logit modeling for studies with general outcome-based matching based on multiple ordinal disease states. We do not require designation of a “control” state, though typically there would be a natural reference state. In the glioblastoma study, each set contains three levels of disease, with the standard survivor level considered as the reference state. We also do not require a fixed number of any level of disease to be contained within a given matched set. Finally, we do not require complete sets for analysis; the likelihood based approach can handle any kind of partial set. These extensions are useful in practice as their flexibility will enable the construction of larger retrospective studies that are based on ordinal disease categories and involve matching on potential confounders. As in the simpler setting of Mukherjee *et al.* [1], we illustrate how the data can be transformed to enable use of standard conditional logistic regression software. We present the pseudo-likelihood approach that separately models short-term versus standard survivors, and long-term versus standard survivors, and compare it to the full likelihood approach. We also consider a full likelihood approach that assumes the effect of covariates to be equivalent between adjacent levels of disease. We identify the advantages and disadvantages of each of these approaches. Finally,

we use the proposed modeling and estimation approach to analyze the MGH glioblastoma study.

2. Notation and models

Let n denote the number of matched sets and n_i denote the number of subjects in set i . Let $Y_{ij} = l$, with $l \in \{1, \dots, c\}$ denote the ordered disease level for subject $j = 1, \dots, n_i$ in matched set $i = 1, \dots, n$; for the glioblastoma study, level 1 denotes short term (less than six months) survival, level 2 denotes standard (six months to three years) survival, and level 3 denotes long term (more than three years) survival. Matched sets contain prespecified numbers of subjects at each disease level; let d_{il} denote the number of subjects at disease level l in matched set i . For 22 of the 30 matched sets in the glioblastoma study, $d_{i1} = d_{i3} = 1$ and $d_{i2} = 3$. This study design greatly generalizes that considered by Mukherjee *et al.* [1], which restricted each matched set to a single subject with disease (at one of several possible ordered levels) and the remaining subjects at the defined control level. Our framework is completely general in its allowance of any pre-defined mixture of disease levels within a set. We assume that one level of disease represents the natural reference level; this may or may not be the lowest or highest level. We denote this level by $l = r$. In the glioblastoma study, the natural reference level is level 2, that of the standard survivors. Let X_{ij} denote the covariate vector for subject j in matched set i . A general, non-proportional odds formulation of the adjacent categories model [26] is given by

$$\log \frac{P(Y_{ij}=l|X_{ij})}{P(Y_{ij}=l+1|X_{ij})} = \gamma_i + \alpha_l + \beta_l X_{ij}, \quad i=1, \dots, n, \quad j=1, \dots, n_i, \quad l=1, \dots, c-1. \quad (1)$$

With respect to the reference level, r , it follows that

$$\log \frac{P(Y_{ij}=l|X_{ij})}{P(Y_{ij}=r|X_{ij})} = \gamma_{il}^* + \beta_l^* X_{ij}, \quad i=1, \dots, n, \quad j=1, \dots, n_i, \quad l=1, \dots, c.$$

where

$$\gamma_{il}^* = (r-l)\gamma_i + \text{sgn}(r-l) \sum_{k=\min(r,l)}^{\max(r-1,l-1)} \alpha_k$$

and

$$\beta_l^* = \text{sgn}(r-l) \sum_{k=\min(r,l)}^{\max(r-1,l-1)} \beta_k, \quad (2)$$

where $\text{sgn}(u) = 1$ if $u > 0$, $\text{sgn}(u) = -1$ if $u < 0$, $\text{sgn}(u) = 0$ if $u = 0$. Note that if the simple adjacent categories model holds, in which $\beta_l \equiv \beta$ for all l , it follows that $\beta_l^* = (r-l)\beta$. An

alternative model that is intermediate between the simple adjacent categories model and the general adjacent categories model is the *stereotype* model, in which the β_l 's are not assumed equal, but rather are assumed parallel: $\beta_l \equiv \phi_l \beta$ [27]. Under this model, it follows that

$\beta_l^* = \beta \sum_{k=\min(r,l)}^{\max(r-1,l-1)} \phi_k$. This model seems particularly useful for settings in which there are several ordered categories, rather than the three that are present in our motivating example. Under the general adjacent categories model formulation and with the transformed coefficients, it follows that

$$P(Y_{ij}=l|X_{ij}) = \frac{\exp(\gamma_{il}^* + \beta_l^* X_{ij})}{1 + \sum_{k \neq r} \exp(\gamma_{ik}^* + \beta_k^* X_{ij})}, l=1, \dots, c, l \neq r,$$

and

$$P(Y_{ij}=r|X_{ij}) = \frac{1}{1 + \sum_{k \neq r} \exp(\gamma_{ik}^* + \beta_k^* X_{ij})}.$$

We proceed to eliminate the set-specific nuisance parameters, γ_{il}^* , $l=1, \dots, c, l \neq r$ through the usual technique of conditioning on their associated sufficient statistics: d_{il} , $l=1, \dots, l, l \neq r$, i.e., the number of subjects in set i at each disease level. Letting $(y_{i1}, \dots, y_{in_i})$ denote the observed disease levels for the n_i subjects in matched set i , the contribution to the conditional likelihood from set i is given by:

$$\begin{aligned} L_i &= P(Y_{ij}=y_{ij}, j=1, \dots, n_i | X_{ij}, j=1, \dots, n_i, d_{i1}, \dots, d_{ic}) \\ &= \frac{\prod_{j=1}^{n_i} P(Y_{ij}=y_{ij} | X_{ij})}{\sum_p \prod_{j=1}^{n_i} P(Y_{ij}=y_{ij}^p | X_{ij})} \\ &= \frac{\exp(\sum_{l=1}^c \gamma_{il}^* d_{il} + \sum_{l=1}^c \beta_l^* \sum_{j=1}^{n_i} X_{ij} I(y_{ij}=l))}{\sum_p \exp(\sum_{l=1}^c \gamma_{il}^* d_{il} + \sum_{l=1}^c \beta_l^* \sum_{j=1}^{n_i} X_{ij} I(y_{ij}^p=l))} \\ &= \frac{\exp(\sum_{l=1}^c \beta_l^* \sum_{j=1}^{n_i} X_{ij} I(y_{ij}=l))}{\sum_p \exp(\sum_{l=1}^c \beta_l^* \sum_{j=1}^{n_i} X_{ij} I(y_{ij}^p=l))}, \end{aligned} \quad (3)$$

where $I(E)$ is the indicator function for the event E and p indexes the set of all possible permutations, $\{(y_{i1}^p, \dots, y_{in_i}^p), p=1, \dots, P\}$, of the observed $(y_{i1}, \dots, y_{in_i})$ that preserve the observed values of the sufficient statistics, $(d_{i1}, \dots, d_{ic})$, i.e., the numbers of set i at each disease level, where

$$P = \binom{n_i}{d_{i1} \dots d_{ic}}.$$

This formulation of the likelihood is equivalent to that from a standard conditional logistic regression, with each matched set consisting of one “case” and $P-1$ “controls,” and with l covariates that are given by:

$$X_{i,case}^l = \sum_{j=1}^{n_i} X_{ij} I(y_{ij}=l)$$

for the set's case, and by

$$X_{i,p}^l = \sum_{j=1}^{n_i} X_{ij} I(y_{ij}^p=l),$$

for each of the set's $P - 1$ "controls" and one "case." We note that this likelihood formulation is complete general and could be applied to incomplete matched sets that do not contain any subjects at some disease levels, including at the reference category level. That is, the model is fully specified and does not require that each matched set contain subjects from this reference level (or from any other level). Of course, each matched set must contain at least two different disease levels in order to contribute non-trivially to the conditional likelihood. This conditional logistic regression formulation may be used to estimate the β_l^* parameters, which then may be transformed into the β_l parameters of interest through inversion of the linear relationship indicated in (2). As noted by Mukherjee *et al.* [1], due to the induced multinomial logit structure, the prospective-retrospective equivalence for estimating β_l holds with this model.

Alternatively, if the simple adjacent categories model holds, in which $\beta_l \equiv \beta$ for all l , the conditional likelihood contribution from set i is given by

$$L_i = \frac{\exp(\beta \sum_l (r-l) \sum_{j:y_{ij}=l} X_{ij})}{\sum_p \exp(\beta \sum_l (r-l) \sum_{j:y_{ij}^p=l} X_{ij})}.$$

To fit this as a simple conditional logistic regression model in which each matched set contains one case and multiple controls, the covariates X_{ij} are transformed via linear combinations of the covariates from set i , so that the covariate associated with the "case" in set i is

$$X_{i,case} = \sum_{l=1}^c (r-l) \sum_{j:y_{ij}=l} X_{ij},$$

and the covariates associated with each of the $p = 1, \dots, n_i \times (n_i - 1)/2 - 1$ "controls" and one "case" in set i are based on permuting the observed disease levels into $(y_{i1}^p, \dots, y_{in_i}^p)$ while holding the observed $(X_{i1}, \dots, X_{in_i})$ fixed, and are given by

$$X_{i,p} = \sum_{l=1}^c (r-l) \sum_{j: y_{ij}^p = l} X_{ij}, p=1, \dots, P.$$

3. Glioblastoma study

The Partners Human Research Committee approved the glioblastoma study. Subjects were accrued from the Massachusetts General Hospital (MGH) Brain Tumor Database, which contains patients consented for inclusion beginning in 2003. Initial inclusion criteria were: patients with GBM or glioblastoma with oligodendroglial features (GBMO); no prior diagnosis of a lower grade tumor; no treatment prior to diagnosis of GBM or GBMO; and archival pathology available at MGH. This yielded 661 patients, of whom 46 patients had lived longer than three years (labeled long-term survivors, or LTS). Overall survival data was calculated from date of surgery to the date of death or date of last follow-up. The remaining 615 patients were separated into those living less than six months (labeled short-term survivors, or STS) and those living between six months and three years (labeled standard survivors, or SS). To be classified as an STS or SS, complete survival data were required, and therefore patients who were still alive, but followed less than three years, were excluded. Paraffin-embedded tumor tissue and clinical information were available for 35 LTS and 52 STS. We then matched on age and KPS at diagnosis to obtain sets of patients that ideally included three standard survivors, one long-term survivor, and one short-term survivor. Age and KPS were chosen as matching variables since they are well-established, important clinical prognostic factors in GBM ([4], [5], [6], [7], [8], [9]). Our final matched cohort included 145 subjects who comprised 22 complete sets (three standard survivors, one LTS, one STS) and eight incomplete sets (four sets with three standard survivors and one LTS, two sets with three standard survivors and one STS, one set with four standard survivors, one LTS and one STS and one set with four standard survivors and one STS).

We let subject $j = 1$ represent the short term survivor, subjects $j = 2, 3, 4$ represent the standard survivors, and subject $j = 5$ represent the long term survivor. The eight incomplete matched sets described above present no difficulty to the proposed method; if a particular survivor type is absent from a given matched set, their index is simply ignored and associated X set to zero. We describe this for the GBM study below. In our development here we treat the standard survivors as the reference category. However, any level of disease could be treated as such with similar results (of course it must be the same level for all subjects). The contribution to the conditional likelihood for a complete matched set i , simplified from (3) for the glioblastoma study, is

$$\begin{aligned} L_i &= P(Y_{i1}=1, Y_{i2}=Y_{i3}=Y_{i4}=2, Y_{i5}=3 | X_{ij}, j=1, \dots, 5, d_{i1}=d_{i3}=1, d_{i2}=3) \\ &= \frac{\exp(\gamma_{i1}^* + \beta_1^* X_{i1} + \gamma_{i3}^* + \beta_3^* X_{i5})}{\sum_{j=1}^5 \sum_{j' \neq j} \exp(\gamma_{i1}^* + \beta_1^* X_{ij} + \gamma_{i3}^* + \beta_3^* X_{ij'})} \\ &= \frac{\exp(\beta_1^* X_{i1} + \beta_3^* X_{i5})}{\sum_{j=1}^5 \sum_{j' \neq j} \exp(\beta_1^* X_{ij} + \beta_3^* X_{ij'})}. \end{aligned}$$

This expression suggests that this model may be fit using standard conditional logistic regression, with appropriate transformation of the data and the parameters of interest. For the glioblastoma study, each complete matched set can be viewed as having a single case, with covariate vector (X_{i1}, X_{i5}) , and 19 controls, with covariate vectors consisting of all pairs, $(X_{ij}, X_{ij'})$, with $(j, j') \in (1, 5)$ and $j \neq j'$. For this study, with $r = 2$ and $c = 3$, $\beta_1 = \beta_1^*$ and $\beta_2 = -\beta_3^*$ from the original model formulation (1). Thus, the likelihood contribution can be written equivalently as

$$L_i = \frac{\exp(\beta_1 X_{i1} - \beta_2 X_{i5})}{\sum_{j=1}^5 \sum_{j' \neq j} \exp(\beta_1 X_{ij} - \beta_2 X_{ij'})}.$$

Thus, to estimate β_1 and β_2 directly from this conditional logistic regression formulation, each case is associated with covariate vector $(X_{i1}, -X_{i5})$ and each of the 19 controls is associated with covariate vector $(X_{ij}, -X_{ij'})$, with $(j, j') \in (1, 5)$ and $j \neq j'$. The interpretation of β_1 is the increase in log odds ratio of short-term survival versus standard survival, associated with a one unit increase in X . The interpretation of β_2 is the increase in log odds ratio of standard survival versus long-term survival, associated with a one unit increase in X . The likelihood contributions from the eight incomplete matched sets are derived similarly. For example, there are four sets that contain three standard-term survivors and one short-term survivor. The likelihood contribution for these sets is given by

$$L_i = \frac{\exp(\beta_1 X_{i1})}{\sum_{j=1}^4 \exp(\beta_1 X_{ij})}.$$

And there are three sets that contain three standard-term survivors and one long-term survivor. The likelihood contribution for these sets is given by

$$L_i = \frac{\exp(-\beta_2 X_{i5})}{\sum_{j=1}^4 \exp(-\beta_2 X_{ij'})}.$$

The remaining sets contain four standard-term survivors and one or two of the other types of survivors. These can be accommodated similarly, though due to the larger sets, the notation is cumbersome and we do not present their likelihood contributions.

Under the single β adjacent categories model, $\beta = \beta_1 = \beta_2$, so that

$$L_i = \frac{\exp[\beta(X_{i1} - X_{i5})]}{\sum_{j=1}^5 \sum_{j' \neq j} \exp[\beta(X_{ij} - X_{ij'})]}.$$

To fit this model, the case for set i is associated with the single covariate $(X_{i1} - X_{i5})$, and each control is likewise associated with the single covariate $(X_{ij} - X_{ij'})$, with $(j, j') \in (1, 5)$ and $j \neq j'$.

Our approach to analyzing the glioblastoma study is to first fit univariate, two- β models for the covariates of interest. These include clinical and demographic features (e.g., extent of surgery, tumor location, initial therapy, MGMT methylation status), hypoxia markers (e.g., HIF1 α , CA9, and GLUT1 staining measures), and vascular markers (e.g., CD31 and SMA). We then test whether $\beta_1 = \beta_2$, under the assumption of model (1), using a likelihood ratio test. If we do not reject this null hypothesis within the framework of the joint model, we fit the single- β model instead. Again, assuming that model (1) holds, this estimator will have optimal efficiency. In the case of univariate analysis of a binary covariate, (1) is the saturated model, and holds trivially. Based on simulation studies that we conducted, the test of $\beta_1 = \beta_2$ has over 70% power with 25 matched sets of the same structure as the GBM study when $\beta_1 = 0$ and $\beta_2 > 0.5$. With 50 matched sets, the test has greater than 73% power when $\beta_1 = 0$ and $\beta_2 > 0.35$.

Table 1 lists selected results from the univariate analyses that we conducted for the glioblastoma study. For those covariates for which we did not reject $H_0 : \beta_1 = \beta_2$, we list both the estimated single β and the estimated β_1 and β_2 from the full likelihood estimation. For these estimates of β_1 and β_2 , we estimate, based on simulations, that we have a range of power, from 30% to 100%. Due to concerns about the power of the test of $\beta_1 = \beta_2$, we report the estimates for β_1 and β_2 even when the null hypothesis is not rejected and we proceed to estimate the single β . The expected gains of the single β model are seen for several of the covariates. For example, the p -value for testing $H_0 : \beta_1 = \beta_2$ for MGMT methylation is 0.995, favoring the single β model. In this model, $\exp(\hat{\beta}) = 0.32$, which is marginally significantly different from zero ($p = 0.074$). In contrast, neither test for $\beta_1 = 0$ or $\beta_2 = 0$ from the two- β model is significant. Similarly, the p -value for testing the single β model for CA9 perinecrotic intensity is 0.639, favoring that model. In this model $\exp(\hat{\beta}) = 0.549$, which is significantly different from zero ($p = 0.048$). In contrast, in the two- β model, $\exp(\hat{\beta}_1) = 0.462$ ($p = 0.109$) and $\exp(-\hat{\beta}_2) = 0.66$ ($p = 0.38$), neither of which is significantly different from zero. For the covariates listed for which we reject $H_0 : \beta_1 = \beta_2$, we report only the results from the two- β model. For example, biopsy versus resection is highly significantly associated with a higher probability of being a STS relative to a standard survivor (OR=34.76, $p < 0.001$), but is not significantly associated with standard versus long term survival (OR=1.17, $p = 0.808$). The single- β model is also rejected for corpus collosum location and for CA9 perinecrotic staining, a marker of hypoxia. In an analysis reported separately [14], we then proceeded to fit a multivariate model, using covariates with significant coefficients in univariate analyses, and using either one or two β 's for each covariate, as determined by our tests of $H_0 : \beta_1 = \beta_2$ in the univariate analyses.

4. Pseudo-likelihood approach

In the context of polychotomous, matched logistic regression, with one case and two types of controls in each set, Liang and Stewart [20] considered an alternative approach to estimation based on separate conditional logistic regression models for each case-control

comparison. For the glioblastoma study, this approach translates into fitting two separate conditional logistic regression models using the STS and standard survivors in one, and the LTS and standard survivors in the other, and amounts to using a pseudo-likelihood function that is given by

$$\begin{aligned}
 PL_i &= P(Y_{i1}=1, Y_{i2}=Y_{i3}=Y_{i4}=2 | X_{ij}, j=1, \dots, 4, d_{i1}=1, d_{i2}=3) \times P(Y_{i2}=Y_{i3}=Y_{i4}=2, Y_{i5}=3 | X_{ij}, j=2, \dots, 5, d_{i3}=1, d_{i2}=3) \\
 &= \frac{\exp(\beta_1 X_{i1})}{\sum_{j=1}^4 \exp(\beta_1 X_{ij})} \times \frac{\exp(-\beta_2 X_{i5})}{\sum_{j=2}^5 \exp(-\beta_2 X_{ij})} \\
 &= \frac{\exp(\beta_1 X_{i1} - \beta_2 X_{i5})}{\sum_{j=1}^4 \sum_{j=2}^5 \exp(\beta_1 X_{ij} - \beta_2 X_{i,j'})}.
 \end{aligned} \tag{4}$$

Thus, estimation of β_1 and β_2 via the pseudo-likelihood approach is equivalent to fitting a single conditional logistic regression model, with one case having covariate vector $(X_{i1}, -X_{i5})$ and with 15 controls having covariate vectors $(X_{ij}, -X_{ij'})$, with $j = 1, \dots, 4, j' = 2, \dots, 5$ and $(j, j') \neq (1, 5)$. A similar result was noted in the context of pseudo-likelihood estimation for a polychotomous logistic regression model[28]. Notice that the denominator of the pseudo-likelihood contribution differs from that of the full likelihood contribution, and produces 15 controls, as compared to the 19 controls produced by the full likelihood (the 15 are not all included among the 19).

An approach of considering every dichotomization of the ordinal outcome to enable conditional logistic regression was proposed in the context of the proportional odds model to enable elimination of nuisance parameters through conditioning, which is not afforded by that model [29]. This pseudo-likelihood is fit through stratified conditional likelihood, with each stratum defined by the threshold for dichotomization. Generalized estimation equations are used for variance estimation. The motivation for our pseudo-likelihood is not to enable elimination of nuisance parameters, as this is already accommodated by the full likelihood based on the adjacent categories model. Rather, it is for potential robustness of estimation in settings in which part of the full model might be misspecified, e.g., the relationship between standard and long term survivors. Operationally, GEE can be implemented in our setting through direct use of a Cox regression fitting procedure (e.g., the *coxph* function in the R survival package, with properly specified strata and time variables to yield the conditional logistic likelihood) with matched sets defined to be clusters. We investigate the robustness of this approach and the need for GEE in our simulation studies in Section 5.

Under standard conditions and assuming model (1), both the full and pseudo-likelihood approaches produce consistent estimators for (β_1, β_2) . The loss of efficiency of the pseudo-likelihood approach relative to the full likelihood approach depends on the true (β_1, β_2) ; e.g., if $\beta_1 = 0$, then there is likely to be a loss in efficiency for estimation of β_2 , as the full conditional likelihood approach effectively includes more controls in each set. Given the simplicity of estimation of the full likelihood model, and its definite advantage for the single β adjacent categories model, there is no advantage to the pseudo-likelihood approach if the adjacent categories model (1) holds. However, while β_1 and β_2 have interpretations as log odds ratios between adjacent levels of disease (i.e., STS to standard, and standard to LTS), the model (1) also implicitly fixes the log odds ratio between non-adjacent levels of disease. That is, it implies that the log odds ratio for being a STS versus a LTS, associated with a one

unit increase in X is $\beta_1 + \beta_2$. This is also implied by the separate models utilized by the pseudo-likelihood approach, though importantly, it does not make any use of this in its estimation. That is, the pseudo-likelihood estimation is not restricted at all by any assumption on the relationship between STS and LTS, and thus is preferred over the full likelihood approach in the case of a non-binary covariate X , the case when it is unknown whether (1) holds. That is, if X is related linearly to the log of the ratio of the probabilities of STS to that of standard survival but a similar linear model does not hold for standard survival versus LTS, then the pseudo-likelihood still provides a consistent estimator for β_1 , but not for β_2 . In contrast, the full likelihood estimator for β_1 is not consistent as it assumes a linear relationship between X and the ratio of probabilities of STS and LTS, which influences $\hat{\beta}_1$.

In the glioblastoma study, all of the variables for which the test $H_0 : \beta_1 = \beta_2$ is rejected are binary. Thus, there is no advantage to applying the pseudo-likelihood estimation for them.

5. Simulation Study

We conducted simulation studies in which we used the full likelihood and pseudo-likelihood to estimate the general adjacent categories model (1), and the full likelihood to estimate the simple adjacent categories model that sets $\beta_1 = \beta_2$ in (1). We had four aims in our simulation studies. First, we aimed to compare the full likelihood estimation procedure to the pseudo-likelihood estimation procedure under a variety of conditions: (1) when the simple adjacent categories model holds ($\beta_1 = \beta_2$; Table 2); (2) when the general adjacent categories model holds ($\beta_1 \neq \beta_2$; Table 3); and (3) when the general adjacent categories model does not hold (see model (5) below; Table 4). Second, we aimed to investigate the magnitude of the expected mean squared error benefit of estimates based on assuming the simple adjacent categories model when the model holds (Table 2) and to investigate the magnitude of the bias that is incurred when the simple model does not hold (Table 3). Third, we aimed to investigate whether the pseudo-likelihood estimation offers robustness to incorrect specification of part of the model, such as the relationship between standard term survivors and long term survivors (Table 4). In this context, we also aimed to investigate further the expected bias of the incorrect assumption of the simple adjacent categories model. And fourth, we aimed to assess the accuracy of the information based standard errors based on the derived conditional logistic likelihood (4) relative to standard errors based on generalized estimating equations, which acknowledge the replicated use of the standard survivors in the pseudo-likelihood.

We patterned our study after the glioblastoma study, and considered $n = 50, 75, 100$ and 150 matched sets, each with one short-term survivor, three standard survivors, and one long-term survivor. We assumed the following model that violates model (1) with nonlinearity in X for the log-odds of level 3 versus level 2:

$$\begin{aligned} \log \frac{P(Y_{ij}=1|X_{ij})}{P(Y_{ij}=2|X_{ij})} &= \gamma_i + \alpha_1 + \beta_1 X_{ij}, \quad i=1, \dots, n, \quad j=1, \dots, n_i, \quad l=1, \dots, c-1. \\ \log \frac{P(Y_{ij}=3|X_{ij})}{P(Y_{ij}=2|X_{ij})} &= -\gamma_i - \alpha_2 - \beta_2 X_{ij}^a, \quad i=1, \dots, n, \quad j=1, \dots, n_i, \quad l=1, \dots, c-1. \end{aligned} \quad (5)$$

When $a = 1$, this model is equivalent to model (1), and when $a \neq 1$, it is not. We then generated the covariate, X from a normal distribution with mean zero and standard deviation of two, the set-specific intercepts, γ_i , from a standard normal distribution, we took the level-specific intercepts, α_i to be zero, and we varied (β_1, β_2, a) . For each simulated study, we estimated (β_1, β_2) and $\beta = \beta_1 = \beta_2$, using the full likelihood approach as in (3), and (β_1, β_2) using the pseudo-likelihood approach as in (4). We also calculated the mean squared error (MSE) of the various estimators for comparison.

Table 2 lists the results of a simulation (3000 repetitions) in which matched sets were generated under model (1), with $\beta_1 = \beta_2$. We report the estimates and MSE's from full and pseudo-likelihood estimation (neither of which assumed $\beta_1 = \beta_2$), as well as from full likelihood estimation that assumed correctly that $\beta_1 = \beta_2$. We first note that the pseudo-likelihood estimates are nearly identical to the full likelihood estimates. The MSE's are also very close for $n = 75$ and nearly identical for larger sample sizes. Secondly, as expected, assumption of the correct model for estimation yields the smallest MSE's; these are roughly 30–50% of the separate estimates of β_1 and β_2 . This demonstrates the advantage to establishing equivalence of the β s, when it holds, followed by use of full-likelihood single- β estimation.

Table 3 lists the results of a simulation (3000 repetitions) in which matched sets were generated under model (1), with $\beta_1 \neq \beta_2$. We report the estimates and MSE's from full and pseudo-likelihood estimation (neither of which assumed $\beta_1 = \beta_2$), as well as from full likelihood estimation that incorrectly assumed that $\beta_1 = \beta_2$. The pseudo-likelihood estimation produces nearly identical estimates as the full likelihood estimation, with slightly higher MSE's for β_2 . As expected, the single β full-likelihood approach displays substantial bias, again emphasizing the need to test for the simple adjacent categories model before using it.

Table 4 lists the results of a simulation (3000 repetitions) in which matched sets were generated under model (5), which deviates from the general adjacent categories model and additionally does not have $\beta_1 = \beta_2$. We report the estimates and MSE's from full and pseudo-likelihood estimation (which did assume the general adjacent categories model, but not that $\beta_1 = \beta_2$), as well as from full likelihood estimation that incorrectly assumed that $\beta_1 = \beta_2$. We considered $n = 75$ in all cases. Not surprisingly, the pseudo-likelihood estimation of β_1 performed well regardless of the true values of β_2 or of the exponent, a . The full likelihood based estimation of β_1 demonstrated bias that is severe for $a = 2$ and moderate for $a = -1$ and $a = -2$. This reflects the larger contribution of X to the probability of long term survival versus standard survival when its exponent is positive rather than negative, and thus the larger influence of the misspecified portion of the model on the full likelihood estimation. Both the full likelihood and pseudo-likelihood estimates of β_2 that assume model (1) with $a = 1$ are highly biased with very high mean squared errors. Not surprisingly, single- β estimation is likewise biased with high mean squared errors.

Table 5 lists the results of a simulation (3000 repetitions) in which matched sets were generated under model (5) and the information based standard errors from the pseudo-likelihood (4) are compared to GEE based standard errors from the pseudo-likelihood. For n

$= 75$ both the standard pseudo-likelihood and the GEE pseudo-likelihood standard errors of $\hat{\beta}_1$ slightly underestimate (by at most 5%) the simulation-based estimates. With $n = 200$, this diminishes to less than 1%. For both sample sizes, the coverage of the 95% confidence intervals calculated from the information based standard errors is very close to 95%, while the coverage of the confidence intervals calculated from the GEE standard errors is approximately 97–98%. For scenarios in which $\alpha = 1$ (and the adjacent categories model does not hold), the GEE standard errors for $\hat{\beta}_2$ better approximate the simulation based standard errors than the information based standard errors. However, given the bias of $\hat{\beta}_2$ in this setting that is seen in Table 4, this is not an important advantage of the GEE estimates.

6. Discussion

We have shown that full likelihood estimation of an adjacent categories model for ordinal, disease-based matched sets is accessible via standard conditional logistic regression procedures available in statistical software packages. We have described the requisite pre-processing of the matched sets to facilitate this translation. For univariate analyses of single, dichotomous covariates, there is no advantage to the alternative pseudo-likelihood approach that involves multiple conditional logistic regression modeling of each disease level versus the reference level over the full likelihood approach that jointly estimates the effects of covariates on each disease level. Furthermore, as expected, if the effects on different disease levels are equivalent, there is a clear advantage to the full likelihood, single- β , estimation. If there are multiple and continuous covariates, then the pseudo-likelihood approach may be more robust than the full likelihood approach, as it will yield a consistent estimator for any part of the full model that does hold, and its estimation of the misspecified parts of the model will not affect those that are correctly specified. Based on our simulation studies, it appears that use of the information based standard errors from the derived conditional logistic likelihood is acceptable and provides close to the nominal coverage. This requires further investigation in other settings, including those with larger matched sets, and in particular, with larger numbers in the reference set, which induces higher correlation between components of the pseudo-likelihood. In such cases, model based variances will be underestimates, and sandwich estimators of variance should be used.

We have demonstrated our fitting of the single and two- β models on the glioblastoma study. We found that CA9 perinecrotic staining is significantly associated with standard survival versus short term survival (OR=0.22; $p=0.002$), but also with standard survival versus long term survival (OR=2.36; $p=0.099$). This result supports the use of the general adjacent categories model, versus a proportional odds model, for these data. We also found that the simple adjacent categories model with respect to MGMT methylation is not rejected by the data, with a nearly significant association between MGMT methylation and longer survival (OR=0.32; $p=0.074$), which was not detected by the general model.

A SAS macro with our example and R code for our simulations are available at <http://www.hsph.harvard.edu/rebecca-betensky/software/>.

Acknowledgments

Contract/grant sponsor: NIH; R01-CA075971, K12-CA090354, P30-CA06516

References

1. Mukherjee B, Liu I, Sinha S. Analysis of matched case-control data with multiple ordered disease states: Possible choices and comparisons. *Statistics in Medicine*. 2007; 26:3240–3257. [PubMed: 17206600]
2. CBTRUS. CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2004–2007. Central Brain Tumor Registry of the United States; 2011. [cited May 25, 2011]; Available from: <http://www.cbtrus.org/2011-NPCR-SEER/WEB-0407-Report-3-3-2011.pdf>
3. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *The New England Journal of Medicine*. 2005; 352:987–96. [PubMed: 15758009]
4. Burton EC, Lamborn KR, Feuerstein BG, Prados M, Scott J, Forsyth P, et al. Genetic aberrations defined by comparative genomic hybridization distinguish long-term from typical survivors of glioblastoma. *Cancer Res*. 2002; 62:6205–10. [PubMed: 12414648]
5. Krex D, Klink B, Hartmann C, von Deimling A, Pietsch T, Simon M, et al. Long-term survival with glioblastoma multiforme. *Brain*. 2007; 130:2596–606. [PubMed: 17785346]
6. Martinez R, Schackert G, Yaya-Tur R, Rojas-Marcos I, Herman JG, Esteller M. Frequent hypermethylation of the DNA repair gene MGMT in long-term survivors of glioblastoma multiforme. *J Neurooncol*. 2007; 83:91–3. [PubMed: 17164975]
7. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol*. 2009; 10:459–66. [PubMed: 19269895]
8. Barbus S, Tews B, Karra D, Hahn M, Radlwimmer B, Delhomme N, et al. Differential retinoic acid signaling in tumors of long- and short-term glioblastoma survivors. *J Natl Cancer Inst*. 2011; 103:598–606. [PubMed: 21346226]
9. Sonoda Y, Kumabe T, Watanabe M, Nakazato Y, Inoue T, Kanamori M, et al. Long-term survivors of glioblastoma: clinical features and molecular analysis. *Acta Neurochir (Wien)*. 2009; 151:1349–58. [PubMed: 19730774]
10. Louis, DN.; Ohgaki, H.; Wiestler, OD.; Cavenee, WK., editors. WHO Classification of Tumours of the Central Nervous System. 4. Lyon: International Agency for Research on Cancer; 2007.
11. Bar EE. Glioblastoma, cancer stem cells and hypoxia. *Brain Pathol*. 2011; 21:119–29.
12. Jain RK. Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. *Science*. 2005; 307:58–62. [PubMed: 15637262]
13. Jain RK, Finn AV, Kolodgie FD, Gold HK, Virmani R. Antiangiogenic therapy for normalization of atherosclerotic plaque vasculature: a potential strategy for plaque stabilization. *Nat Clin Pract Cardiovasc Med*. 2007; 4:491–502. [PubMed: 17712362]
14. Lee EQ, Jansen M, Yip S, Szymonifka J, Camelo-Piragua SI, Ryg P, Provencher KA, Laban AS, Kim JCM, McCormack K, Louis DN, Betensky RA, Batchelor TT, Nutt CL. Comprehensive analysis of clinical and tissue biomarker correlates of survival in a matched case-control study of patients with glioblastoma. Unpublished manuscript. 2013
15. Rothman, KJ.; Greenland, S.; Lash, TL. *Modern Epidemiology*. Lippincott Williams & Wilkins; Philadelphia: 2008.
16. Schlesselman, JJ. *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press; Oxford: 1982.
17. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
18. Agresti A, Lang JB. A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika*. 1993; 80:5237–534.
19. Whitehead J. Sample size calculations for ordered categorical data. *Statistics in Medicine*. 1993; 12:2257–2271. [PubMed: 8134732]

20. Liang KY, Stewart WF. Polychotomous logistic regression methods for matched case-control studies with multiple case or control groups. *Am J Epidemiol.* 1987; 125:720–730. [PubMed: 3826049]
21. Becher H, Jockel KH. Bias adjustment with polychotomous logistic regression in matched case-control studies with two control groups. *Biometrical Journal.* 1990; 7:801–816.
22. Becher H. Alternative parameterization of polychotomous models: theory and application to matched case-control studies with two control groups. *Statistics in Medicine.* 1991; 10:375–382. [PubMed: 2028121]
23. Levin B. The saddlepoint correction in conditional logistic likelihood analysis. *Biometrika.* 1990; 77:275–285.
24. Chatterjee N. A two-stage regression model for epidemiological studies with multivariate disease classification data. *Journal of the American Statistical Association.* 2005; 99:127–138.
25. Gebregziabher M, Guimaraes P, Cozen W, Conti DV. A polytomous conditional likelihood approach for combining matched and unmatched case-control studies. *Statistics in Medicine.* 2010; 29:1004–1013. [PubMed: 20069625]
26. Agresti, A. Modeling ordinal categorical data. 2010. http://www.stat.ufl.edu/~aa/ordinal/agresti_ordinal_tutorial.pdf
27. Anderson JA. Regression and ordered categorical variables. *JR Statist Soc B.* 46:1–30.
28. Risch HA, Tibshirani RJ. Re: “Polychotomous logistic regression methods for matched case-control studies with multiple case or control groups” (Letter.). *Am J Epidemiol.* 1988; 128:446–448.
29. Mukherjee B, Ahn J, Liu I, Rathouz PJ, Sanchez BN. Fitting stratified proportional odds models by amalgamating conditional likelihoods. *Statistics in Medicine.* 2008; 27:4950–4971. [PubMed: 18618428]

Table 1

Selected results of univariate analyses.

covariate	$H_0 : \beta_1 = \beta_2$ <i>p</i> -value	single β $\exp(\beta)$ (<i>p</i> -value)	STS vs standard $\exp(\beta_1)$ (<i>p</i> -value)	standard vs LTS $\exp(-\beta_2)$ (<i>p</i> -value)
gross total resection	0.110	0.382 (0.014)	0.095 (0.028)	0.70 (0.503)
biopsy	0.009		34.75 (< 0.001)	1.17 (0.808)
MGMT methylation	0.995	0.320 (0.074)	< 0.001 (0.995)	0.612 (0.526)
cerebral hemisphere	0.227	0.126 (< 0.001)	0.088 (< 0.001)	0.461 (0.473)
corpus collosum	0.060		19.74 (< 0.001)	0.64 (0.723)
XRT alone	0.991	2.785 (0.053)	1.837 (0.354)	NA
XRT & TMZ	0.110	0.363 (0.003)	0.162 (0.006)	0.66 (0.390)
CA9 perinecrotic intensity	0.639	0.549 (0.048)	0.462 (0.109)	0.66 (0.380)
CA9 perinecrotic staining	0.004		0.222 (0.002)	2.36 (0.099)
GLUT1 MVP staining	0.447	2.667 (0.018)	5.290 (0.115)	2.038 (0.191)
GLUT1 necrosis intensity	0.646	0.504 (0.028)	0.433 (0.066)	0.61 (0.354)
α SMA stain intensity	0.564	0.438 (0.010)	0.359 (0.029)	0.56 (0.279)
α SMA MVP stain intensity	0.651	0.532 (0.046)	0.645 (0.406)	0.46 (0.090)

Estimates and MSE's under model (1) with $l_f = l_g = \beta_2$: 3000 repetitions, n matched sets containing one short term survivor, three standard survivors, and one long-term survivor.

Table 2

n	full likelihood ^l			pseudo-likelihood ²			single β		full likelihood			pseudo-likelihood		
	β_1	β_2	$\hat{\beta}_1$	MSE $\hat{\beta}_1$	$\hat{\beta}_1$	MSE $\hat{\beta}_1$	$\hat{\beta}$	MSE $\hat{\beta}$	$\hat{\beta}_2$	MSE $\hat{\beta}_2$	$\hat{\beta}_2$	$\hat{\beta}_2$	MSE $\hat{\beta}_2$	$\hat{\beta}_2$
50	0.25	0.25	0.256	0.009	0.257	0.009	0.253	0.004	0.253	0.009	0.253	0.253	0.009	0.009
	0.50	0.50	0.511	0.014	0.510	0.015	0.507	0.006	0.515	0.015	0.516	0.516	0.015	0.015
	0.75	0.75	0.773	0.027	0.772	0.028	0.761	0.013	0.772	0.028	0.772	0.772	0.028	0.028
	1.00	1.00	1.043	0.046	1.043	0.047	1.021	0.021	1.034	0.046	1.034	1.034	0.046	0.046
	1.50	1.50	1.580	0.122	1.580	0.124	1.533	0.047	1.562	0.111	1.562	1.562	0.111	0.111
	2.00	2.00	2.112	0.233	2.113	0.237	2.051	0.099	2.114	0.264	2.114	2.114	0.267	0.267
75	0.25	0.25	0.254	0.006	0.254	0.006	0.253	0.002	0.255	0.006	0.255	0.255	0.006	0.006
	0.50	0.50	0.510	0.009	0.511	0.010	0.506	0.004	0.509	0.010	0.509	0.509	0.010	0.010
	0.75	0.75	0.764	0.017	0.764	0.017	0.760	0.008	0.770	0.018	0.770	0.770	0.018	0.018
	1.00	1.00	1.026	0.028	1.026	0.029	1.014	0.013	1.024	0.029	1.024	1.024	0.029	0.029
	1.50	1.50	1.545	0.070	1.545	0.070	1.521	0.031	1.541	0.065	1.541	1.541	0.065	0.065
	2.00	2.00	2.058	0.135	2.059	0.135	2.025	0.059	2.061	0.128	2.061	2.061	0.129	0.129
100	0.25	0.25	0.251	0.004	0.251	0.004	0.250	0.002	0.252	0.004	0.253	0.253	0.004	0.004
	0.50	0.50	0.505	0.007	0.505	0.007	0.503	0.003	0.507	0.007	0.507	0.507	0.008	0.008
	0.75	0.75	0.758	0.012	0.758	0.013	0.754	0.006	0.760	0.012	0.760	0.760	0.012	0.012
	1.00	1.00	1.016	0.020	1.016	0.021	1.007	0.009	1.015	0.021	1.016	1.016	0.021	0.021
	1.50	1.50	1.533	0.049	1.533	0.049	1.515	0.021	1.531	0.047	1.530	1.530	0.047	0.047
	2.00	2.00	2.056	0.097	2.056	0.097	2.023	0.044	2.044	0.095	2.044	2.044	0.096	0.096
150	0.25	0.25	0.252	0.177	0.252	0.177	2.534	0.079	2.575	0.170	2.576	2.576	0.171	0.171
	0.25	0.25	0.253	0.003	0.253	0.003	0.252	0.001	0.252	0.003	0.252	0.252	0.003	0.003
	0.50	0.50	0.505	0.005	0.506	0.005	0.503	0.002	0.504	0.005	0.504	0.504	0.005	0.005
	0.75	0.75	0.756	0.008	0.756	0.008	0.753	0.004	0.757	0.008	0.757	0.757	0.008	0.008
	1.00	1.00	1.008	0.013	1.008	0.013	1.005	0.006	1.012	0.013	1.012	1.012	0.014	0.014
	1.00	1.00	1.008	0.013	1.008	0.013	1.005	0.006	1.012	0.013	1.012	1.012	0.014	0.014

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

<i>n</i>	full likelihood ¹		pseudo-likelihood ²		single β^3		full likelihood		pseudo-likelihood	
	$\hat{\beta}_1$	MSE $\hat{\beta}_1$	$\hat{\beta}_1$	MSE $\hat{\beta}_1$	$\hat{\beta}$	MSE $\hat{\beta}$	$\hat{\beta}_2$	MSE $\hat{\beta}_2$	$\hat{\beta}_2$	MSE $\hat{\beta}_2$
1.50	1.50	1.513	1.512	0.029	1.506	0.014	1.521	0.032	1.522	0.032
2.00	2.00	2.027	2.027	0.059	2.013	0.027	2.034	0.060	2.034	0.060
2.50	2.50	2.550	2.550	0.110	2.521	0.050	2.547	0.106	2.547	0.106

¹ Full likelihood uses (3) for estimation.

² Pseudo-likelihood uses (4) for estimation.

³ Single β uses (3) for estimation, setting $\beta_1 = \beta_2$

Table 3

Estimates and MSE's under model (1) with I_f β_1 β_2 : 3000 repetitions, n matched sets containing one short term survivor, three standard survivors, and one long-term survivor.

n	full likelihood ¹				pseudo-likelihood ²				single β^3				full likelihood				pseudo-likelihood			
	β_1	β_2	$\hat{\beta}_1$	MSE $\hat{\beta}_1$	$\hat{\beta}_1$	MSE $\hat{\beta}_1$	$\hat{\beta}_1$	MSE $\hat{\beta}_1$	$\hat{\beta}_1$	MSE $\hat{\beta}_1$	$\hat{\beta}_1$	MSE $\hat{\beta}_1$	$\hat{\beta}_2$	MSE $\hat{\beta}_2$	$\hat{\beta}_2$	MSE $\hat{\beta}_2$	$\hat{\beta}_2$	MSE $\hat{\beta}_2$	$\hat{\beta}_2$	MSE $\hat{\beta}_2$
50	0.00	0.25	-0.001	0.007	-0.001	0.007	0.007	0.018	0.125	0.018	0.018	0.008	0.259	0.008	0.259	0.008	0.259	0.008	0.259	0.008
	0.00	0.50	-0.004	0.008	-0.004	0.008	0.008	0.058	0.234	0.058	0.058	0.012	0.512	0.012	0.512	0.012	0.512	0.012	0.512	0.012
	0.00	0.75	0.001	0.009	0.001	0.009	0.009	0.112	0.330	0.112	0.112	0.021	0.776	0.021	0.776	0.021	0.776	0.021	0.776	0.021
	0.00	1.00	0.002	0.010	0.002	0.010	0.010	0.165	0.402	0.165	0.165	0.038	1.036	0.038	1.036	0.038	1.036	0.038	1.036	0.038
	0.00	1.50	-0.002	0.013	-0.002	0.013	0.013	0.253	0.497	0.253	0.253	0.136	1.591	0.136	1.591	0.136	1.591	0.136	1.591	0.136
75	0.00	0.25	0.001	0.005	0.001	0.005	0.005	0.017	0.124	0.017	0.017	0.005	0.254	0.005	0.254	0.005	0.254	0.005	0.254	0.005
	0.00	0.50	-0.001	0.005	-0.001	0.005	0.005	0.057	0.234	0.057	0.057	0.008	0.507	0.008	0.507	0.008	0.507	0.008	0.507	0.008
	0.00	0.75	0.001	0.006	0.001	0.006	0.006	0.110	0.328	0.110	0.110	0.014	0.765	0.014	0.765	0.014	0.765	0.014	0.765	0.014
	0.00	1.00	-0.001	0.007	-0.001	0.007	0.007	0.161	0.398	0.161	0.161	0.023	1.022	0.023	1.022	0.023	1.022	0.023	1.022	0.023
	0.00	1.50	0.001	0.008	0.001	0.008	0.008	0.249	0.496	0.249	0.249	0.077	1.557	0.077	1.557	0.077	1.557	0.077	1.557	0.077
100	0.00	0.25	-0.001	0.009	-0.001	0.009	0.009	0.308	0.552	0.308	0.308	0.271	2.112	0.271	2.112	0.271	2.112	0.271	2.112	0.271
	0.00	0.50	0.002	0.003	0.002	0.003	0.003	0.017	0.125	0.017	0.017	0.004	0.253	0.004	0.253	0.004	0.253	0.004	0.253	0.004
	0.00	0.75	-0.002	0.004	-0.002	0.004	0.004	0.057	0.235	0.057	0.057	0.006	0.505	0.006	0.505	0.006	0.505	0.006	0.505	0.006
	0.00	1.00	-0.001	0.005	-0.001	0.005	0.005	0.107	0.325	0.107	0.107	0.010	0.761	0.010	0.761	0.010	0.761	0.010	0.761	0.010
	0.00	1.50	-0.001	0.006	-0.001	0.006	0.006	0.247	0.495	0.247	0.247	0.049	1.542	0.049	1.542	0.049	1.542	0.049	1.542	0.049
150	0.00	0.25	-0.002	0.007	-0.002	0.007	0.007	0.306	0.551	0.306	0.306	0.132	2.082	0.132	2.082	0.132	2.082	0.132	2.082	0.132
	0.00	0.50	0.001	0.002	0.001	0.002	0.002	0.016	0.124	0.016	0.016	0.003	0.252	0.003	0.252	0.003	0.252	0.003	0.252	0.003
	0.00	0.75	-0.001	0.003	-0.001	0.003	0.003	0.056	0.234	0.056	0.056	0.004	0.505	0.004	0.505	0.004	0.505	0.004	0.505	0.004
	0.00	1.00	0.001	0.003	0.001	0.003	0.003	0.108	0.326	0.108	0.108	0.006	0.757	0.006	0.757	0.006	0.757	0.006	0.757	0.006
	0.00	1.50	-0.001	0.004	-0.001	0.004	0.004	0.245	0.493	0.245	0.245	0.030	1.527	0.030	1.527	0.030	1.527	0.030	1.527	0.030
I_f	0.00	0.25	0.001	0.005	0.001	0.005	0.005	0.306	0.551	0.306	0.306	0.072	2.052	0.072	2.052	0.072	2.052	0.072	2.052	0.072
	0.00	0.50	-0.001	0.003	-0.001	0.003	0.003	0.016	0.124	0.016	0.016	0.003	0.252	0.003	0.252	0.003	0.252	0.003	0.252	0.003
	0.00	0.75	0.001	0.003	0.001	0.003	0.003	0.108	0.326	0.108	0.108	0.006	0.757	0.006	0.757	0.006	0.757	0.006	0.757	0.006
	0.00	1.00	0.002	0.003	0.002	0.003	0.003	0.159	0.397	0.159	0.159	0.011	1.011	0.011	1.011	0.011	1.011	0.011	1.011	0.011
	0.00	1.50	-0.001	0.004	-0.001	0.004	0.004	0.245	0.493	0.245	0.245	0.030	1.527	0.030	1.527	0.030	1.527	0.030	1.527	0.030

¹ Full likelihood uses (3) for estimation.

Author Manuscript
Pseudo-likelihood uses (4) for estimation.
Single β uses (3) for estimation, setting $\beta_2 = \beta_1$

Author Manuscript

Author Manuscript

Author Manuscript

Estimates and MSE's under model (5): 3000 repetitions, $n = 75$ matched sets containing one short term survivor, three standard survivors, and one long-term survivor.

Table 4

β_1	β_2	$a^\dagger$	full likelihood ¹			pseudo-likelihood ²			single β^3			full likelihood			pseudo-likelihood		
			$\hat{\beta}_1$	MSE	$\hat{\beta}_1$	$\hat{\beta}_1$	MSE	$\hat{\beta}_1$	$\hat{\beta}$	MSE	$\hat{\beta}$	$\hat{\beta}_2$	MSE	$\hat{\beta}_2$	$\hat{\beta}_2$	MSE	$\hat{\beta}_2$
-1.00	0.50	-1	-1.028	0.021	-1.027	0.028	0.028	-0.314	0.473	0.198	0.100	0.201	0.099	0.201	0.099	0.099	0.099
-1.00	1.00	-1	-1.048	0.023	-1.032	0.030	0.030	-0.258	0.553	0.358	0.423	0.371	0.407	0.371	0.407	0.407	0.407
-1.00	1.50	-1	-1.069	0.024	-1.033	0.030	0.030	-0.215	0.617	0.488	1.036	0.517	0.981	0.517	0.981	0.981	0.981
-1.00	2.00	-1	-1.093	0.029	-1.030	0.030	0.030	-0.182	0.671	0.601	1.971	0.652	1.839	0.652	1.839	1.839	1.839
0.00	0.50	-1	0.001	0.005	0.002	0.005	0.005	0.056	0.005	0.110	0.157	0.111	0.156	0.111	0.156	0.156	0.156
0.00	1.00	-1	0.002	0.005	0.002	0.005	0.005	0.099	0.011	0.196	0.651	0.199	0.646	0.199	0.646	0.646	0.646
0.00	1.50	-1	0.003	0.005	0.003	0.005	0.005	0.136	0.020	0.269	1.520	0.275	1.506	0.275	1.506	1.506	1.506
0.00	2.00	-1	-0.002	0.006	-0.002	0.005	0.005	0.167	0.030	0.337	2.772	0.346	2.742	0.346	2.742	2.742	2.742
1.00	0.50	-1	1.046	0.026	1.028	0.029	0.029	0.411	0.349	0.032	0.226	0.032	0.226	0.032	0.226	0.226	0.226
1.00	1.00	-1	1.055	0.029	1.031	0.030	0.030	0.436	0.321	0.087	0.839	0.087	0.839	0.087	0.839	0.839	0.839
1.00	1.50	-1	1.058	0.030	1.030	0.031	0.031	0.459	0.295	0.138	1.860	0.138	1.860	0.138	1.860	1.860	1.860
1.00	2.00	-1	1.059	0.032	1.029	0.032	0.032	0.480	0.274	0.184	3.303	0.184	3.303	0.184	3.303	3.303	3.303
1.00	0.50	-2	0.926	0.024	1.027	0.029	0.029	0.500	0.254	0.204	0.097	0.203	0.097	0.203	0.097	0.097	0.097
1.00	1.00	-2	0.900	0.028	1.022	0.028	0.028	0.533	0.224	0.271	0.542	0.270	0.544	0.270	0.544	0.544	0.544
1.00	2.00	-2	0.881	0.031	1.026	0.027	0.027	0.580	0.183	0.359	2.704	0.357	2.711	0.357	2.711	2.711	2.711
1.00	3.00	-2	0.868	0.034	1.028	0.029	0.029	0.613	0.158	0.422	6.661	0.419	6.675	0.419	6.675	6.675	6.675
1.00	-3.00	2	0.222	0.610	1.016	0.079	0.079	0.094	0.834	-0.031	8.858	-0.021	8.907	-0.021	8.907	8.907	8.907
1.00	-2.00	2	0.281	0.522	1.005	0.062	0.062	0.120	0.785	-0.039	3.883	-0.027	3.919	-0.027	3.919	3.919	3.919
1.00	-1.00	2	0.406	0.359	1.022	0.040	0.040	0.197	0.654	-0.011	1.008	-0.003	1.016	-0.003	1.016	1.016	1.016
1.00	-0.50	2	0.544	0.217	1.015	0.030	0.030	0.307	0.488	0.081	0.358	0.077	0.351	0.077	0.351	0.351	0.351
1.00	1.00	2	1.289	0.121	1.025	0.030	0.030	0.225	0.602	-0.518	2.312	-0.549	2.410	-0.549	2.410	2.410	2.410
1.00	2.00	2	1.341	0.157	1.032	0.031	0.031	0.214	0.619	-0.591	6.722	-0.634	6.953	-0.634	6.953	6.953	6.953

[†] a is exponent of X in model for standard versus long-term survivors.

¹ Full likelihood uses (3) for estimation.

² Pseudo-likelihood uses (4) for estimation.

$\beta_2 = \beta_1$ for estimation, setting $\beta_2 = \beta_1$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

Standard errors and coverage of pseudo-likelihood estimation under model (5): 3000 repetitions, matched sets containing one short term survivor, three standard survivors, and one long-term survivor.

n^+	β_1	β_2	$a^{\dagger}$	empirical ^l		pseudo-likelihood		pseudo-likelihood (GEE)		empirical		pseudo-likelihood		pseudo-likelihood GEE	
				SE $\hat{\beta}_1$	SE $\hat{\beta}_1$	SE $\hat{\beta}_1$	coverage ²	SE $\hat{\beta}_1$	coverage	SE $\hat{\beta}_2$	coverage	SE $\hat{\beta}_2$	coverage	SE $\hat{\beta}_2$	coverage
75	0.25	0.25	1	0.076	0.076	0.076	0.952	0.075	0.982	0.078	0.978	0.076	0.948	0.075	0.977
	0.50	0.50	1	0.101	0.098	0.098	0.948	0.097	0.976	0.098	0.976	0.097	0.955	0.096	0.979
	0.75	0.75	1	0.132	0.127	0.127	0.952	0.126	0.975	0.128	0.975	0.127	0.957	0.126	0.979
	1.00	1.00	1	0.163	0.162	0.162	0.960	0.159	0.982	0.164	0.982	0.160	0.957	0.158	0.982
	1.50	1.50	1	0.257	0.245	0.245	0.953	0.240	0.976	0.256	0.976	0.245	0.960	0.238	0.981
	2.00	2.00	1	0.361	0.342	0.342	0.955	0.332	0.981	0.362	0.981	0.342	0.958	0.331	0.985
	2.50	2.50	1	0.513	0.460	0.460	0.958	0.443	0.983	0.489	0.983	0.455	0.956	0.437	0.980
	0.00	0.25	1	0.068	0.069	0.069	0.951	0.068	0.973	0.074	0.973	0.073	0.951	0.073	0.975
	0.00	0.50	1	0.074	0.072	0.072	0.950	0.072	0.976	0.091	0.976	0.091	0.958	0.090	0.980
	0.00	0.75	1	0.077	0.077	0.077	0.952	0.076	0.970	0.123	0.970	0.120	0.954	0.119	0.981
	0.00	1.00	1	0.082	0.081	0.081	0.950	0.081	0.973	0.168	0.973	0.161	0.958	0.158	0.979
	0.00	1.50	1	0.091	0.089	0.089	0.947	0.089	0.973	0.298	0.973	0.274	0.961	0.264	0.982
200	1.00	-3.00	2	0.292	0.286	0.286	0.949	0.284	0.963	0.182	0.963	0.112	0.000	0.176	0.000
	1.00	-2.00	2	0.295	0.286	0.286	0.948	0.284	0.954	0.178	0.954	0.112	0.000	0.176	0.000
	1.00	-1.00	2	0.245	0.244	0.244	0.953	0.242	0.975	0.165	0.975	0.107	0.000	0.165	0.000
	1.00	-0.50	2	0.170	0.170	0.170	0.956	0.168	0.978	0.132	0.978	0.091	0.000	0.132	0.001
	1.00	1.00	2	0.170	0.161	0.161	0.958	0.158	0.982	0.101	0.982	0.110	0.000	0.098	0.000
	1.00	2.00	2	0.171	0.162	0.162	0.952	0.158	0.976	0.111	0.976	0.120	0.000	0.108	0.000
	0.25	0.25	1	0.047	0.046	0.046	0.955	0.046	0.979	0.046	0.979	0.046	0.949	0.046	0.972
	0.50	0.50	1	0.061	0.059	0.059	0.949	0.059	0.976	0.060	0.976	0.059	0.948	0.059	0.978
	0.75	0.75	1	0.077	0.077	0.077	0.952	0.076	0.980	0.075	0.980	0.077	0.956	0.076	0.982
	1.00	1.00	1	0.097	0.097	0.097	0.952	0.097	0.982	0.100	0.982	0.097	0.942	0.097	0.975
	1.50	1.50	1	0.148	0.145	0.145	0.951	0.144	0.978	0.146	0.978	0.145	0.953	0.144	0.979
	2.00	2.00	1	0.203	0.202	0.202	0.957	0.200	0.981	0.210	0.981	0.202	0.945	0.201	0.977
200	2.50	2.50	1	0.283	0.269	0.269	0.950	0.266	0.977	0.277	0.977	0.268	0.953	0.264	0.979
	0.00	0.25	1	0.042	0.042	0.042	0.952	0.042	0.975	0.045	0.975	0.044	0.950	0.044	0.977

n	β_1	β_2	$a^\dagger$	empirical ¹		pseudo-likelihood		pseudo-likelihood (GEE)		empirical		pseudo-likelihood		pseudo-likelihood GEE	
				SE $\hat{\beta}_1$	SE $\hat{\beta}_1$	SE $\hat{\beta}_1$	coverage ²	SE $\hat{\beta}_1$	coverage	SE $\hat{\beta}_2$	coverage	SE $\hat{\beta}_2$	coverage	SE $\hat{\beta}_2$	coverage
0.00	0.50	1	0.044	0.044	0.044	0.957	0.044	0.044	0.977	0.057	0.942	0.055	0.942	0.055	0.975
0.00	0.75	1	0.047	0.047	0.047	0.951	0.047	0.047	0.975	0.072	0.949	0.072	0.949	0.072	0.981
0.00	1.00	1	0.050	0.050	0.049	0.952	0.049	0.049	0.971	0.096	0.952	0.095	0.952	0.095	0.980
0.00	1.50	1	0.054	0.054	0.054	0.950	0.054	0.054	0.971	0.162	0.955	0.158	0.955	0.156	0.979
1.00	-3.00	2	0.176	0.176	0.174	0.948	0.174	0.174	0.976	0.111	0.000	0.068	0.000	0.110	0.000
1.00	-2.00	2	0.174	0.174	0.174	0.951	0.174	0.174	0.876	0.111	0.000	0.068	0.000	0.110	0.000
1.00	-1.00	2	0.150	0.150	0.148	0.948	0.148	0.148	0.979	0.103	0.000	0.065	0.000	0.102	0.000
1.00	-0.50	2	0.104	0.104	0.102	0.949	0.102	0.102	0.977	0.080	0.000	0.055	0.000	0.081	0.000
1.00	1.00	2	0.100	0.100	0.096	0.950	0.096	0.096	0.974	0.062	0.000	0.067	0.000	0.060	0.000
1.00	2.00	2	0.100	0.100	0.096	0.945	0.095	0.095	0.975	0.067	0.000	0.073	0.000	0.066	0.000

¹ empirical standard errors are simulation standard deviations of the estimates

² coverage of 95% confidence intervals calculated from associated SE's

[†] n is number of matched sets.

^{*} a is exponent of X in model for standard versus long-term survivors.