Analysis of a molecular genetic neuro-oncology study with partially biased selection

REBECCA A. BETENSKY†
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA
betensky@hsph.harvard.edu

DAVID N. LOUIS
Department of Pathology and Neurosurgical Service, Massachusetts General Hospital and Harvard Medical School, Boston MA, USA

J. GREGORY CAIRNCROSS
Department of Oncology and Clinical Neurological Sciences, University of Western Ontario and London Regional Cancer Centre, London, Ontario, Canada

SUMMARY
Oligodendrogliomas are a common variant of malignant brain tumors, and are unique for their relative sensitivity to chemotherapy and better prognosis. For these reasons, the identification of an objective oligodendroglial marker has been a long sought-after goal in the field of neuro-oncology. To this end, 75 patients who received chemotherapy at the London Regional Cancer Centre between 1984 and 1999 were studied (Ino et al., Clinical Cancer Research, 7, 839–845, 2001). Of these 75 patients, 50 were initially treated with chemotherapy (the current practice) and comprise a population-based sample. The remaining 25 patients were initially treated with radiation and were included in the study only because their tumor recurred, at which time they received chemotherapy. Because this group of 25 patients included neither those radiation patients whose tumors never recurred nor those radiation patients whose tumors recurred but were not treated with chemotherapy, issues of selection bias were of concern. For this reason, the initial analysis of these data included only the 50 population-based patients. This was unsatisfying given the rarity of this disease and of genetic information on this disease and led us to question whether we could undertake an analysis that includes all of the patients.

Here we examine approaches for utilizing the entire study population, as well as the assumptions required for doing so. We illustrate that there are both costs and benefits to using the 25 selected patients.

Keywords: Missing covariate; Omitted covariate; Selection bias.

1. INTRODUCTION

Malignant gliomas are the most common type of primary human brain tumor, and comprise the bulk of most clinical neuro-oncology practices. Each year in the United States, 12,000 new cases of malignant glioma are diagnosed (CBTRUS, 1997). These lesions are associated with high morbidity and constitute one of the most expensive forms of human cancer. Oligodendroglialomas constitute up to 25% of all malignant gliomas and are clinically unique because approximately two-thirds of them are

*To whom correspondence should be addressed

responsive to chemotherapy. The histological diagnosis of anaplastic oligodendroglioma is fraught with difficulty as it may share histological features with the most aggressive and most common malignant glioma, glioblastoma. Since glioblastoma is notoriously recalcitrant to available therapies, the distinction from anaplastic oligodendroglioma at initial diagnosis is of major clinical importance. Identification of an objective oligodendrogliomal genetic marker has therefore been a primary goal in the field of neuro-oncology (Cairncross et al., 1998).

To investigate potential genetic markers, 75 patients who were treated for an anaplastic oligodendroglioma at the London Regional Cancer Centre between 1984 and 1999 were studied (Ino et al., 2001). Standard treatment for this disease changed over these 15 years; the current practice is to initially treat patients with chemotherapy, while the former practice was to initially treat with radiation therapy. Because genetic predictors of response to chemotherapy were of interest in this study, only those patients who eventually received chemotherapy were included in the study. For the 25 patients from the earlier time period who were included, chemotherapy was used to treat a tumor that had recurred following initial radiation therapy. Because this group of patients with recurrent lesions included neither those patients whose tumors never recurred nor those patients whose tumors recurred but were not treated with chemotherapy, it is not a population-based sample of all patients from the earlier time period. We refer to this group as the ‘selected’ group. The remaining 50 patients are from the later time period and constitute a population-based sample.

Current medical opinion is that initial treatment with radiation therapy, potentially followed by chemotherapy, is equivalent to initial treatment with chemotherapy, potentially followed by radiation. Thus, if we had a population-based sample of patients who were treated using either strategy, we could evaluate the effects of genetic alterations and other baseline predictors on response and survival, after accounting for delayed entry into the study. However, as only part of our sample is population-based, we cannot analyze all of the patients together without regard for the way in which they were sampled. We must be concerned with potential biases in the selection of the patients in the ‘selected’ group.

Selection bias has been studied extensively in the econometrics literature, where it is defined as non-random selection with respect to a specific outcome, controlling for explanatory variables (e.g. Heckman, 1974, 1976). As defined by Kleinbaum et al. (1982), selection bias ‘refers to a distortion in the estimate of effect resulting from the manner in which subjects are selected for the study population’. The potential selection bias in the selected subsample of the neuro-oncology study may have substantial effects on the primary analyses of the genetic alterations and their association with the clinical endpoints of interest. Table 1 lists the distributions of the important patient and tumor features at diagnosis, as well as p-values for testing the association between these features and sampling group (i.e. selected versus population-based). The two sampling groups of patients are significantly different with respect to many of these features, implying that the selected patients are significantly different from the non-selected patients with respect to these features (assuming there to be no difference between treatment strategies). It is therefore of particular concern that these patients may differ with respect to other, unmeasured, features that may be associated with response and survival.

Other authors have recognized the problem of selection bias in glioma studies (e.g. Winger et al., 1989; Florell et al., 1992; Irish et al., 1997; Barker et al., 1998; Huncharek and Muscat, 1998; Razack et al., 1998; Videtic et al., 1999). Irish et al. (1997) and Florell et al. (1992) tested for selection bias in trials of patients who received adjuvant treatment or treatment at recurrence by comparing the survival of eligible and ineligible patients in an untreated database. Improved survival for the eligible patients, even without treatment, provided evidence for selection bias. For our data, there was no explicitly defined ‘eligibility’ for the selected subsample, and we are unable to identify the ‘eligible’ and ‘ineligible’ patients among the population-based subsample. Nonetheless, we approach the problem in a similar spirit by viewing the population-based subsample as a mixture of potentially selected and non-selected patients.

In many studies in which selection bias is of concern, the selection bias can be assessed only by
Table 1. Description of neuro-oncology data

<table>
<thead>
<tr>
<th>Feature</th>
<th>Population-based group</th>
<th>Selected group</th>
<th>Fisher’s exact test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>age ≤ 45</td>
<td>26/50 (52)</td>
<td>19/25 (76)</td>
<td>0.051</td>
</tr>
<tr>
<td>enhancement at diagnosis</td>
<td>34/50 (68)</td>
<td>14/22 (64)</td>
<td>0.789</td>
</tr>
<tr>
<td>ring enhancement at diagnosis</td>
<td>9/50 (18)</td>
<td>0/22 (0)</td>
<td>0.029</td>
</tr>
<tr>
<td>1pLOH</td>
<td>29/50 (58)</td>
<td>22/25 (88)</td>
<td>0.008</td>
</tr>
<tr>
<td>19qLOH</td>
<td>33/50 (66)</td>
<td>21/25 (84)</td>
<td>0.054</td>
</tr>
<tr>
<td>10qLOH</td>
<td>12/47 (26)</td>
<td>2/22 (9)</td>
<td>0.050</td>
</tr>
<tr>
<td>PTEN mutation$^2$</td>
<td>6/49 (12)</td>
<td>0/25 (0)</td>
<td>0.076</td>
</tr>
<tr>
<td>CDKN2A deletion$^3$</td>
<td>7/50 (14)</td>
<td>5/25 (20)</td>
<td>0.740</td>
</tr>
<tr>
<td>EGFR amplification$^4$</td>
<td>7/50 (14)</td>
<td>1/25 (4)</td>
<td>0.108</td>
</tr>
<tr>
<td>TP53 mutation$^5$</td>
<td>9/50 (18)</td>
<td>4/25 (16)</td>
<td>0.750</td>
</tr>
</tbody>
</table>

1 Loss of heterozygosity using polymorphic genetic markers, which reflects allelic loss of a chromosomal arm (such as here for 1p, 19q, 10q).
2 Mutation or deletion of the PTEN gene, a tumor suppressor on chromosome 10 with a number of proposed functions.
3 Homozygous deletion of the CDKN2A gene on chromosome 9p, which encodes the p16 cell cycle control protein.
4 Increased copy number of the epidermal growth factor receptor gene, a receptor tyrosine kinase oncogene.
5 Mutation of the TP53 gene, a tumor suppressor on chromosome 17p with many functions.

imposing strong assumptions or through comparisons with external data. Kleinbaum et al. (1982) noted that the assessment of selection bias is ‘quite difficult, since it usually requires either information from another study’, in this case, a study of all patients initially treated with radiation therapy, or ‘knowledge of selection probabilities for related studies’. The use of external sources of information is not nearly as reliable as the internal comparisons afforded by a population-based subsample. The existence of this population-based subsample, in conjunction with the actually selected subsample, places us in the unique position of being able to assess and adjust for the selection bias internally to our study, allowing for the estimation of the marginal effects of genetic alterations of the tumor and other baseline predictors.

Due to our concerns about potential selection bias, we reported only on the 50 population-based patients in our initial analysis of this study (Ino et al., 2001). It was unsatisfying to completely omit the 25 actually selected patients from our analyses, especially given the rarity of this disease and of genetic information on this disease. In this paper, we explore the possibility of making better use of these selected patients and we assess the advantages and disadvantages of including them. In Section 2, we discuss the possible approaches for analysis and the corresponding required assumptions. In Section 3 we derive an upper bound for the proportion of the population with disease who are in the ‘select’ group, which is necessary for some of the analyses. In Section 4 we introduce notation and describe in detail the possible approaches to analyzing response to chemotherapy. We describe the analyses of survival in Section 5. We apply these methods and results to the neuro-oncology data in Section 6.

2. APPROACHES FOR ANALYSIS

Our primary goal is to assess the univariate relationships between radiographic and genetic features of the patients’ tumors at diagnosis and their subsequent responses to chemotherapy and survival times. In the presence of selection bias, marginal models for response and survival are derived by marginalizing full models that include selection status and genetic alteration as predictors, with respect to selection status. Our approach to the potential selection effect is to view the 50 population-based patients as a mixture...
of patients who would have been among the 25 had they received radiation therapy first (i.e. those who would have had recurrences and who would have then been treated with chemotherapy), and patients who would not have been among the 25 had they received radiation therapy first (i.e. those who would have not had recurrences, or who would have had recurrences but would not have been treated with chemotherapy). To capture this difference between patients, we define $S = 1$ for patients who were actually among the 25 selected patients, or who would have been a part of this group had they received radiation therapy first and we define $S = 0$ for patients who would not have been a part of this group had they received radiation therapy first. Clearly, we do not observe $S$ for the 50 population-based patients.

Even if we did observe $S$ for the population-based patients, the population-based patients with $S = 1$ are comparable to the actually selected patients only if there is no difference in the endpoints of interest with respect to actual treatment received and recurrence status. In fact, the current medical belief is that there is no difference between treatment with chemotherapy of recurrent tumors (previously treated with radiation) and treatment with chemotherapy of primary tumors, with respect to response to chemotherapy. There is also no difference with respect to survival, after adjustment for the delayed entry into the study. This assumption implies also that, given the baseline covariates, the population-based patients and the actually selected patients are comparable.

These assumptions of comparability leads to a few possibilities for analysis. The first is simply to restrict analysis to the population-based subsample. We refer to this approach as the restricted sample approach (see Sections 4.1 and 5.1). While this approach is valid and operationally simple, it may be inefficient as it makes no use of the 25 selected patients. A second approach, that uses all of the data, is to fit saturated models with genetic alteration and sampling group (i.e. population-based versus actually selected) as predictors. This approach allows for identification of the marginal effects of the genetic alteration, as well as the selection effect, via generalized moment estimation. We refer to this approach the moment estimation approach (see Sections 4.2 and 5.2). This is the ideal approach for the analysis of response, as both the selection effect and the effect of interest are identifiable without any additional assumptions. The drawbacks of this approach are that the model is not estimable if $S$ and response are highly associated and that it is not applicable to the analysis of survival. A third approach, that also uses all of the data, is to use maximum likelihood to fit full models with selection status and genetic alteration as predictors. Of course, selection status is missing for the population-based sample. These models are then marginalized with respect to selection status to obtain models with genetic alteration as the only predictor. We refer to this approach as the maximum likelihood approach (see Sections 4.3, 5.3). The appeal of this approach is that it uses all of the data. Its drawbacks are that it requires full specification of the likelihood, it requires identification of the probability of selection, $P(S = 1)$, and it is unstable for large $P(S = 1)$. We describe these approaches in more detail and evaluate them in Sections 4 and 5.

3. Bounding $P(S = 1)$

The maximum likelihood approach described above requires knowledge of the proportion of the earlier patients who were initially treated with radiation therapy who recurred and received chemotherapy, $P(S = 1)$, for identification of the selection effect. Although in some applications the probability of selection may be known, for the neuro-oncology data it was not known. Although we cannot estimate $P(S = 1)$ from our data, we can compute an upper bound for it.

Let $X_1, \ldots, X_J$ be indicators for the $J$ baseline features of interest. To bound $P(S = 1)$, note that

$$P(S = 1) = \frac{P(X_j = x)}{P(X_j = x \mid S = 1)} P(S = 1 \mid X_j = x) \quad \text{for } j = 1, \ldots, J, \text{ and } x \in \{0, 1\}$$

$$\leq \min_{j,x} \frac{P(X_j = x)}{P(X_j = x \mid S = 1)}. \quad (1)$$
This upper bound is identifiable from the data: \( P(X_j = x) \) can be consistently estimated from the population-based group and \( P(X_j = x | S = 1) \) can be consistently estimated from the actually selected group. The bound will be equal to \( P(S = 1) \) if there is at least one baseline feature that determines selection, i.e. \( P(S = 1 | X_k = 1) = 1 \). In this case, \( P(S = 1) = P(X_k = 1) / P(X_k = 1 | S = 1) \), which by (1), implies that \( P(S = 1) = \min_{i,x} P(X_j = x) / P(X_j = x | S = 1) \). If there is no association between selection and any of the baseline features, this upper bound is theoretically equal to one and is uninformative about \( P(S = 1) \). A bootstrap percentile upper confidence bound (Efron and Tibshirani, 1993) can be computed for the upper bound by resampling separately from the actually selected and population-based groups.

For the neuro-oncology data, the upper bound for \( P(S = 1) \) is 0.66. Because this upper bound was computed over a small subset of all possible sets (i.e. all sets in the \( \sigma \)-algebra generated by the \( J \) indicators), it is is unlikely to be sharp. The bootstrap percentile upper 95% confidence limit (based on 2000 bootstrap samples) for the upper bound is 0.70. As an upper confidence bound on a crude upper bound, this value is quite conservative. Based on their experiences with the patients, and without knowledge of these estimates, the physicians who conducted the study guessed that the selected group of 25 was drawn from a larger group of 35–40 patients and thus \( P(S = 1) \) is between 25/40 (0.63) and 25/35 (0.71).

4. Analysis of binary response

One important measure of treatment efficacy is the rate of neuroradiological response, as defined by a decrease in tumor size of more than 50%. Let \( Y \) indicate response to chemotherapy and let \( R \) indicate membership in the population-based subsample (versus the actually selected subsample). We assume that the covariate, \( X \), and selection status, \( S \), explain the response \( Y \) through the following saturated logistic regression model:

\[
\mu_{xs} = P(Y = 1 | X = x, S = s) = \frac{\exp(b_0 + b_1 x + b_2 s + b_3 xs)}{1 + \exp(b_0 + b_1 x + b_2 s + b_3 xs)}, \quad (2)
\]

Underlying this model is the assumption of comparability of the actually selected patients and the potentially selected patients (i.e. population-based patients with \( S = 1 \)) with respect to response. We introduce the following additional notation: \( p = P(S = 1) \), \( \tau = P(X = 1) \), \( \tau_1 = P(X = 1 | S = 1) \), \( \pi_{xs} = P(S = s | X = x) \), \( \mu_s = \mu_s x \pi_1 + \mu_s 0 \pi_0 \), and \( q = P(R = 0) \). The primary goal of our analysis is to compare \( \mu_0 \) and \( \mu_1 \) for each covariate, \( X \). A secondary objective is to assess whether there is a selection effect, i.e. whether \( \beta_2 \) in model (2) is significantly different from zero.

4.1 Restricted sample approach

We first consider analysis restricted to the population-based subsample. This analysis proceeds by fitting the model

\[
\text{logit} \ P(Y = 1 | X = x) = \beta_0^* + \beta_1^* x. \quad (3)
\]

With analysis is restricted to the population-based group, valid estimates of \( \mu_0 \) and \( \mu_1 \) result, as \( \beta_0^* = \log[\mu_0/(1 - \mu_0)] \) and \( \beta_1^* = \log[(1 - \mu_0)\mu_1/[\mu_0(1 - \mu_1)]] \). However, if this analysis is extended to the entire sample, without regard for selection status, then \( \mu_0 \) and \( \mu_1 \) are not identified. Instead, \( \beta_0^* \) and \( \beta_1^* \) are functions of \( \mu_0, \mu_1, \mu_01, \) and \( \mu_11 \).
4.2 Moment estimation approach

A natural question to ask with regard to these data is whether we can achieve our objectives through adjustment for $R$, the completely observed sampling group indicator, instead of $S$, the incompletely observed selection indicator. The rationale for this approach is that $R = 1$ represents a mixture of patients with $S = 0$ and $S = 1$, while $R = 0$ represents only patients with $S = 1$, and so the probability of response given $R = 0$ is equal to that given $R = 1$ if and only if the probability of response given $S = 1$ is equal to that given $S = 0$. That is, $P(Y = y \mid X = x, R = 1) = P(Y = y \mid X = x, R = 0)$ if and only if $\mu_{s0} = \mu_{s10} + \mu_{s1} \tau_{1x}$, if and only if $\mu_{s0} = \mu_{s1}$, implying that $P(Y = y \mid X = x, S = 0) = P(Y = y \mid X = x, S = 1)$. It turns out that this approach is feasible in situations in which selection is not too highly predictive of response.

Assuming model (2) to be true, and assuming that $Y$ is conditionally independent of $R$ given $S$ and $X$, it follows that

$$
\text{logit } P(Y = 1 \mid X = x, R = r) = \log \frac{\mu_{01}}{1 - \mu_{01}} + \log \frac{\mu_{11}(1 - \mu_{11})}{(1 - \mu_{01})\mu_{01}}x + \log \frac{\mu_{0}(1 - \mu_{01})r}{(1 - \mu_{0})\mu_{01}} + \log \frac{\mu_{1} \mu_{01}(1 - \mu_{0})(1 - \mu_{11})}{(1 - \mu_{1})(1 - \mu_{01})\mu_{01}} x r.
$$

Thus, it is apparent that this interaction model enables identification of $\mu_{0}$ and $\mu_{1}$, the quantities of interest. Further, we can use this model to test for a selection effect, namely that $\mu_{0} = \mu_{01}$ and $\mu_{1} = \mu_{10}$ (implying that $\mu_{0} = \mu_{01}$ and $\mu_{10} = \mu_{11}$). The advantages of this approach over the maximum likelihood approach are that it is operationally simple and it does not require knowledge of the selection probability, $P(S = 1)$. The drawback of this approach is that if selection is highly predictive of response, as it is for the neuro-oncology data (23 responded out of 24 evaluable for response from the $R = 0$ group), then there will be insufficient separation in the data to allow for the fitting of this model. This, of course, may be a problem for the maximum likelihood approach as well, although with larger numbers and with missing $S$, it may be less severe.

4.3 Maximum likelihood approach

The maximum likelihood approach treats the selection status indicator, $S$, as missing for the 50 patients in the population-based group, and fits the correct model (2) using maximum likelihood estimation. This is feasible subject to knowledge of the selection probability, $p = P(S = 1)$, for which an upper bound is available. We do not need to assume that $S$ is missing at random, as do most approaches for regression analysis with missing covariates (e.g. Vach and Schumacher, 1993). In fact, it is clear that $S$ is not missing at random since conditional on $S = 0$, it is missing with probability one and conditional on $S = 1$, it is missing with probability strictly less than one.

We assume that $Y$ is conditionally independent of $R$, the indicator for membership in the population-based sample, given $S$ and $X$, and that $X$ is independent of $R$, given $S$. The contribution to the likelihood by patients from the actually selected subsample is $P(R = 0, Y = y, X = x, S = 1)$, which is equal to $q\mu_{y1}(1 - \mu_{1})(1 - \tau_{1})(1 - r_{1})^{1-x}$. The contribution to the likelihood by patients from the population-based subsample is $P(R = 1, Y = y, X = x)$, which is equal to $(1 - q)\mu_{y1}(1 - \mu_{1})(1 - \tau_{1})(1 - r_{1})^{1-x}$. If $p = 1$, this contribution reduces to $(1 - q)\mu_{y1}(1 - \mu_{1})(1 - \tau_{1})(1 - r_{1})^{1-x}$, in which case $\beta_{2}$ is nonidentifiable, as expected. Thus, this approach will become less and less stable as $p$ increases to one. We maximize this likelihood with respect to $\beta_{0}, \beta_{1}, \beta_{2}, q, \tau_{1}$, and $r$, and invert the information matrix to obtain the asymptotic covariance matrix. This approach allows for estimation of $\mu_{0}$ and $\mu_{1}$ and the selection effect, $\beta_{2}$.
5. Analysis of Survival

The other important measure of treatment efficacy is overall survival from diagnosis. Complicating the analysis of the survival data from the neuro-oncology data is delayed entry into the study. That is, patients were included in the study only if they had received chemotherapy, and thus their survival times are left-truncated by their times from diagnosis to start of chemotherapy. We assume that the survival times and times to chemotherapy are independent, given that the survival times exceed the times to chemotherapy. Let \( T \) denote survival time, let \( C \) denote censoring time, let \( Y = \min(T, C) \) denote observed survival time, and let \( \delta \) indicate whether failure was observed. Further, let \( Z \) denote time to start of chemotherapy.

Analogous to our model for response, we assume that the covariate, \( X \), and selection, \( S \), explain the survival time \( T \) through the proportional hazards regression model:

\[
\lambda_{xs}(t) = \lambda(t) \exp(\beta_1 x + \beta_2 s + \beta_3 xs). \tag{5}
\]

Letting \( \Lambda(t) = \int_0^t \lambda(u) \, du \), it follows that

\[
S_{xs}(t) = P(T > t \mid X = x, S = s) = \exp\left[ -\exp(\beta_1 x + \beta_2 s + \beta_3 xs) \Lambda(t) \right]
\]

and

\[
f_{xs}(t) = P(T \in dt \mid X = x, S = s) = \lambda_{xs}(t)S_{xs}(t).
\]

Thus, the survivor function given \( X = x \) is

\[
S_x(t) = S_{x0}(t)\pi_0 x + S_{x1}(t)\pi_1 x,
\]

the density of \( T \) given \( X = x \) is

\[
f_x(t) = f_{x0}(t)\pi_0 x + f_{x1}(t)\pi_1 x,
\]

and the hazard function given \( X = x \) is

\[
\lambda_x(t) = \lambda_0(t)S_x(t)/S_{xs}(t).
\]

5.1 Restricted sample approach

Assuming model (5) to be correct, we can express the marginal hazard function as

\[
\lambda_x(t) = \lambda_0(t) \exp(\beta_1 g(t)x),
\]

where \( \exp[\beta_1 g(t)] \) is simply the true marginal hazard ratio, \( \lambda_1(t)/\lambda_0(t) \). This implies that under model (5), and non-exponential hazards, we cannot simply apply a proportional hazards regression analysis to the population-based sample, as the propotional hazards assumption does not hold for the marginalized hazard functions. Thus, this approach necessarily entails assuming exponential hazard functions or deriving a nonparametric estimate of the marginal hazard functions. Alternatively, a proportional hazards model could be fit to the data, entailing a loss of estimation efficiency for \( \beta_1 \) (e.g. Lagakos and Schoenfeld, 1984).

5.2 Moment estimation approach

Because the marginalized hazard function (with respect to \( S \)) that is conditional on \( X \) is not a weighted average of the hazard functions for \( S = 0 \) and for \( S = 1 \), it is not possible to derive a proportional hazards model given \( X \) and \( R \), assuming model (5) to be true. Thus, in contrast to the analysis of response, this approach cannot be used to identify the effect of interest in the analysis of survival.

5.3 Maximum likelihood approach

Assuming no treatment effect, we can use maximum likelihood to fit the model (5), taking into account the missing values of \( S \) for the 50 patients in the population-based subsample. In so doing, we must additionally specify the baseline hazard function in (5). Similar to the analysis of response, the contribution to the
likelihood by patients from the actually selected subsample is given by $q \lambda_{x_1}(y) S_{x_1}(z) \tau_x^x (1 - \tau_1)^{1-x}$ and the contribution to the likelihood by patients from the population-based subsample is given by $(1 - q) \lambda_{x}(y) S_{x}(z) \tau_x^x (1 - \tau)^{1-x}$. Note that the delayed entry is taken into account in that the survivor functions are conditional on failure occurring after initiation of chemotherapy (i.e., $Z$). Again, for large $p < 1$, this procedure will be unstable, with $\beta_2$ being nonidentifiable for $p = 1$. This approach enables direct analysis of the effects of selection via $\beta_2$.

6. APPLICATION TO NEURO-ONCOLOGY DATA

In this section, we apply the approaches outlined in the previous sections to the neuro-oncology data. We evaluate the differences in probability of response and in 5-year survival as functions of baseline genetic alterations using the restricted sample and maximum likelihood approaches. Due to the small sample size of this study, we were unable to fit the full interaction models for response and survival for many values of interest. Thus, we fit instead the models without the interaction terms (between baseline covariate of interest and selection). Underlying this simplification is the assumption that selection did not alter the relationships between response or survival and the baseline covariate. Further, the maximum likelihood approach depends on the parameter $P(S = 1)$, about which the data only provide an upper bound of 0.66, as derived in Section 3. Under these assumptions, we use the maximum likelihood approach, for $P(S = 1)$ ranging from 0.1 to 0.6, to derive the marginal relationships between response or survival and the baseline covariates and to evaluate the importance of adjusting for selection in the analyses of these data.

6.1 Analysis of binary response

There were 62 patients who were evaluable for radiographic response, 48 of whom responded. Of the 24 evaluable patients from the actually selected group of 25, 23 responded. For this reason, it is not possible to use the moment estimation approach to assess the effect of selection or to estimate the effects of the baseline features.

Figure 1 contains the estimates and 95% confidence intervals for the difference in probabilities of response with respect to presence or absence of the predictor of interest and as a function of the method of estimation. The estimate labeled ‘(50)’ is from the restricted sample approach using the 50 population-based patients. The remaining estimates are from the maximum likelihood approach and are listed as a function of $P(S = 1)$. When $P(S = 1)$ is small, the maximum likelihood approach is preferable as its inclusion of all 75 patients appears to reduce the width of the confidence intervals of the probability differences. In fact, for 19qLOH and TP53 mutation and for small $P(S = 1)$, the use of the 50 population-based patients leads to confidence intervals that include zero, whereas the use of all 75 patients via maximum likelihood leads to confidence intervals that exclude zero. In contrast, for large $P(S = 1)$, the restricted sample analysis of the 50 population-based patients is preferable. In this case, maximum likelihood is unstable, as seen with the wide confidence intervals.

Figure 2 displays the estimated values of $\beta_2$ from the maximum likelihood approach, along with 95% confidence limits. There are some values of $P(S = 1)$ for which no estimates are displayed; at these values a maximum likelihood did not exist. Similarly, the very wide confidence intervals for $\beta_2$ for some values of $P(S = 1)$ indicate insufficient data for reliable estimation. There is evidence of marginally significant selection effects for the smaller values of $P(S = 1)$ for the analyses of all predictors, with the exception of ring enhancement.
Fig. 1. Differences in probabilities (and 95% confidence intervals) of response (blue) and 5-year survival (red) for $X_j = 1$ and $X_j = 0$ as a function of $P(S = 1)$ (results plotted at ‘(50)’ are based on the 50 population-based patients only, other results are based on all patients via maximum likelihood)
Fig. 2. Estimates and confidence limits for $\beta_2$ from logistic regression models (blue) and hazard models (red) as a function of $P(S = 1)$.
6.2 Analysis of survival

As the logarithm of the Kaplan–Meier estimate of the survivor function appeared to be nearly linear over most of the time range, we assumed a constant baseline hazard function in our analyses of survival. Further, we assumed an exponential censoring distribution with mean 167 months, as in the actual data.

Figure 1 displays the differences in 5-year survival and 95% confidence intervals with respect to presence or absence of the predictor of interest and as a function of the method of estimation. The principal advantage to using the maximum likelihood approach in this case is in providing slightly narrower confidence intervals for the survival differences. This is particularly true for the analysis of TP53 mutation; using only the 50 population-based patients the confidence interval is (−0.67, 0.00), whereas it is (−0.61, −0.12) at \( P(S = 1) = 0.4 \), for example.

Figure 2 displays the estimated values of \( \beta_2 \) from the maximum likelihood approach, along with 95% confidence limits. There is strong evidence for selection effects in the analyses of all predictors except for enhancement, across the range of \( P(S = 1) \) and nearly independent of \( P(S = 1) \). In contrast to those for the analysis of response, all of the confidence intervals displayed here are narrow. This is due to the greater power of the procedure with the continuous endpoint of survival versus the dichotomous endpoint of response.

6.3 Discussion

While in many studies Ellenberg’s (1994) description of selection bias as ‘difficult or impossible to quantify’ is accurate, the situation in a study of oligodendrogliomas was not so grim. This was because the sampling scheme of the study afforded us the opportunity to adjust for and estimate the selection effect. In particular, this was enabled through the presence of the population-based subsample, in conjunction with the actually selected subsample. Additionally, we required the one key assumption that there was no effect of treatment strategy on response or survival, which reflects the current medical belief. The approaches that we have considered are applicable more generally to any study in which there is a subsample of population-based participants and a subsample of participants who entered the study through some selection mechanism and in which the same treatment is applied to both subsamples.

Following the current medical understanding that the two treatment strategies are equivalent with respect to response and survival, we were able to take a maximum likelihood approach that makes use of the 25 selected patients. While this approach would seemingly increase the estimation efficiency of the effects of baseline predictors, we found that other factors may negate this advantage. For one, the maximum likelihood approach depends on the generally nonidentifiable probability of selection. Further, its stability, especially in the analysis of a dichotomous outcome such as response, suffers from high associations of selection with response and/or the baseline predictor. The level of associations, together with the small sample size, also forced us to assume there to be no interaction between selection and the baseline predictors in the models for response and survival. Lastly, in adjusting for the missing selection status, it requires a full likelihood analysis of survival. The alternative restricted sample approach has the advantages of being operationally simple and not relying on any nonidentifiable assumptions regarding selection. Its disadvantages are that it does not use the 25 selected patients and it leads to a misspecified proportional hazards analysis of survival, except under the restrictive assumption of a constant hazard function. Nonetheless, for the neuro-oncology study, given the small sample size and given that \( P(S = 1) \) is likely to be close to 0.6, where the maximum likelihood approach is unstable, we prefer the restricted sample approach for analysis. For a larger study, however, the efficiency gains of the maximum likelihood approach, especially if there is some external information available regarding the probability of selection, may be more substantial and thereby recommend it for analysis.
ACKNOWLEDGEMENTS

This research was supported in part by NIH grants CA75971 and CA57683. We thank the referees for helping to clarify the goals and exposition of this paper.

REFERENCES


[Received July 30, 2001; revised December 28, 2001; accepted for publication February 1, 2002]