

Testing and Adjusting for Selection Bias in a Partially Retrospective Molecular Genetic Neuro-Oncology Study

Rebecca A. Betensky¹, David N. Louis² and J. Gregory Cairncross³

¹ Department of Biostatistics, Harvard School of Public Health,
655 Huntington Avenue, Boston, MA 02115,
e-mail: betensky@hsph.harvard.edu, telephone: 617-432-2821, fax: 617-432-2832

² Department of Pathology and Neurosurgical Service, Massachusetts General Hospital and
Harvard Medical School, Boston MA

³ Department of Medical and Experimental Oncology, London Regional Cancer Centre and
University of Western Ontario, London, Ontario

May 23, 2000

Testing and Adjusting for Selection Bias in a Partially Retrospective Molecular Genetic Neuro-Oncology Study

ABSTRACT

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., 2000). For 50 of the patients, chemotherapy was planned from diagnosis, whereas for the remaining 25 patients, chemotherapy was not planned at the outset, but was used to treat a tumor that had recurred following initial radiation therapy. Because the group of 25 patients included neither those patients whose tumors never recurred nor those patients whose tumors recurred but were not treated with chemotherapy, issues of selection bias were of concern. We propose a test for selection bias based on a minimally selected p -value, analyzed via a refined Bonferroni correction derived by Worsley (1982). The test relies heavily on the presence of the randomly selected subsample of the study. We find there to be significant evidence of selection bias in the neuro-oncology study. Further, we propose estimators for the overall probability of selection and for the probability of selection given baseline predictors, and find that much of the selection bias can be explained by one baseline genetic feature. Lastly, we show that it is essential to adjust comparisons of treatment strategy for the selection bias; naive comparisons of response and of survival are significant, whereas properly adjusted comparisons are not. We assess the performance of the test and estimators in a simulation study.

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. Unfortunately, using current conventional diagnostic approaches, therapeutically sensitive variants are difficult to identify in a consistent and reproducible way. As a result, patients may receive inappropriate therapies, with resultant additional morbidity and loss of health care resources.

Histopathological examination remains the gold standard of classification for malignant

⁰ *Key words:* Bonferroni correction, spanning tree, glioma

gliomas, since these neoplasms are classified and graded microscopically. Oligodendrogliomas constitute up to 25% of all malignant gliomas and are clinically unique because approximately two-thirds of them are responsive to chemotherapy. The histological diagnosis of oligodendroglioma, and particularly of anaplastic oligodendroglioma, is fraught with difficulty. In particular, anaplastic oligodendrogliomas 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 is of major clinical importance. Identification of an objective oligodendrogliomal “marker” has therefore been a primary goal in the field of neuro-oncology (Cairncross et al., 1998).

To investigate this question, 75 patients who received chemotherapy for a newly diagnosed or recurrent anaplastic oligodendroglioma at the London Regional Cancer Centre between 1984 and 1999 were studied (Ino et al., 2000). For 50 of the patients, chemotherapy was used as an integral part of an overall patient management strategy from diagnosis, whereas for the remaining 25 patients, chemotherapy was not planned at the outset, but was used to treat a tumor that had recurred following initial radiation therapy. Because the group of 25 patients with recurrent lesions included neither those patients whose tumors never recurred nor those patients whose tumors recurred but were not treated with chemotherapy, issues of selection bias are of concern. In particular, the group of 50 patients is a random sample of all patients for whom chemotherapy, potentially followed by radiation therapy, is planned from diagnosis, whereas the group of 25 patients is a non-random sample of all patients for whom radiation therapy, potentially followed by chemotherapy, is planned from diagnosis. In this regard, the study is partially retrospective.

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.” Williams (1978) noted that the effects of selection bias can be serious in magnitude, they can be subtle, and they can change apparent relationships. Thus, the potential selection bias in the retrospective 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 response, duration of response, and survival. Further, it may impact the comparison of the two treatment strategies of initial treatment with radiation therapy versus initial treatment with chemotherapy. These complications aside, the selection bias itself is of interest, as it is the result of a largely unknown mechanism by which patients who were initially treated with radiation, whose lesions

recur, whose physicians recommend treatment with chemotherapy, and who agree to undergo this treatment are selected. Identification of the baseline genetic and clinical predictors of this selection may add to the understanding of the determinants of physician recommendations and patient decisions.

Because a subset of the data is a random sample (i.e., the 50 patients), we are in the unique position of being able to learn about the nature of the selection bias in ways that would not be possible without the random subsample. In most studies in which selection bias is of concern, the selection bias can be assessed only 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 random subsample.

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 pre-defined “eligibility” for the retrospective sample, and so we could not carry out this kind of an analysis. Instead, we test for whether selection depends on baseline features known to be associated with the outcomes of interest.

In Section 2, we propose an exact test for selection bias based on a minimally selected p -value. We analyze the test using a refined Bonferroni correction based on the minimal spanning tree derived by Worsley (1982). In Section 3, we provide an upper bound for the probability of selection and estimates of the association between baseline features and selection. In Section 4, we compare the two treatment strategies with respect to response to chemotherapy and survival, with adjustment for the selection bias. We demonstrate that these adjustments are essential; the naive comparisons are significant, whereas the adjusted comparisons are not. We assess the performance of the test and estimators via several simulation studies in Section 5.

2. Testing for selection bias

Suppose there are J baseline features, all dichotomous, whose occurrences for a given patient are indicated by the Bernoulli random variables, X_1, \dots, X_J . Let S denote the event that the patient is a part of the retrospective sample, and let R denote the event that the patient is a part of the random sample. Under the null hypothesis of no selection bias, S is independent of all Borel measurable sets contained in \mathcal{A} , the σ -algebra generated by the random variables X_1, \dots, X_J . In particular, $P(S|A_j) = P(S|A_k)$ for all $A_j, A_k \in \mathcal{A}$. Writing

$$P(S|A_j) = \frac{P(A_j|S)P(S)}{P(A_j)},$$

it is apparent that $P(S|A_j)$ is non-identifiable from the data due to its dependence on the overall probability of selection, $P(S)$. The other two components of this probability are estimable from the data: $P(A_j|S)$ is estimated from the retrospective sample and $P(A_j) = P(A_j|R)$ is estimated from the random sample. Thus, the ratio of the probabilities, $P(S|A_j)/P(S|A_k)$ is estimable from the data and we rewrite the null hypothesis of no selection bias in terms of these ratios as

$$H_0 : \frac{P(S|A_j)}{P(S|A_k)} = 1, \text{ for all } A_j, A_k \in \mathcal{A},$$

or, equivalently, as

$$H_0 : \frac{P(A_j|S)P(A_k|R)}{P(A_k|S)P(A_j|R)} = 1, \text{ for all } A_j, A_k \in \mathcal{A}. \quad (1)$$

Note that for sets A_j and A_k with $A_j \cap A_k = \emptyset$, (1) is simply the usual odds-ratio null hypothesis for independence in a 2×2 table. Under H_0 , the statistic

$$T_{jk} = \frac{\log[\hat{P}(S|A_j)/\hat{P}(S|A_k)]}{\sqrt{\text{Var}(\log[\hat{P}(S|A_j)/\hat{P}(S|A_k)])}}$$

is approximately distributed as a standard normal random variable, where the variance can be approximated using the delta method, n_S is the number of patients in the retrospectively sampled group, and n_R is the number of patients in the randomly sampled group.

A natural test statistic for testing (1) is the maximal statistic, $\max_{1 \leq j < k \leq N} |T_{jk}|$, where $N = n(n-1)/2$, and n is the number of Borel measurable sets contained in \mathcal{A} . As N is likely to be an impracticably large number, especially for computing the Bonferroni bounds on the p -value described below, we consider a restricted null hypothesis that leads to manageable computing. In particular, we test

$$H_0 : \frac{P(X_j = 1|S)P(X_j = 0|R)}{P(X_j = 0|S)P(X_j = 1|R)} = 1, j = 1, \dots, J \quad (2)$$

using the statistic

$$T_j = \frac{\log[P(S|X_j = 1)/P(S|X_j = 0)]}{\sqrt{(\frac{1}{n_S} + \frac{1}{n_R})\frac{1}{p_j(1-p_j)}}}$$

where

$$p_j = \frac{n_S P(X_j = 1|S) + n_R P(X_j = 1|R)}{n_S + n_R}.$$

Maximal statistics, such as $\max_j |T_j|$, have been used in the analysis of a sequence of $k \times 2$ tables in which the columns of the table are created by dichotomizing a continuous outcome on the basis of a sequence of thresholds and a maximal chi-square or odds-ratio statistic is used to test for independence (e.g., Miller and Siegmund, 1982, Halpern, 1982, Betensky and Rabinowitz, 1999). They have been used also in the analysis of the transmission/disequilibrium test for markers with multiple alleles (Betensky and Rabinowitz, 2000). Unlike the individual T_j , $\max_j |T_j|$ does not follow a standard normal distribution, as it is maximally selected. One approach to obtaining a correct p -value on the basis of this statistic is to use a Bonferroni correction. Letting t denote the observed value of $\max_j |T_j|$, the p -value, $P(\max_j |T_j| \geq t)$, is contained in the interval,

$$\left(\sum_{j=1}^J P(|T_j| \geq t) - \sum_{j=1}^{J-1} \sum_{k=j+1}^J P(|T_j| \geq t, |T_k| \geq t), \sum_{j=1}^J P(|T_j| \geq t) \right), \quad (3)$$

where the lower bound was derived by Feller (1968).

The crude Bonferroni upper bound can be inaccurate if J is large, or if the T_j are highly correlated. A more accurate upper bound was derived by Worsley (1982) based on simple results in graph theory. Let A_j be the event that $|T_j| \geq t$ and represent the events $\{A_j, j = 1, \dots, J\}$ as vertices of a graph, G . Let $\{e_{jk}\}$ represent the edges of the tree, namely the straight line segments joining the vertices A_j and A_k . Associated with each edge, e_{jk} , is a length, l_{jk} . A spanning tree is defined as a set of edges joining pairs of vertices such that no closed loops occur, each vertex is visited by at least one edge, and the tree is connected. A maximal spanning tree, M , is a spanning tree of maximum length, where lengths are assigned to the edges. See Gower and Ross (1969) for more details and for an algorithm for finding the maximal spanning tree. Defining the lengths of the edges to be the probabilities of the intersections of their associated node events, i.e., $l_{ij} = P(A_i \cap A_j)$, Worsley's (1982) improved upper bound is given by

$$\sum_{j=1}^J P(|T_j| \geq t) - \sum_{(j,k): e_{jk} \in M} P(|T_j| \geq t, |T_k| \geq t). \quad (4)$$

Note that the computation of this bound would be prohibitive if the statistic $\max_{j,k} |T_{jk}|$ for testing (1) were used.

Table 1 lists the 11 baseline features that were measured in the neuro-oncology study and that were found to be significantly associated with at least one of the primary outcomes of response, duration of response, or survival (Ino et al., 2000). These include clinical features such as age, gender, and radiographic contrast enhancement and ring-enhancement, and genetic features such as allelic loss of various chromosome arms and alterations of various genes. Explanations of the genetic features are listed in footnotes to the table. Table 1 lists also the value of T_j for each feature, X_j . The observed value of $\max_j |T_j|$ is infinity, due to there being no ring enhancement and no PTEN mutations in the retrospectively sampled group.

The positive probability of an infinite value for T_j suggests that the normal approximation may not be appropriate for calculating accurate p -values. In fact, with a low overall response rate, such as there is for ring enhancement, PTEN mutation, and EGFR amplification, the probability that $|T_j|$ is infinite is nonnegligible. For this reason, we used exact calculations, conditioning on the margins of the corresponding 2×2 tables, to compute the probabilities that comprise the Bonferroni bounds. In particular,

$$P(|T_j| \geq t) = \sum_{k=0}^{o_j} \mathbf{1}_{\{|T_j(k)| \geq t\}} \frac{\binom{n_S}{k} \binom{n_R}{o_j - k}}{\binom{n_S + n_R}{o_j}}$$

and

$$P(|T_j| \geq t, |T_k| \geq t) = \sum_{a=0}^{m_j} \sum_{b=0}^{m_k} \sum_{c=0}^{m_{jk}} \mathbf{1}_{\{|T_j(a+c)| \geq t\}} \mathbf{1}_{\{|T_k(b+c)| \geq t\}} \frac{\binom{n_S}{a \quad b \quad c} \binom{n_R}{m_j - a \quad m_k - b \quad m_{jk} - c}}{\binom{n_S + n_R}{m_j \quad m_k \quad m_{jk}}},$$

where $\mathbf{1}_A$ is the indicator function of the event A , $T_j(k)$ is the value of the normalized log odds-ratio based on observing k patients with $X_j = 1$ in the retrospectively sampled group, o_j is the number of subjects with $X_j = 1$, m_j is the number of subjects with $X_j = 1$ and $X_k = 0$, m_k is the number of subjects with $X_j = 0$ and $X_k = 1$, and m_{jk} is the number of subjects with $X_j = 1$ and $X_k = 1$ (notice that $o_j = m_j + m_{jk}$).

For the neuro-oncology study, the p -value is contained in (0.104,0.166), with a crude Bonferroni upper bound of 0.189. Thus, Worsley's (1982) correction leads to considerable improvement in this case. This range for the p -value suggests that there is not strong evidence for selection bias in these data. However, it is possible that there is a dependence between selection

and an unmeasured feature, such as the physician's perception of the relative value of the most aggressive treatment versus the best quality of life to the patient. Another possibility for this unexpectedly large p -value is that this test statistic does not best measure extreme deviations from an odds-ratio of one. This is because the log odds-ratio is infinite if there is a zero in one of the cells of the table, without regard for the overall response rate. Thus, even if $X_j = 1$ for just a single patient among all $n_S + n_R$ patients, $|T_j|$ would be infinite because either $P(X_j = 1|R) = 0$ or $P(X_j = 1|S) = 0$.

For this reason, we measured extreme deviations from an odds-ratio of one via the exact p -values of the $|T_j|$. In particular, we used the minimally selected statistic, $\min_j P_j$, where $P_j = P(|T_j| \geq t_j)$ and t_j is the observed value of $|T_j|$. Table 1 lists the values of P_j for the 11 baseline features that were measured in the neuro-oncology study. The p -value based on this test statistic is $P(\min_j P_j \leq p)$, where p is the observed value of $\min_j P_j$. It can be bounded as in (3) and (4), where

$$P(P_j \leq p) = \sum_{k=0}^{o_j} 1_{\{P_j(k) \leq p\}} \frac{\binom{n_S}{k} \binom{n_R}{o_j - k}}{\binom{n_S + n_R}{o_j}}$$

and

$$P(P_j \leq p, P_k \leq p) = \sum_{a=0}^{m_j} \sum_{b=0}^{m_k} \sum_{c=0}^{m_{jk}} 1_{\{P_j(a+c) \leq p\}} 1_{\{P_k(b+c) \leq p\}} \frac{\binom{n_S}{a \quad b \quad c} \binom{n_R}{m_j - a \quad m_k - b \quad m_{jk} - c}}{\binom{n_S + n_R}{m_j \quad m_k \quad m_{jk}}},$$

and $P_j(k)$ is the exact p -value based on observing k patients with $X_j = 1$ in the retrospectively sampled group, given that there are o_j patients with $X_j = 1$ in both groups combined.

Indeed, this test is more powerful than that based on the normalized log odds-ratios for the neuro-oncology study. As seen in Table 1, the value of $\min_j P_j$ is 0.008, which occurs for the baseline feature of allelic loss of chromosomal arm 1p (1pLOH). As this p -value was minimally selected, it would be incorrect to report 0.008 as the p -value for the test of selection bias. Instead, the Bonferroni interval for the true p -value, correcting for the multiple comparisons, is (0.037,0.038). The crude Bonferroni upper bound is 0.039, which is almost identical to the refined upper bound based on the maximal spanning tree. Thus, based on the minimum exact p -value, it is likely that selection bias is present in the retrospective subsample of the neuro-

oncology study and that this bias is primarily due to the relatively high probability of selection among patients with allelic loss of 1p.

3. Quantifying the selection process

Having established that selection bias was operative in the retrospective sampling for this study, it is of interest to study the nature of the selection. Again, the presence of the random subsample enables us to do this. We first propose an upper bound for the overall probability of selection, $P(S)$, that is informative in the case of selection bias. We then explore the baseline correlates of selection.

3.1 Estimating $P(S)$

To bound $P(S)$, note that

$$\begin{aligned} P(S) &= \frac{P(A_j)}{P(A_j|S)}P(S|A_j) \text{ for all } A_j \in \mathcal{A} \\ &\leq \min_j \frac{P(A_j)}{P(A_j|S)} \\ &= \min_j \frac{P(A_j|R)}{P(A_j|S)}. \end{aligned} \tag{5}$$

This bound will be equal to $P(S)$ if there is selection bias in the sense that for at least one function of the baseline features, A_k , all patients having A_k are selected, i.e., $P(S|A_k) = 1$. In this case, $P(S) = P(A_k)/P(A_k|S)$, which by (5), implies that $P(S) = \min_j P(A_j)/P(A_j|S)$. If there is no selection bias, this upper bound is theoretically equal to one and is uninformative about $P(S)$. These features of the bound are illustrated in the simulation study described in Section 5. A naive 95% upper confidence bound for $P(S)$ could be taken to be the minimum of the 95% upper confidence bounds for each $P(A_j)/P(A_j|S)$. This, however, will be larger than the true 95% upper confidence bound. A correct bootstrap percentile upper confidence bound (Efron and Tibshirani, 1993) can be computed for the upper bound on $P(S)$ by resampling separately from the retrospective and random groups.

For the neuro-oncology data, the upper bound given in (5), restricted to sets A_j of the form $\{X_j = x\}$, for $x = 0, 1$, is 0.66. Because this upper bound was computed over a small subset of all possible sets, it is unlikely to be a sharp bound. 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. Without knowledge of these estimates, the physicians who conducted the study guessed that $P(S)$ is between 25/40

(0.63) and 25/35 (0.71).

3.2 Understanding the baseline correlates of selection

In Section 2 we found significant evidence for selection bias among the retrospective subsample of the neuro-oncology study. This was due primarily to the increased probability of selection given allelic loss of 1p relative to the probability of selection given intact 1p (no loss of chromosome 1). The associated odds-ratio is 5.31, with exact (univariate) p -value of 0.008. It is of interest to explore whether there are any other important baseline predictors of selection, beyond 1p. To do this, we subsetted the data according to 1p status. We then re-calculated the univariate p -values for testing for selection bias. These values are listed in Table 2. There appears to be no evidence for further selection bias within these subsets.

We then considered the relative probabilities of selection for one feature versus another. Under the full null hypothesis (1),

$$\frac{P(S|X_j = x)}{P(S|X_k = y)} = \frac{P(X_j = x|S) P(X_k = y|R)}{P(X_k = y|S) P(X_j = x|R)} = 1, \quad j, k \in \{1, \dots, J\}, \quad x, y \in \{0, 1\}.$$

For the neuro-oncology study, there are $2J \times (2J - 1)/2 = 231$ odds-ratios and associated exact p -values to consider. Using a simple Bonferroni correction, any one of these p -values should only be considered significant at the 0.05 level if it is less than 0.0002. In fact, none of the p -values was this small. Table 3 lists the comparisons for which the p -value was less than 0.01. While the p -values listed in the table must be interpreted in light of the multiplicity of the comparisons, the comparisons are informative for the relative contribution of the baseline features to the selection process. For example, the likelihood of selection given intact 1p is less than or equal to the likelihood of selection given all other baseline features, except for age at diagnosis greater than 45.

Lastly, we examined the probability of selection given a baseline feature as a function of the overall probability of selection, using the relationship

$$P(S|X_j = x) = \frac{P(X_j = x|S)}{P(X_j = x)} P(S).$$

We derived asymptotic confidence intervals for $P(X_j = x|S)/P(X_j = x)$ using a variant of Fieller's (1954) method for the confidence interval for the ratio of normal means. Table 4 lists the estimates and bounds for $P(S|X_j = x)$ based on taking $P(S) = 0.61$ and 0.67 , the estimated upper bound for $P(S)$ and the bootstrap upper 95% confidence limit for the upper bound. The

baseline features that appear to drive the selection bias, in terms of having lower confidence limits for the conditional probability of selection larger than the estimated upper bound for $P(S)$, are $\text{age} \leq 45$, no ring enhancement, 1pLOH, and no PTEN mutation. This analysis, however, does not correct for multiple comparisons, and should be viewed as descriptive.

4. Adjusting for the selection process in the comparison of treatment strategies

Although the overall probability of selection, $P(S)$, is not identifiable from the observed data, it was shown in Section 3 that an upper bound for it is identifiable. Alternatively, it may be known more accurately from external sources, such as the physicians who conducted the study. This probability is essential to obtaining correct analyses of the endpoints of interest, with respect to which there is selection bias. Here we show that it is essential to properly adjust for selection bias in the comparison of the treatment strategies of initial treatment with radiation therapy versus initial treatment with chemotherapy, with respect to response and survival.

4.1 Adjusting for selection in the comparison of response

The observed rates of response to chemotherapy for the two treatment strategies are quite different: 96% of the patients who were initially treated with radiation therapy responded, versus 66% of the patients who were initially treated with chemotherapy. The naive 95% confidence intervals for these rates are (88%,100%) and (51%,81%), respectively, and a naive analysis would conclude that the strategies are significantly different with respect to response. However, an analysis that corrects for the selection bias cannot conclude any difference in response rates. Note that the naive analysis is correct, as long as the comparison is understood as being between the strategies of initial treatment with radiation followed by chemotherapy at recurrence versus initial treatment with chemotherapy, and as long as follow-up is long enough to include patients with the longest times to recurrence.

Given the presence of selection bias in the neuro-oncology study, as was demonstrated in Section 2, the estimate of a 96% response rate for the retrospective sample (i.e., those who initially received radiation therapy), is incorrect. This follows from writing

$$\begin{aligned} P(\text{response}|\text{radiation first}) &= P(\text{response}|S, \text{radiation first})P(S|\text{radiation first}) \\ &\quad + P(\text{response}|S^c, \text{radiation first})P(S^c|\text{radiation first}). \end{aligned}$$

Even if we knew the overall probability of selection, $P(S)$, the probability of response among patients who were treated with radiation therapy first but who were not selected for the study,

$P(\text{response}|S^c, \text{radiation first})$, is not identifiable from the data. Thus, at best, we can bound $P(\text{response}|\text{radiation first})$ above by

$$P(\text{response}|S, \text{radiation first})P(S|\text{radiation first}) + 1 - P(S)$$

and below by

$$P(\text{response}|S, \text{radiation first})P(S|\text{radiation first}).$$

Note that if there were no selection bias, even if $P(S) < 1$, the naive estimate and confidence interval would be correct, as in this case,

$$P(\text{response}|\text{radiation first}) = P(\text{response}|S, \text{radiation first}) = P(\text{response}|S^c, \text{radiation first}).$$

If it were of interest to compare response rates conditioning further on certain baseline features, similar bounds could be obtained. Likewise, similar bounds could be obtained for response rates conditional on baseline features, but not conditional on treatment strategy.

We obtain a 95% confidence interval for $P(\text{response}|\text{radiation first})$ that correctly accounts for the selection bias by calculating an approximate upper 97.5% confidence limit for the upper bound and an approximate lower 97.5% confidence limit for the lower bound. This is given by

$$\left(\left[\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n_S} \right] P(S), \left[\hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/n_S} \right] P(S) + 1 - P(S) \right),$$

where \hat{p} is the proportion of responses among the study patients who received radiation therapy first (i.e., the retrospectively sampled group). When $P(S)$ is set to 0.66, its estimated upper bound, the confidence interval for the probability of response for those who received radiation therapy first is (0.58,1.00). When compared with the confidence interval for those who received chemotherapy first, (0.51,0.81), no difference in response rates can be inferred. In fact, for the confidence intervals for the two treatment strategies to be non-overlapping, $P(S)$ would have to be 0.92 or larger. Thus, given that there is selection bias, and that the overall probability of selection is less than 0.70 with probability 0.95, we cannot conclude that the two treatment strategies differ with respect to response.

4.2 Adjusting for selection in the comparison of survival

The same procedure that was used to obtain a corrected confidence interval for the probability of response for the retrospective sample can be used to obtain a corrected confidence interval for survival probabilities. For example, it is of interest to compare the probabilities of surviving

two years from diagnosis for the two treatment strategies. The naive 95% confidence interval for this probability is (88.6%,100%) for the patients who received radiation therapy first and it is (62.3%,87.9%) for the patients who received chemotherapy first. As these are non-overlapping, a naive analysis would conclude that the strategies differ with respect to two-year survival. When $P(S)$ is set to 0.66, its estimated upper bound, the confidence interval for the patients who received radiation therapy first becomes (58.5%,100%) and no difference between strategies can be concluded. In fact, for these confidence intervals to be non-overlapping, $P(S)$ would have to be greater than 0.99. Figure 1 contains the Kaplan-Meier plot for the randomly sampled group and confidence bounds for the retrospectively sampled group.

5. Simulations

We performed several simulations to assess the test for selection bias and the upper bound for the probability of selection derived in Sections 2 and 3. Because of the computational burden involved in calculating the p -value for the test for even a single dataset with several baseline predictors, we considered there to be only the single predictor of 1p status. We chose 1p because of its role in the selection bias (see Section 2) and because it has been shown to be a strong predictor of response and survival (Cairncross et al., 1998, Ino et al., 2000). The first set of simulations was designed so that $P(S) < 1$, with no selection bias. In particular, in each of 1000 repetitions, we sampled with replacement from the 50 randomly sampled patients to create a retrospectively sampled group. In a second set of simulations, we incorporated selection bias with respect to allelic loss of 1p. In these simulations, in each of 1000 repetitions, we sampled with replacement from the 50 randomly sampled patients on the basis of 1p status to create a retrospectively sampled group. For each simulated data set, we calculated the p -value for the test for selection bias, as well as the upper bound for $P(S)$ and its 95% upper confidence limit, and the ratio of the probabilities of selection given 1p status. We then computed the rejection rate as the proportion of p -values that were less than 0.05.

The results of the simulations are reported in Table 5. Under the scenarios chosen for the simulations, the test for selection bias has a type I error of less than 0.03. The power of the test is a function of the degree of selection bias, measured by $P(S|1p = 1)/P(S|1p = 0)$. The power appears to be greater than 70% for $P(S|1p = 1)/P(S|1p = 0) > 3.33$, it drops to 33% for $P(S|1p = 1)/P(S|1p = 0) = 2.33$, to 16% for $P(S|1p = 1)/P(S|1p = 0) = 1.8$, and to 8% for $P(S|1p = 1)/P(S|1p = 0) = 1.5$. As noted in Section 3.1, the upper bound for $P(S)$, labelled

$\hat{P}(S)$ in the table, is uninformative when there is no selection bias, and estimates $P(S)$ itself if $P(S|1p = 1) = 1$. As predicted, for the first five entries in the table for which there is no selection bias, $\hat{P}(S)$ is not close to $P(S)$. For the three entries in the table for which $P(S|1p = 1) = 1$, $\hat{P}(S)$ is an excellent estimator of $P(S)$, whereas it is not as good an estimator of $P(S)$ for the remaining four entries for which $P(S|1p = 1) < 1$. Had we considered all of the baseline features, with selection bias due only to 1p status, the power would likely have been slightly lower than what we observed.

6. 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 study contained a randomly selected subsample, which afforded us the opportunity to test for selection bias, and to quantify it. Our test for selection bias exploits the fact that the probabilities of events defined by the baseline features can be estimated using the random subsample. Based on our simulations, it appears to have low type I error rate and high power for moderate selection bias. Some limitations of the test are that it assumes that all of the baseline factors that influence selection and outcome have been recorded and that selection depends at least partially on baseline features, and not solely on post-baseline outcomes.

Further, we have proposed an estimator for the overall probability of selection that is useful if a selection bias is found. This estimator will be close to the true value under an extreme form of selection bias, and will be an upper bound under less extreme forms. An estimator of this probability is useful for estimating the probability of selection conditional on baseline features, and thus for understanding the baseline correlates of the selection process. It is useful also for comparisons of the retrospectively sampled patients with the randomly sampled patients.

In the neuro-oncology study, we found significant evidence for selection bias in the retrospective subsample. We estimated that the overall probability of selection was less than 0.61, and that the selection bias was primarily driven by the preferential selection of patients who had allelic loss of 1p. It may have been driven also by the preferential selection of patients who were young, had tumors that were not ring-enhancing, and did not have PTEN mutations. While 1p and PTEN status may have affected a patient’s recurrence following initial treatment with radiation therapy, the first event in the selection process, they could not have overtly impacted the subsequent events in the selection process. This is because they were unknown to both the

physician and the patient at the time of recurrence, when the physician would have recommended chemotherapy and the patient would have decided to receive it. Thus, this analysis increases our understanding of the selection process in that it identifies the baseline features that are predictive of selection, some of which are unknown at the time of selection. Lastly, we found that while the patients who were treated initially with radiation, followed by chemotherapy at recurrence, (the retrospective group) and the patients who were treated initially with chemotherapy (the random group) appeared to differ with respect to response and two-year survival, these differences were not significant when the selection bias was taken into account.

ACKNOWLEDGEMENTS

This research was supported in part by NIH grants CA75971 and CA57683.

REFERENCES

- Barker, F.G. 2nd, Chang, S.M., Gutin, P.H., Malec, M.K., McDermott, M.W., Prados, M.D., Wilson, C.B. (1998), "Survival and functional status after resection of recurrent glioblastoma multiforme," *Neurosurgery* 42, 709-720.
- Betensky, R.A. and Rabinowitz, D. (1999), "Maximally selected χ^2 statistics for $k \times 2$ tables," *Biometrics*, 55, 317-320.
- Betensky, R.A. and Rabinowitz, D. (2000), "Simple approximations for the maximal transmission/disequilibrium test with a multi-allelic marker." To appear, *Annals of Human Genetics*.
- Cairncross, J.G., Ueki, K., Zlatescu, M.C., Lisle, D., Finkelstein, D.M., Hammond, R.R., Silver, J.S., Stark, P.C., Macdonald, D.R., Ino, Y., Ramsay, D.A. and Louis, D.N. (1998), "Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas," *Journal of the National Cancer Institute*, 90, 1473-1479.
- Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.
- Ellenberg, J.H. (1994), "Selection bias in observational and experimental studies," *Statistics in Medicine*, 13, 557-567.
- Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, New York: Wiley.
- Fieller, E.C. (1954), "Some problems in interval estimation," *Journal of the Royal Statistical Society, Series B*, 16, 175-185.
- Florell, R.C., Macdonald, D.R., Irish, W.D., Bernstein, M., Leibel, S.A., Gutin, P.H. and Cairncross, J.G. (1992), "Selection bias, survival, and brachytherapy for glioma," *J Neurosurg* 76, 179-183.

- Gower, J.C. and Ross, G.J.S. (1969), "Minimum spanning trees and single linkage cluster analysis," *Applied Statistics*, 18, 54-64.
- Halpern, J. (1982), "Maximally selected chi square statistics for small samples," *Biometrics*, 38, 1017-1023.
- Huncharek, M. and Muscat, J. (1998), "Treatment of recurrent high grade astrocytoma; results of a systematic review of 1,415 patients," *Anticancer Res* 18, 1303-1311.
- Ino, Y., Betensky, R.A., Zlatescu, M.C., Sasaki, H., Macdonald, D.R., Stemmer-Rachamimov, A.O., Ramsay, D.A., Cairncross, J.G., Louis, D.N. (2000), "Molecular diagnosis in the clinical management of patients with malignant gliomas," in preparation.
- Irish, W.D., Macdonald, D.R. and Cairncross, J.G. (1997), "Measuring bias in uncontrolled brain tumor trials – to randomize or not to randomize?" *Can J Neurol Sci* 24, 307-312.
- Kleinbaum, D.G., Kupper, L.L. and Morgenstern, H. (1982), *Epidemiologic Research*, New York: Van Nostrand Reinhold.
- Miller, R. and Siegmund, D. (1982), "Maximally selected chi square statistics," *Biometrics*, 38, 1011-1016.
- Razack, N., Baumgartner, J. and Bruner, J. (1998), "Pediatric oligodendrogliomas," *Pediatric Neurosurgery* 28, 121-129.
- Videtic, G.M., Gaspar, L.E., Zamorano, L., Fontanesi, J., Levin, K.J., Kupsky, W.J., Tekyi-Mensah, S. (1999), "Use of the RTOG recursive partitioning analysis to validate the benefit of iodine-125 implants in the primary treatment of malignant gliomas," *Int J Radiat Oncol Biol Phys* 45, 687-692.
- Williams, W.H. (1978), "How bad can 'good' data really be?" *The American Statistician*, 32, 61-65.
- Winger, M.J., Macdonald, D.R., Schold, S.C. Jr. and Cairncross, J.G. (1989), "Selection bias in clinical trials of anaplastic glioma," *Ann Neurol* 26, 531-534.
- Worsley, K.J (1982), "An improved Bonferroni inequality and applications," *Biometrika*, 69, 297-302.

Table 1: *Description of data and univariate test statistics and p-values for selection bias*

Feature	Random group	Retrospective group	t_j^\dagger	P_j^\ddagger
age \leq 45	26/50 (52)	19/25 (76)	2.15	0.051
male	26/50 (52)	14/25 (56)	0.33	0.809
enhancement at diagnosis	34/50 (68)	14/22 (64)	0.36	0.789
ring enhancement at diagnosis	9/50 (18)	0/22 (0)	∞	0.029
1pLOH ¹	29/50 (58)	22/25 (88)	3.18	0.008
19qLOH	33/50 (66)	21/25 (84)	1.82	0.054
10qLOH	12/47 (26)	2/22 (9)	1.92	0.050
PTEN mutation ²	6/49 (12)	0/25 (0)	∞	0.076
p16 deletion ³	7/50 (14)	5/25 (20)	0.64	0.740
EGFR amplification ⁴	7/50 (14)	1/25 (4)	1.72	0.108
p53 mutation ⁵	9/50 (18)	4/25 (16)	0.22	0.750

$\dagger t_j$ is the normalized absolute log odds-ratio

$\ddagger P_j$ is the univariate exact p -value

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.

Table 2: *Univariate p-values for selection bias among subgroups defined by 1p status*

Feature	1pLOH [†]	intact 1p [‡]
age \leq 45	0.250	0.223
male	0.782	0.273
enhancement at diagnosis	1.000	1.000
ring enhancement at diagnosis	1.000	0.304
19qLOH	1.000	0.413
10qLOH	1.000	0.544
PTEN mutation	NA	0.413
p16 deletion	0.152	1.000
EGFR amplification	NA	1.000
p53 mutation	1.000	0.223

See note to Table 1 for explanations of genetic events.

Table 3: Ratios of probabilities of selection given events A_j and A_k with univariate p -value less than 0.01

A_j	A_k	$P(S A_j)/P(S A_k)$	p -value
1pLOH	1p intact	5.31	0.008
1pLOH	no p16 deletion	1.63	0.005
1pLOH	no p53 mutation	1.48	0.009
1p intact	age > 45	3.03	0.004
1p intact	age \leq 45	0.20	0.001
1p intact	male	0.27	0.005
1p intact	19qLOH	0.22	0.005
1p intact	non-ring enhancing	0.27	0.006
1p intact	no p16 deletion	0.31	0.007
1p intact	no EGFR amplification	0.26	0.006
1p intact	no p53 mutation	0.28	0.008
non-ring enhancing	no PTEN mutation	0.25	0.003

See note to Table 1 for explanations of genetic events.

Table 4: *Probabilities of selection given baseline features*

A_j	$P(S) = 0.66$		$P(S) = 0.70$	
	$P(S A_j)$	95% confidence interval	$P(S A_j)$	95% confidence interval
age \leq 45	0.96	(0.68,1.00)	1.00	(0.72,1.00)
age $>$ 45	0.33	(0.16,0.70)	0.35	(0.16,0.74)
male	0.71	(0.46,1.00)	0.75	(0.49,1.00)
female	0.60	(0.36,1.00)	0.64	(0.38,1.00)
enhancement	0.62	(0.43,0.89)	0.66	(0.45,0.95)
no enhancement	0.75	(0.38,1.00)	0.80	(0.40,1.00)
ring enhancement	0.00	(0.00,1.00)	0.00	(0.00,1.00)
no ring enhancement	0.80	(0.71,0.92)	0.85	(0.75,0.97)
1pLOH	1.00	(0.76,1.00)	1.00	(0.81,1.00)
1p intact	0.19	(0.06,0.57)	0.20	(0.07,0.61)
19qLOH	0.84	(0.65,1.00)	0.89	(0.69,1.00)
19q intact	0.31	(0.12,0.83)	0.33	(0.12,0.88)
10qLOH	0.24	(0.06,0.96)	0.25	(0.06,1.00)
10q intact	0.81	(0.65,1.00)	0.85	(0.69,1.00)
PTEN mutation	0.00	(0.00,1.00)	0.00	(0.00,1.00)
no PTEN mutation	0.75	(0.68,0.84)	0.80	(0.72,0.89)
p16 deletion	0.94	(0.33,1.00)	1.00	(0.35,1.00)
no p16 deletion	0.61	(0.49,0.77)	0.65	(0.52,0.82)
EGFR amplification	0.19	(0.02,1.00)	0.20	(0.03,1.00)
no EGFR amplification	0.74	(0.64,0.85)	0.78	(0.68,0.90)
p53 mutation	0.59	(0.20,1.00)	0.62	(0.21,1.00)
no p53 mutation	0.68	(0.55,0.84)	0.72	(0.58,0.89)

See note to Table 1 for explanations of genetic events.

Table 5: *Simulation study (1000 repetitions)*

$P(S)$	Design parameters			rejection rate	Results		
	$P(S 1p=1)$	$P(S 1p=0)$	$P(S 1p=1)/P(S 1p=0)$		$\hat{P}(S)$ (upper 95% bound)	$\hat{P}(S 1p=1)/\hat{P}(S 1p=0)$	
0.20	0.20	0.20	1.00	0.020	0.81 (0.97)	1.09	1.09
0.40	0.40	0.40	1.00	0.027	0.85 (0.97)	0.89	0.89
0.50	0.50	0.50	1.00	0.019	0.87 (0.97)	1.09	1.09
0.60	0.60	0.60	1.00	0.010	0.88 (0.97)	1.09	1.09
0.80	0.80	0.80	1.00	0.009	0.89 (0.99)	0.98	0.98
0.56	0.90	0.10	9.00	0.957	0.63 (0.71)	9.05	9.05
0.64	1.00	0.15	6.67	0.996	0.64 (0.70)	6.88	6.88
0.66	1.00	0.20	5.00	0.966	0.67 (0.73)	5.18	5.18
0.55	0.80	0.20	4.00	0.730	0.69 (0.80)	4.10	4.10
0.71	1.00	0.30	3.33	0.770	0.71 (0.80)	3.26	3.26
0.53	0.70	0.30	2.33	0.328	0.76 (0.92)	2.40	2.40
0.73	0.90	0.50	1.80	0.159	0.82 (0.94)	1.86	1.86
0.52	0.60	0.40	1.50	0.076	0.84 (0.98)	1.54	1.54

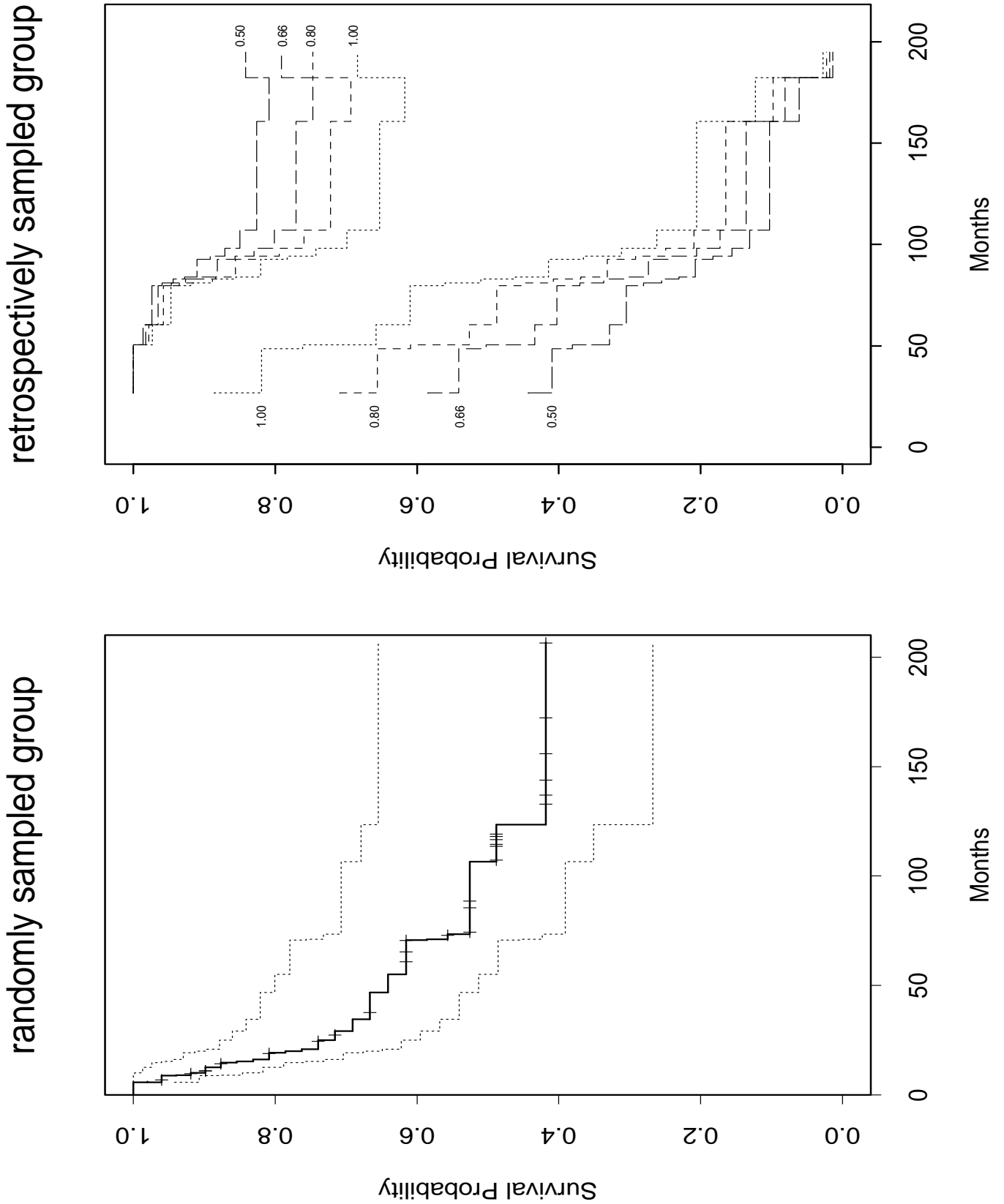


Figure 1: Kaplan-Meier curve and 95% confidence intervals from the Neuro-oncology Study: the four sets of intervals for the retrospectively sampled group are based on hypothetical probabilities of selection of 0.50, 0.66, 0.80, and 1.00.