

The influence of central review on outcome associations in childhood malignant gliomas: Results from the CCG-945 experience¹

Ian F. Pollack,² James M. Boyett, Allan J. Yates, Peter C. Burger, Floyd H. Gilles, Richard L. Davis, and Jonathan L. Finlay, for the Children's Cancer Group

Department of Neurosurgery, University of Pittsburgh School of Medicine and the Children's Hospital of Pittsburgh, Pittsburgh, PA 15213 (I.F.P.); Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105 (J.M.B.); Department of Pathology, Ohio State University, Columbus, OH 43210 (A.J.Y.); Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205 (P.C.B.); Department of Pathology, University of Southern California, Los Angeles, CA 90027 (F.H.G.); Department of Pathology, University of California, San Francisco, CA 94143 (R.L.D.); Department of Pediatrics, New York University Medical Center, New York 10016 (J.L.F.); USA

To examine the influence of the pathology review mechanism on the results of analyses of therapeutic efficacy and biological prognostic correlates for pediatric high-grade gliomas, we evaluated the effects of using single-expert review or consensus review, as alternatives to institutional classification, in determining outcome results of a large randomized trial. The study group was the randomized cohort of Children's Cancer Group study 945, which compared efficacy of 2 chemotherapy regimens adjuvant to surgery and radiation. Trial eligibility required institutional histopathologic diagnosis of high-grade glioma. Sections of study tumors also were centrally reviewed, initially by a study review neuropathologist and subsequently by 5 neuropathologists, including the review pathologist. Reviews were independent, and reviewers were masked to clinical factors and outcomes, and consensus diagnoses of the panel were then established.

Among 172 eligible patients, 42 tumors were classified as discordant on single-expert review and 51 on consensus review. Progression-free survival probabilities calculated for patients with tumors classified as high-grade gliomas by either single-expert or consensus review were inferior to those for the overall, institutionally diagnosed cohort. However, conclusions of the study regarding relative efficacy of treatment and clinical and molecular outcome correlates were unaffected by diagnosis method. Resection extent, proliferation index, and p53 expression were associated strongly with outcome, regardless of diagnosis method. However, comparisons between arms in which inclusion was determined by different review criteria for each arm caused spurious conclusions about efficacy differences between treatments. We conclude that the pathology review mechanism had little effect on within-trial comparisons of therapeutic effects or prognostic correlates in this randomized study, but strongly influenced survival distributions that were calculated for each treatment arm. These results support the implementation of expedited central review in therapeutic studies involving childhood malignant gliomas as a way to prospectively identify and exclude cases with discordant diagnoses and indicate the need for additional measures, such as molecular assessments, to increase the reproducibility of neuropathologic classification for these tumors. *Neuro-Oncology* 5, 197–207, 2003 (Posted to *Neuro-Oncology* [serial online], Doc.03-009, May 13, 2003. URL <http://neuro-oncology.mc.duke.edu>; DOI: 10.1215/S1152 8517 03 00009 7)

Received March 17, 2003; accepted March 27, 2003.

¹ This work was supported in part by NIH grants NS37704 (I.F.P.), CA13539 (Children's Cancer Group), and CA21765 (J.M.B.) and by the American Lebanese Syrian Associated Charities (J.M.B.).

² Address correspondence to Ian F. Pollack, Children's Oncology Group, P.O. Box 60012, Arcadia, CA 91066-6012, USA (Ian.Pollack@chp.edu).

³ Abbreviations used are as follows: 8-in-1, 8-drugs-in-one-day; AA, anaplastic astrocytomas; CCG-945, Children's Cancer Group 945 study; GBM, glioblastoma multiforme; HGG, high-grade glioma; pCV, prednisone, CCNU (lomustine), and vincristine.

Classifying gliomas is challenging and requires judgment, experience, and meticulous adherence to established nosologic guidelines. Because of the histologic heterogeneity of these tumors, it is not uncommon for experienced reviewers to reach different diagnoses from specimens of the same tumor (Aldape et al., 2000; Childhood Brain Tumor Consortium, 1989; Yung et al., 1999). Pediatric gliomas pose additional difficulties because there are several low-grade variants that arise specifically in children, exhibit many morphologic features of malignancy, and are easily mistaken for high-grade gliomas, yet are associated with generally favorable prognoses (Kleihues et al., 1993; Papahill et al., 1996). These diagnostic challenges have complicated the design of clinical trials of new therapeutic approaches to improve the generally poor prognoses of malignant gliomas. Most studies to date have been designed using an “intent-to-treat” approach, which is generally accepted by biostatisticians (although not necessarily by all clinical investigators) as the preferred basis for analysis in comparative clinical trials (Schwartz and Lellouch, 1967; Tsiatis, 1990). This approach acknowledges that there is no absolute standard by which inclusion can be validated, and inclusion criteria established on an institutional basis provide a real-world approximation of efficacy of a given treatment under conditions in which it is most likely to be applied. Some favor an alternate approach that involves restriction of analyses to patients judged by post hoc central review to have met a protocol’s eligibility criteria. To evaluate the influence of both analytic approaches on study interpretations, we critically examined the results of the Children’s Cancer Group 945 study (CCG-945),³ the largest randomized study reported to date of pediatric malignant gliomas (Finlay et al., 1995), in the context of central review by an independent panel of 5 senior neuropathologists. Although the study found striking discrepancies between institutional, individual-reviewer, and consensus diagnoses, these disparities had no influence on interpretation of the within-trial comparisons, or identification of important clinical and biological prognostic factors. However, we found that potentially serious interpretational errors could arise when outcome results defined by different reviewers were compared, which mandates that caution be exercised in comparing published studies for high-grade gliomas (HGGs) in which different review criteria have been employed. This observation supports the use of prospective, panel-based, central review for cooperative group studies involving these tumors to facilitate exclusion of discordant histologies and thereby enhance the reliability of cross-study comparisons of outcome data.

Methods

Patient Population

We examined the randomly assigned patients of the HGG study CCG-945 (Finlay et al., 1995). Eligible children with malignant gliomas were treated with a combination of surgery, radiotherapy (5400 cGy in 180-cGy

fractions), and chemotherapy with (a) the postirradiation regimen of prednisone, CCNU (lomustine), and vincristine (pCV) or (b) the 8-drugs-in-one-day (8-in-1) regimen, administered for 2 cycles before and continued after irradiation. This more complex regimen incorporated 7 agents with some activity against pediatric brain tumors: lomustine, vincristine, hydroxyurea, procarbazine, cisplatin, cytosine arabinoside, and dacarbazine; methylprednisolone was added to reduce cerebral edema. The goal of this approach was to circumvent resistance to individual agents by exposing the tumor to multiple agents, albeit at low-dose intensity. Details of these regimens have been previously reported (Finlay et al., 1995; Pendergrass et al., 1987). Patients between 18 months and 21 years old who had intracranial HGGs (N = 172) were assigned randomly between the 2 treatment regimens (Finlay et al., 1995). CCG-945 was open to participating institutions between April 1, 1985, and May 31, 1990. Eligibility mandated institutional histopathologic confirmation of HGG that arose primarily outside the brainstem. Tumors were categorized as glioblastoma multiforme (GBM), anaplastic astrocytoma (AA), or other eligible high-grade glioma (other HGG) such as anaplastic oligoastrocytoma. No therapy other than surgery was permitted before entry. Decisions concerning the extent of tumor resection were left to the discretion of the treating neurosurgeon; however, the protocol recommended removal of as much tumor as was safely feasible.

Institutional histologic diagnoses and clinical factors were recorded for all patients, and long-term follow-up was achieved. In a subset of patients for whom there was sufficient archival histologic material for biological analyses, immunohistochemical and molecular data also were collected, masked, and recorded in a separate database, which was merged with the clinical database for outcome analyses. For this analysis of review histologic classification, the Pediatric Branch of the Cooperative Human Tissue Network coordinated tissue accrual and distribution for each case. Specimens were coded so that investigators other than statisticians were masked to clinical and outcome results.

Central Review Process

In initial reports of this series (Allen et al., 1998; Finlay et al., 1995; Geyer et al., 1995; Wisoff et al., 1998), the review neuropathologist (A.J.Y) examined all cases and assigned a review diagnosis. Tumors were categorized into 4 groups: GBM, AA, other HGG, and not-HGG (discordant, e.g., low-grade astrocytoma). This initial review predated the revised WHO criteria published in 1993 (Kleihues et al., 1993). In the initial central review, only 24 tumors were considered to be overtly discordant with a diagnosis of HGG, and in 3 others the amount of histopathologic material was inadequate to establish a definitive review histological diagnosis.

For the current report, our diagnostic standard was more rigorous. The review neuropathologist, who was masked to institutional diagnoses and his original review diagnoses, provided revised review diagnoses based on the revised WHO criteria (Kleihues et al., 1993), and that

review was used to establish the consensus diagnosis with the independent, concurrent reviews of 4 other experienced neuropathologists who were masked to outcome. Cases in which at least 3 of 5 pathologists reached a common histologic diagnosis of malignant glioma with an identical histological subtype—AA, GBM, or other HGG—were deemed “consensus AA,” “consensus GBM,” or “consensus other HGG,” respectively. Cases in which at least 3 of 5 pathologists reached a diagnosis of HGG but fewer than 3 agreed about the exact subtype classification were considered to be “HGG, without subtype consensus.” Cases in which at least 3 of 5 pathologists classified the tumor as discordant were considered to be “consensus not HGG.”

Primary Outcome Comparison

The primary goal of the original clinical study was to compare efficacy of the pCV regimen with that of the 8-in-1 regimen. That outcome, based upon institutional diagnoses, was among the best reported for these tumors, with a 5-year overall survival of $36 \pm 6\%$ and a 5-year progression-free survival of $33 \pm 5\%$ (Finlay et al., 1995). Neither distributions of progression-free survival nor overall survival differed significantly between treatment arms based on institutional diagnoses or initial review diagnoses (Finlay et al., 1995). Here we repeated treatment comparisons of eligible patients using those who had consensus diagnoses of HGGs and investigated differences in sensitivity and specificity of expert reviewers in identifying these tumors using the WHO criteria (Kleihues et al., 1993).

Additional Outcome Correlates

Previous reports on this cohort that used institutional and initial review diagnoses noted a strong association between the amount of postoperative residual disease and outcome (Finlay et al., 1995; Wisoff et al., 1998). More recent reports of biologic and molecular prognostic factors that used institutional and revised review diagnoses found that proliferation index (assessed by using the MIB-1 antibody to label the Ki-67 antigen) and p53 expression status were associated strongly with progression-free survival (Pollack et al., 2002a, b). To determine whether those factors were associated with outcome when consensus diagnoses were employed, we repeated the analyses using that more stringent inclusion criterion.

Statistical Considerations

Survival was defined as the time from randomized assignment to death or date of last contact. The primary end point for correlating biological and molecular prognostic factors was time to tumor progression, defined as the time from randomized assignment to disease progression. Patients who died from causes other than primary tumor progression ($n = 5$) were censored at the time of death. Kaplan-Meier estimates of distributions of time to tumor progression and survival (Kalbfleisch and Prentice, 1980;

Kaplan and Meier, 1958) were provided with standard errors as suggested by Peto et al. (1976). Distributions of time to tumor progression were compared by using the Mantel-Haenszel statistic (Mantel, 1966), stratified when appropriate. No patient was censored within 6.5 years of randomized assignment, so 5-year survival rates were compared by using Fisher's exact test. Reviewers' sensitivity and specificity for identifying HGGs were estimated and compared by using the method of Qu et al. (1996). This random-effects model is applicable in situations in which there is no criterion standard and when dependence, conditioned upon true diagnostic state, exists among reviewers. In these analyses, reviewer diagnoses were dichotomized as HGG or not.

Results

Patient Characteristics

Archival tumor specimens were submitted for central review for the 172 eligible children who were assigned randomly on CCG-945. In 169 of them, histopathologic material was adequate for review, and diagnoses were provided by all 5 review neuropathologists. A comparison of initial review diagnoses of the study neuropathologist (reported in the original clinical summary of the series [Finlay et al., 1995]) and the revised review diagnoses (based upon the 1993 revision of the WHO criteria [Kleihues et al., 1993]) is shown in Table 1. The masked rereview resulted in diagnoses changing for 21 (12.4%) of 169 patients, mostly reclassifications of tumors as Not HGG (18 of 21). Sixteen percent of cases originally diagnosed as AA were changed to Not HGG. In the revised review, 42 cases were determined to be Not HGG. Comparisons between the revised review and the original institutional diagnoses are provided in Table 2.

A comparison of the classification of the cohort based on institutional and consensus review criteria showed an even greater number of cases that were discordant by consensus diagnoses ($n = 51$, 29.6%), as shown in Table 3. The panel did not agree on specific histologies of 16 other cases, but for all 16 cases the consensus was consistent with diagnosis of HGG. There was substantial discrepancy among experts in the classification of individual tumors. Figure 1A shows the percentages of tumors classified into the 4 diagnostic groups by individual expert reviewers, illustrating substantial variance in frequency of tumors within categories. Of particular note was the variance among the 5 neuropathologists in the proportion of the 169 patients who had tumors inconsistent with diagnoses of HGG (42, 56, 42, 61, and 35, ranging from 20% to 36%). That variance affected tumors that had consensus classifications of GBM (Fig. 1B), AA (Fig. 1C), other HGG (Fig. 1D), and Not HGG (Fig. 1E). For all groups, at least 1 reviewer reached an alternate diagnosis in over half of the cases. This percentage was particularly high for patients with consensus diagnoses of Not HGG, in which more than 76% of cases were classified as HGG by at least 1 reviewer. Based on the random-effects model, Figs. 2A and 2B show 95% con-

Table 1. Initial review diagnoses versus revised review diagnoses for the CCG-945 randomized cohort

Initial Review Diagnosis	Revised Review Diagnosis Based on 1993 WHO Criteria					Totals
	GBM	AA	Other HGG	Not HGG	Insufficient Tissue	
GBM	53	0	0	3	0	56
AA	0	61	2	12 (16%)	0	75
Other HGG	0	0	11	3	0	14
Not HGG	0	1	0	23	0	24
Insufficient Tissue	0	0	0	1	2	3
Totals	53	62	13	42	2	172

Abbreviations: AA, anaplastic astrocytomas; GBM, glioblastoma multiforme; HGG, high-grade glioma.

Table 2. Revised review and original institutional diagnoses for the CCG-945 randomized cohort

Institutional Diagnosis	Single Expert Reviewer Diagnosis					Totals
	GBM	AA	Other HGG	Not HGG	Insufficient Tissue	
GBM	37	12	3	5	0	57
AA	11	42	4	24	1	82
Other HGG	5	8	6	13	1	33
Totals	53	62	13	42	2	172

Abbreviations: AA, anaplastic astrocytomas; GBM, glioblastoma multiforme; HGG, high-grade glioma.

Table 3. Classification of the cohort based on institutional and on consensus review criteria

Institutional Diagnosis	Consensus Panel Diagnosis						Totals
	GBM	AA	Other HGG	Not HGG	HGG, No Consensus	Insufficient Tissue	
GBM	31	11	0	6	8	1	57
AA	7	39	1	28	6	1	82
Other HGG	3	7	3	17	2	1	33
Totals	41	57	4*	51 (29.6%)	16	3	172

Abbreviations: AA, anaplastic astrocytomas; GBM, glioblastoma multiforme; HGG, high-grade glioma.

*Other eligible HGG diagnoses included anaplastic mixed glioma (2), anaplastic oligodendroglioma (1), and anaplastic mixed versus anaplastic oligodendroglioma (1).

fidence interval estimates of sensitivity and specificity for each of the 5 reviewers for their diagnoses of HGG in relation to consensus diagnoses. Only reviewer #1 appeared different from the other 4 in sensitivity, but all reviewers had good sensitivity for diagnosing HGG in this cohort. The relatively high sensitivity for all reviewers was not surprising because each knew that the patients had been treated on CCG-945; however, there were statistically significant differences in specificity between reviewers #2 and #4 compared with reviewers #3 and #5 ($P \leq 0.0001$ for all 4). Reviewer #1 was statistically significantly different from #3 and #5 ($P < 0.040$ and $P < 0.031$, respectively) for specificity but not from reviewers #2 and #4 ($P > 0.05$ for both).

Effect of Classification Criteria on Primary Outcome Comparison

Given the substantial discrepancies in the classifications reached by the institutional and consensus diagnoses, the

variability between reviewers in the classification of individual cases, and the alterations in the diagnoses of a single expert review neuropathologist, who examined the same cases at different points in time in the context of evolving classification criteria, a valid concern could be raised regarding the integrity of the original study conclusions in terms of comparisons of efficacy between the 2 treatment regimens. As reported, there were no differences in distributions of progression-free or overall survival when the regimens were compared using institutional or original review diagnoses (Finlay et al., 1995). Figures 3A and 3B illustrate comparisons of survival distributions for children assigned randomly to the 8-in-1 and pCV regimens, using institutional diagnosis of HGG (shown here for reference) and consensus diagnosis of HGG, respectively, and neither comparison is statistically significant ($P > 0.7$ and $P > 0.6$, respectively) by Mantel-Haenszel tests. The 5-year survival estimates for the 8-in-1 and pCV regimens were $40 \pm 5\%$ and $36 \pm 5\%$, respectively, using institutional diagnoses, but only $23 \pm 5\%$

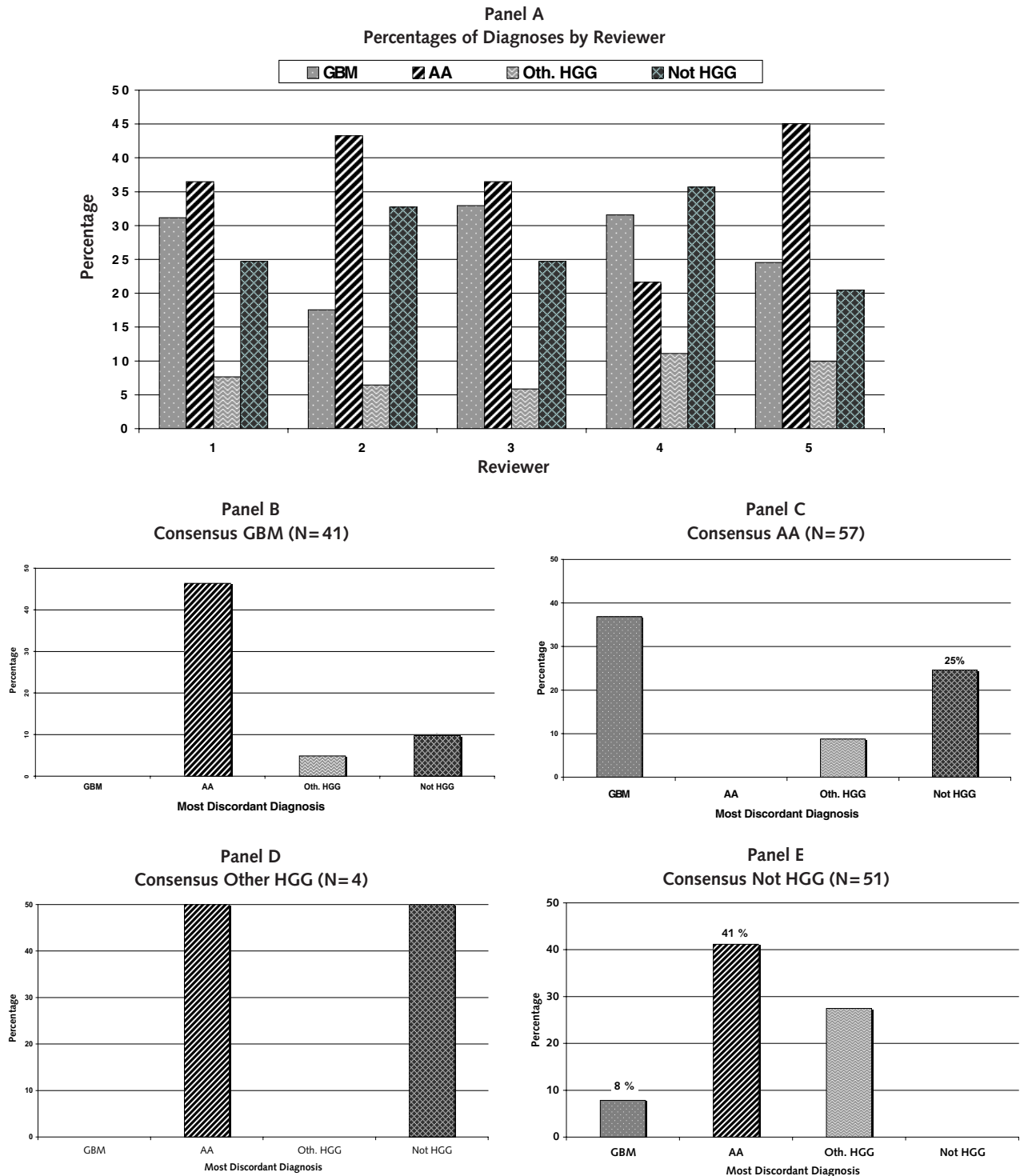


Fig. 1. Classification of tumors by 5 individual expert reviewers and by consensus review. Panel A. The percentages of tumors classified within different diagnostic groups by individual reviewers. For tumors within individual consensus diagnostic groups, a substantial percentage of reviewers reached alternate diagnoses. The most discordant alternate diagnoses are shown for GBM (Panel B), AA (Panel C), other HGG (Panel D), and Not HGG (Panel E).

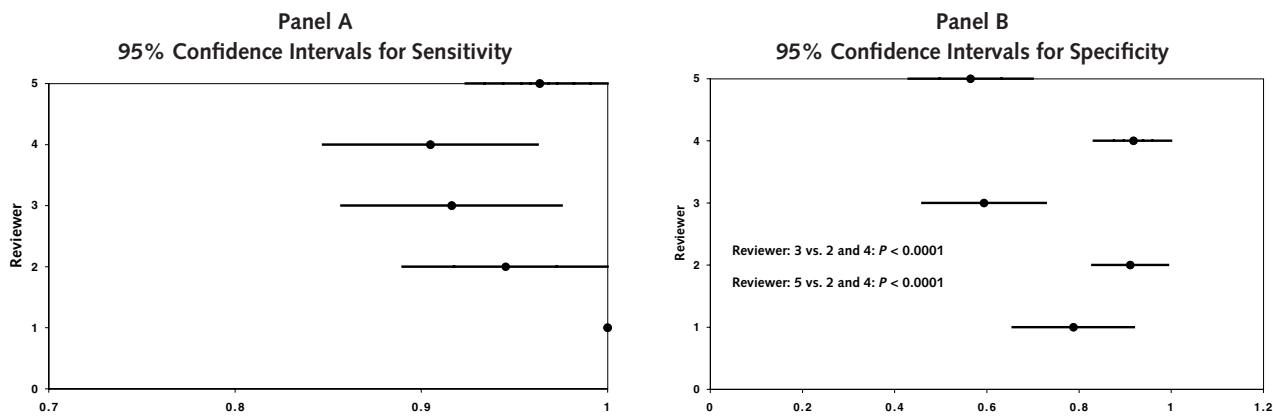


Fig. 2. Sensitivity and specificity for the 5 individual HGG diagnoses versus the consensus diagnosis. Based on the random-effects model, 95% confidence interval estimates are shown for sensitivity (Panel A) and specificity (Panel B) for the individual reviewers relative to the consensus diagnosis of HGG.

and $19 \pm 5\%$, respectively, using consensus diagnoses to define patients with HGG. The hazard ratio for the 8-in-1 to pCV regimens was 1.06 (95% CI, 0.7–1.5) using the institutional diagnoses, compared with 1.08 (95% CI, 0.7–1.6) using the consensus diagnoses. The strikingly lower survival results calculated with the HGG cohort deemed eligible by consensus review largely reflected that most of the excluded (Not HGG) tumors probably represented low-grade gliomas (Table 4), which have more favorable prognoses. Although only a small percentage of GBMs were determined to be Not HGG on consensus review, a sizeable percentage of them were determined either to be AA or to have HGG diagnoses without a consensus regarding histology. Accordingly, survival percentages were substantially lower in the subset of 31 patients deemed to have GBM on both consensus and institutional review than in the 26 with institutional diagnoses of GBM and alternate diagnoses on consensus review (Fig. 3C). An intermediate survival outcome was

observed in the small group of patients ($n = 10$) classified as AA or other HGG on institutional review but classified as GBM on consensus review.

We next repeated the outcome analysis using diagnoses made by each of the review neuropathologists, and the difference in survival for patients assigned between the 2 treatments was not significant for any of the 5 reviewers ($P > 0.50$ for each reviewer; data not shown). Although these results indicated that within-study comparisons of outcome between therapeutic arms were unaffected by review mode, they also showed the potential for erroneous conclusions if an investigator attempted to compare results for a given treatment using one histologic inclusion approach with those calculated using another. For example, Fig. 4A illustrates the differences in survival estimates if consensus HGG patients treated with the pCV regimen were compared with institutionally diagnosed HGG patients treated with the 8-in-1 regimen, which could lead to an incorrect attribution of superior efficacy to the latter. The converse interpretation would be reached if the review groupings were reversed.

Even more worrisome is that although no differences exist in event-free survival, progression-free survival, or overall survival between regimens using diagnoses reached by any expert member of the consensus panel (data not shown), comparisons between regimens using different reviewers could cause erroneous conclusions. Figure 4B shows comparison of survival distributions for patients assigned to 8-in-1 with HGG defined by reviewer 5 versus those assigned to pCV with HGG defined by reviewer 4, showing an apparent significant difference in efficacy. Figures 4C and 4D show the results for both arms for each reviewer, confirming that differences in Fig. 4B indicate differences between the reviewers' stringency in diagnosing HGG.

Table 4. Consensus diagnosis versus institutional diagnosis

Non-HGG Consensus Diagnoses*	Institutional Diagnosis		
	GBM	AA	Other HGG
Low-grade fibrillary astrocytoma	2	11	4
Pilocytic astrocytomas	2	5	1
Oligodendroglioma – mixed glioma	0	0	4
Ependymoma	1	2	3
Pleomorphic xanthoastrocytoma	1	3	0
Medulloblastoma – PNET	0	0	1
Ganglioglioma	0	0	2
Insufficient material for diagnosis	0	0	1
No consensus	0	7	1
Total	6	28	17

Abbreviations: AA, anaplastic astrocytoma; GBM, glioblastoma multiforme; HGG, high-grade glioma; PNET, primitive neuroectodermal tumor.

*Diagnoses represent the majority diagnoses reached by the consensus panel members among cases classified as “Not HGG.” Cases in which multiple non-HGG diagnoses were given without a clear majority for any one diagnosis are listed as “no consensus.”

Additional Outcome Correlations

Previous analyses based upon institutional diagnoses, original review diagnoses, and revised review diagnoses of the study neuropathologist showed several factors

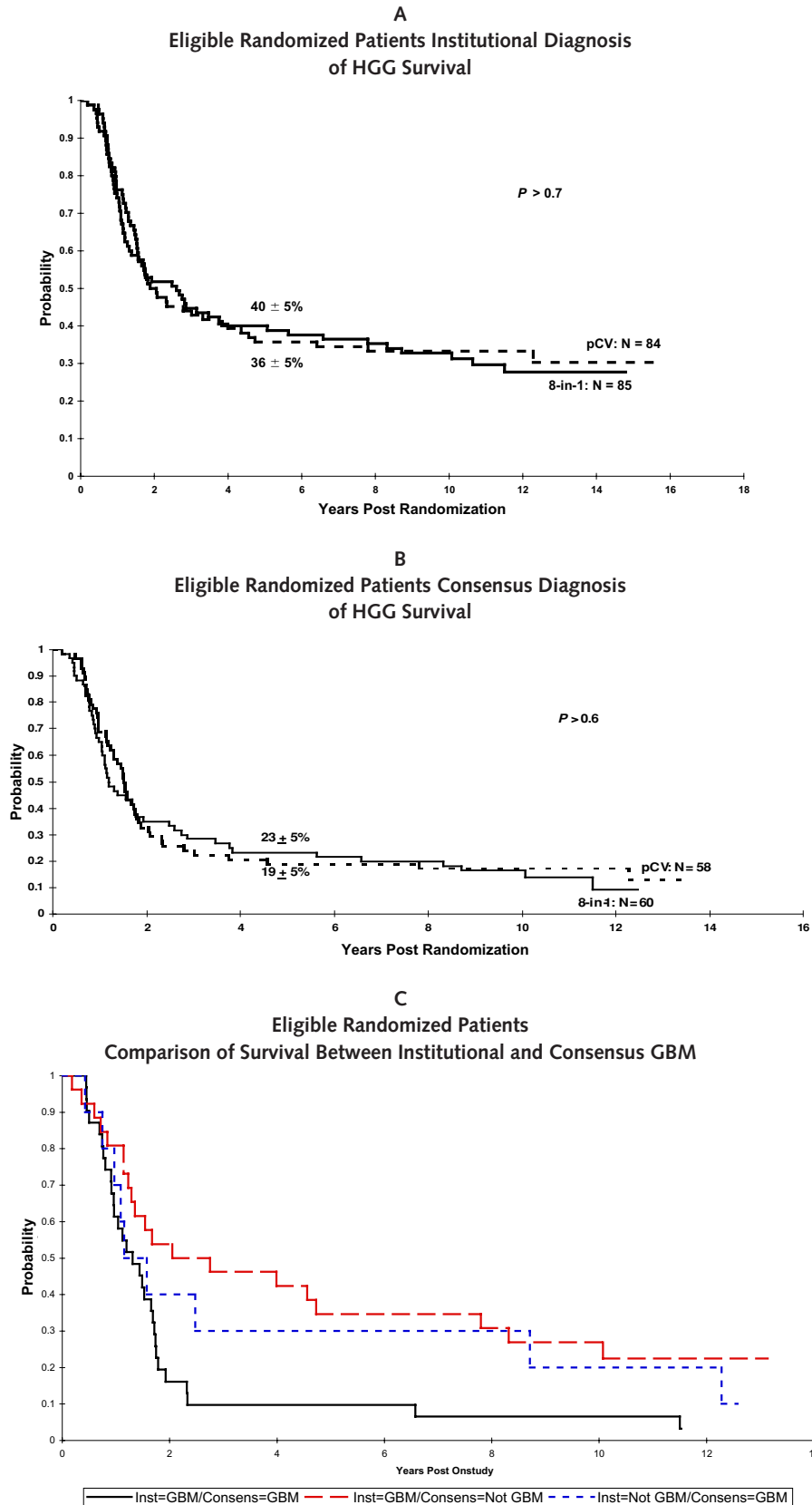


Fig. 3. Survival distributions. Comparisons of distributions of survival for children randomized to treatment with pCV or 8-in-1 adjuvant chemotherapy in the CCG-945 clinical trial based on (A) institutional diagnosis of HGG and (B) consensus diagnosis of HGG. (C) Survival distribution for the 31 children with tumors classified as GBM on both institutional and consensus review, the 26 children with tumors institutionally classified as GBM in which the consensus diagnosis differed, and the 10 children with consensus diagnoses of GBM in which the institutional diagnosis was AA or other HGG.

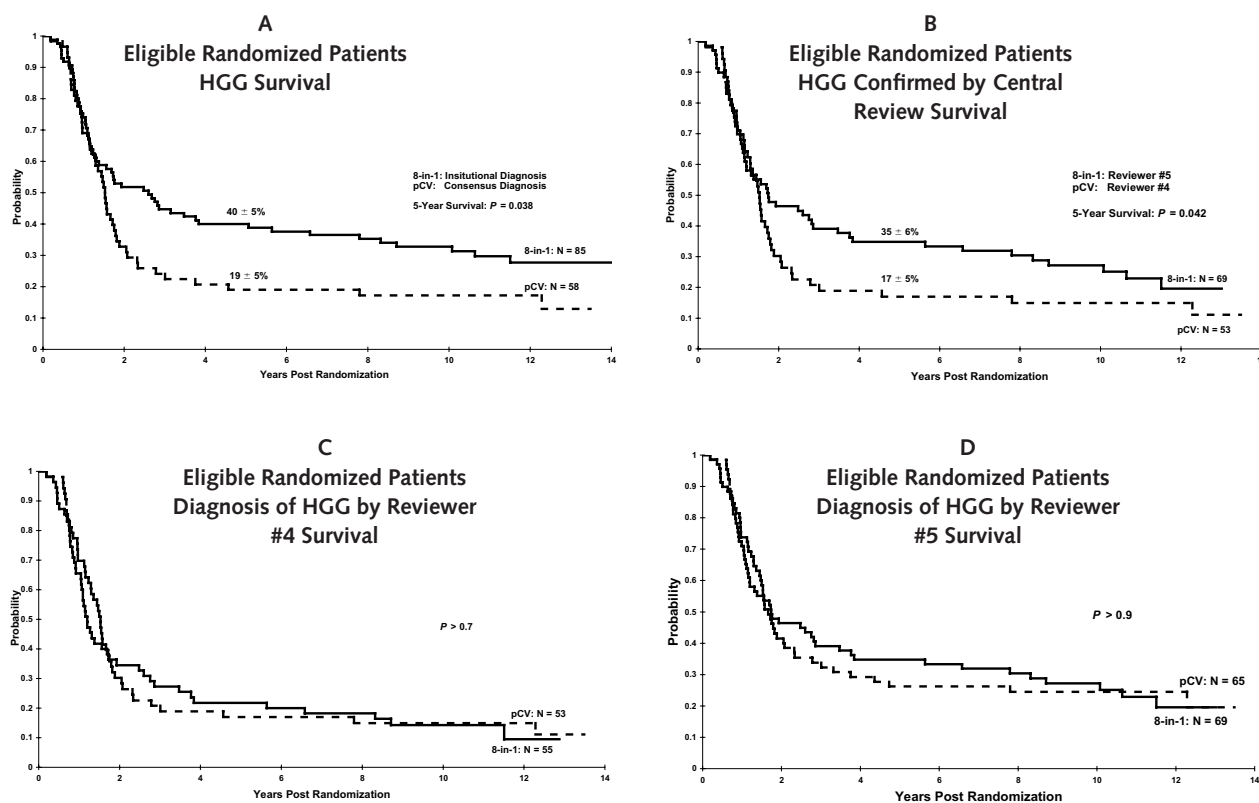


Fig. 4. Differences in survival estimates and distribution. A. Estimates artificially obtained if comparisons were made among randomized patients with consensus-diagnosed HGG treated with pCV versus institutionally diagnosed HGG treated with 8-in-1. B. Comparison between HGG confirmed by reviewer #4 and treated with pCV versus HGG confirmed by reviewer #5 and treated with 8-in-1. C. HGG confirmed by reviewer #4 and randomized to receive either regimen. D. HGG confirmed by reviewer #5 and randomized to receive either regimen.

associated with outcome in this cohort. The clinical factor that had the strongest association with outcome was extent of tumor resection, graded as $\geq 90\%$ or $< 90\%$ in the initial clinical report (Finlay et al., 1995; Wisoff et al., 1998). MIB-1 labeling index ($< 18\%$ vs. $18\%–36\%$ vs. $\geq 36\%$) (Pollack et al., 2002a) and p53 expression (expression similar to normal brain vs. overexpression relative to normal brain) (Pollack et al., 2002b) also were associated strongly with outcome, independent of tumor histology. Patients with consensus diagnoses of HGG who had less extensive resections had significantly worse 5-year progression-free survival than those who had extensive resections (Fig. 5A: stratified by histology, $P = 0.0004$), which agreed with observations using less stringent inclusion criteria. Similarly, patients with consensus diagnoses of HGG whose tumors had relatively low MIB-1 indices ($< 18\%$) had more favorable prognoses than those with high ($18\%–36\%$) or very high ($\geq 36\%$) indices (Fig. 5B: $P = 0.02$), and children whose tumors did not overexpress p53 had better prognoses than those with overexpressing tumors (Fig. 5C: $P < 0.01$).

Discussion

Classification of gliomas in general, and of childhood gliomas in particular, is associated with frequent discrepancies between reviewers (Childhood Brain Tumor

Consortium, 1989). The current analysis of CCG-945 showed that regardless of central review criteria, cases were frequent in which the original institutional diagnosis was viewed as ineligible on further review. There also was frequent discordance among reviewers, which suggests that despite availability of established, widely recognized diagnostic criteria (Kleihues et al., 1993), applying such guidelines to a topographically heterogeneous tumor can be challenging, even in the best of hands. The recent observation that molecular and biological markers can help refine prognostic assessments for these tumors suggests that eventually such supplemental analyses may improve tumor classification and risk-based treatment stratification (Pollack et al., 1997, 2002a, b); however, as with histologic classification criteria, no factors to date absolutely predict outcome.

In view of these results, clinicians are faced with a quandary in deciding how to integrate pathological classification into the design of prospective clinical studies. One approach is intent-to-treat, whereby all patients eligible by institutional criteria are included in outcome analyses for a given treatment. A theoretical advantage of this design is that it mimics the real-world situation, in which central review of histological data is rarely pursued, and distributes bias among the broadest possible pool of reviewers, rather than potentially skewing the analysis by basing it on review criteria of a single expert

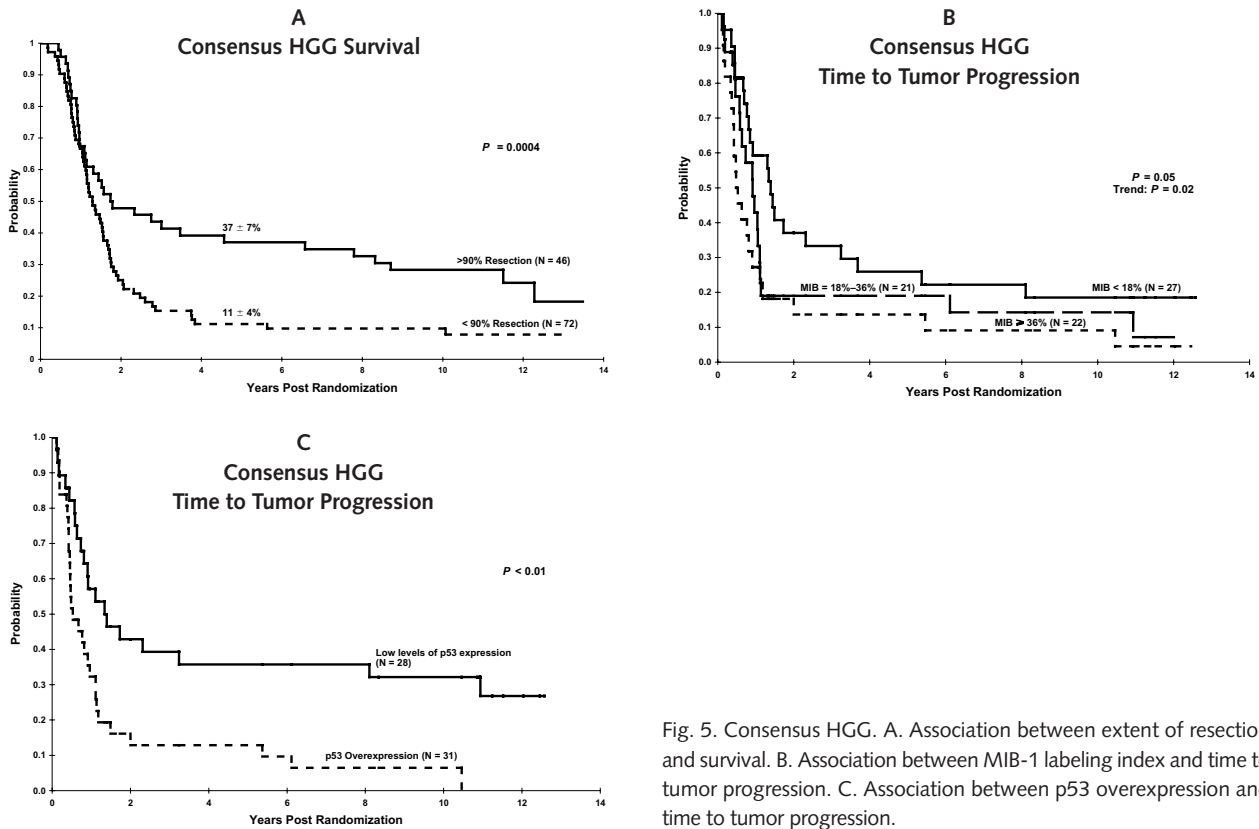


Fig. 5. Consensus HGG. A. Association between extent of resection and survival. B. Association between MIB-1 labeling index and time to tumor progression. C. Association between p53 overexpression and time to tumor progression.

examiner. However, an intent-to-treat design allows substantial classification errors in situations in which there is a tendency for more stringent inclusion criteria to be applied by experts, which can influence efficacy assessments of treatment approaches. In this context, the high frequency of discordance in the current study in part reflects evolution in classification criteria for pediatric gliomas during the past 2 decades. It is likely that many cases considered ineligible by the consensus review would have been considered ineligible by the institutional neuropathologist had they been rereviewed institutionally using contemporary criteria, an exercise confirmed at the first author's center in recent years, using a subset of this cohort (Pollack et al., 1997). It is also important to emphasize that the pathology reviews in this study were done in a blinded fashion, whereas institutional pathological diagnoses are typically established in the context of available clinical and imaging results and operative findings, which may introduce distinct insights and biases into the classification process. It is reassuring that the principal conclusions of CCG-945 were unaffected (i.e., the lack of efficacy difference between regimens and clinical and biological factors correlated with outcome) regardless of whether institutional, single-examiner central review, or consensus review diagnoses were applied. However, this observation may indicate that factors that influenced outcome in the consensus Not HGG group, many of which were low-grade gliomas, were fortuitously similar to those relevant in HGGs. For example, the strong effect of resection extent on outcome of consensus-confirmed childhood HGGs is mirrored by the

strong association between resection extent and outcome among a centrally reviewed cohort of 516 children with low-grade gliomas (Sanford et al., 2002). Thus, it is not surprising that such association would be apparent regardless of whether institutional or review diagnoses were used.

Although these observations do not refute intent-to-treat study analysis for HGGs, they highlight diagnostic classification issues that must be considered in the evaluation of such studies. Most biostatisticians prefer this analysis as the primary treatment comparison in a randomized trial because it is based on a sample taken from the population to which the inference is intended, but also acknowledge that a secondary analysis should be done that compares only those patients who were treated exactly as prescribed by the protocol, if possible to discern (Peto et al., 1976, 1977). This second subset analysis evaluates efficacy of treatments in an ideal setting, excluding cases not treated according to protocol. In the current study, the ideal setting also reflected cases in which independent histopathologic reviews from an experienced panel yielded consensus diagnoses of HGG, excluding institutionally eligible cases in which consensus diagnoses were deemed to be ineligible. The conduct of the CCG-945 study in a way that facilitated both analyses provided important insights into the diagnostic difficulties involving these tumors. Our results should alert institutional investigators to diagnostic challenges in such cases and indicate that outcomes may be influenced by cases that pose dilemmas. These findings also should sensitize readers of the medical literature to be

cautious when interpreting results for HGG based on single-institution studies and to critically evaluate studies in which inclusion criteria are not articulated.

Our results also highlight the importance of incorporating prospective central review in the design of future studies for HGGs to minimize inclusion of discordant cases. Although the primary conclusions of CCG-945 were not influenced by the review mechanism, the survival percentages calculated for the eligible cohort defined by revised review or consensus diagnoses were substantially lower than those calculated for the institutionally eligible cohort. Review by a panel of experts does not guarantee absolute diagnostic accuracy, but it does ensure adherence to guidelines based on contemporary classification paradigms, and it highlights cases that may have received outlier diagnoses by one reviewer, an issue that can be resolved most efficiently by a multireviewer process. This design does not negate the importance of meticulous institutional diagnostic classification as an initial step in study entry, but incorporates the advantage of a tiered review process with a panel of reviewers to enhance identification of discordant cases. It has the added benefit of providing ongoing, real-time feedback to pathologists, including reviewers themselves. By eliminating discordant subgroups (which in the case of HGGs are usually low-grade glioma variants with much more favorable prognoses), panel review may improve a study's power to detect treatment effects. Central review also helps identify heretofore unrecognized histologic characteristics of treatment responders that may help refine tumor classification and guide stratification in subsequent studies.

Protocol-directed central review also provides an important safety measure for potential study participants. Given the poor survival of children with malignant gliomas who are treated with standard therapies such as the CCG-945 regimens, most new studies are examining innovative, potentially high-risk treatments such as intensive chemotherapy or novel agents. It is important to ensure that patients who are given such therapies meet the most rigorous histologic eligibility criteria possible. Given the findings of the current study, the use of an

expedited panel review process will be applied prospectively as a criterion for study eligibility in Children's Oncology Group HGG trials such as the recently opened ACNS0126 study. The panel review is done immediately after initial institutional review as a way of confirming histological eligibility prior to enrollment. This approach places a priority on optimizing the risk/benefit ratio for potential participants, and on evaluating efficacy in a clearly defined cohort.

Our results also highlight several caveats to be considered when attempting to compare findings of a new study with those of a historical control group. A practical problem is that clinical trials that incorporate different review mechanisms may promulgate a degree of interreviewer variation (and intrareviewer variation if criteria for review have changed over time) in classification of individual cases that may preclude valid cross-study comparisons. Thus, for a study design in which a historical control group is believed to be essential (as a practical alternative to including a control treatment arm in which outcome results may, at best, be suboptimal), it is critical to ensure that both populations are reviewed histologically by the same individual(s), avoiding interreviewer bias, and that such reviews be based on the same review criteria, minimizing intrareviewer bias as inclusion criteria evolve. Recognition of both potential pitfalls is essential to optimize the robustness of cross-study comparisons.

Acknowledgements

The authors recognize the collaboration of Judith Burnham, Ronald L. Hamilton, and Sydney D. Finkelstein (University of Pittsburgh) in the biological analyses of this cohort and Emiko J. Holmes and Richard Sposto (Children's Oncology Group) and Dana Wallace (St. Jude Children's Research Hospital) in the statistical analyses involving those studies. The late Laurence E. Becker was a contributor to this paper but died before its submission. We are thankful for his contribution.

References

- Aldape, K., Simmons, M.L., Davis, R.L., Miike, R., Wiencke, J., Barger, G., Lee, M., Chen, P., and Wrensch, M. (2000) Discrepancies in diagnoses of neuroepithelial neoplasms: The San Francisco Bay Area Adult Glioma Study. *Cancer* **88**, 2342–2349.
- Allen, J.C., Aviner, S., Yates, A.J., Boyett, J.M., Cherlow, J.M., Turski, P.A., Epstein, F., and Finlay, J.L. (1998) Treatment of high-grade spinal cord astrocytoma of childhood with "8-in-1" chemotherapy and radiotherapy: A pilot study of CCG-945. *J. Neurosurg.* **88**, 215–220.
- Childhood Brain Tumor Consortium. (1989) Intraobserver reproducibility in assigning brain tumors to classes in the World Health Organization diagnostic scheme. The Childhood Brain Tumor Consortium. *J. Neurooncol.* **7**, 211–224.
- Finlay, J.L., Boyett, J.M., Yates, A.J., Wisoff, J.H., Milstein, J.M., Geyer, J.R., Bertolone, S.J., McGuire, P., Cherlow, J.M., and Tefft, M. (1995) Randomized phase III trial in childhood high-grade astrocytoma comparing vincristine, lomustine, and prednisone with the eight-drugs-in-1-day regimen. Children's Cancer Group. *J. Clin. Oncol.* **13**, 112–123.
- Geyer, J.R., Finlay, J.L., Boyett, J.M., Wisoff, J., Yates, A., Mao, L., and Packer, R.J. (1995) Survival of infants with malignant astrocytomas: A report from the Children's Cancer Group. *Cancer* **75**, 1045–1050.
- Kalbfleisch, J.D., and Prentice, R.I. (1980) *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons.
- Kaplan, E.L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481.

- Kleihues, P., Burger, P.C., and Scheithauer, B.W. (1993) *Histological typing of tumours of the central nervous system*, 2nd ed. International Histological Classification of Tumours **21**. Berlin: Springer, pp. 11–16.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50**, 163–170.
- Papahill, P.A., Ramsay, D.A., and Del Maestro, R.F. (1996) Pleomorphic xanthoastrocytoma: Case report and analysis of the literature concerning efficacy of resection and the significance of necrosis. *Neurosurgery* **38**, 822–829.
- Pendergrass, T.W., Milstein, J.M., Geyer, J.R., Mulne, A.F., Kosnik, E.J., Morris, J.D., Heideman, R.L., Ruymann, F.B., Stuntz, J.T., and Bleyer, W.A. (1987) Eight drugs in one day chemotherapy for brain tumors: Experience with 107 children and rationale for preradiation chemotherapy. *J. Clin. Oncol.* **5**, 1221–1231.
- Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., and Smith, P.G. (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design *Br. J. Cancer* **34**, 585–612.
- Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., and Smith, P.G. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br. J. Cancer* **35**, 1–39.
- Pollack, I.F., Hamilton, R.L., Finkelstein, S.D., Campbell, J.W., Martinez, A.J., Sherwin, R.N., Bozik, M.E., and Gollin, S.M. (1997) The relationship between *TP53* mutations and overexpression of p53 and prognosis in malignant gliomas of childhood. *Cancer Res.* **57**, 304–309.
- Pollack, I.F., Hamilton, R.L., Burnham, J., Holmes, E.J., Finkelstein, S.D., Sposto, R., Yates, A.J., Boyett, J.M., and Finlay, J.L. (2002a) The impact of proliferation index on outcome in childhood malignant gliomas: Results in a multi-institutional cohort. *Neurosurgery* **50**, 1238–1244.
- Pollack, I.F., Finkelstein, S.D., Woods, J., Burnham, J., Holmes, E.J., Hamilton, R.L., Yates, A.J., Boyett, J.M., Finlay, J.L., and Sposto, R. (2002b) Expression of p53 and prognosis in children with malignant gliomas. *N. Engl. J. Med.* **346**, 420–427.
- Qu, Y., Tan, M., and Kutner, M.H. (1996) Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52**, 797–810.
- Sanford, A., Kun, L., Sposto, R., Holmes, E., Wisoff, J.H., Heier, L., and McGuire-Cullen, P. (2002) Low-grade gliomas of childhood: Impact of surgical resection. A report from the Children's Oncology Group. *J. Neurosurg.* **96**, 427–428.
- Schwartz, D., and Lellouch, J. (1967) Explanatory and pragmatic attitudes in therapeutical trials. *J. Chronic Dis.* **20**, 637–648.
- Tsiatis, A. (1990) Analysis and interpretation of trial results: Intent-to-treat analysis. *J. Acquir. Immune Defic. Syndr.* **3** (Suppl. 2), S120–S123.
- Wisoff, J.H., Boyett, J.M., Berger, M.S., Brant, C., Li, H., Yates, A.J., McGuire-Cullen, P., Turski, P.A., Sutton, L.N., Allen, J.C., Packer, R.J., and Finlay, J.L. (1998) Current neurosurgical management and the impact of the extent of resection in the treatment of malignant gliomas of childhood: A report of the Children's Cancer Group trial no. CCG-945. *J. Neurosurg.* **89**, 52–59.
- Yung, W.K., Prados, M.D., Yaya-Tur, R., Rosenfeld, S.S., Brada, M., Friedman, H.S., Albright, R., Olson, J., Chang, S.M., O'Neill, A.M., Friedman, A.H., Bruner, J., Yue, N., Dugan, M., Zaknoen, S., and Levin, V.A. (1999) Multicenter phase II trial of temozolomide in patients with anaplastic astrocytoma or anaplastic oligoastrocytoma at first relapse. Temodal Brain Tumor Group [erratum in *J. Clin. Oncol.* **17**, 3693, 1999]. *J. Clin. Oncol.* **17**, 2762–2771.