

Proposal for the Use of Progression-Free Survival in Unblinded Randomized Trials

Boris Freidlin, Edward L. Korn, Sally Hunsberger, Robert Gray, Scott Saxman, and Jo Anne Zujewski

ABSTRACT

Progression-free survival is an attractive end point for clinical trials when an overall survival end point may be confounded by additional treatments administered after progression. When a trial is performed in an unblinded manner, however, there is the potential for bias between the treatment arms because of the subjective aspects of the progression end point. We discuss the magnitude of this potential bias and suggest methods for lessening it. We propose the carrying forward of any progression information to two designated time points for the statistical analysis for trials that are not blinded. This proposal, possibly combined with central review of progression scans for these two time points, essentially eliminates any bias, with little risk of major efficiency loss compared with using the reported progression times.

J Clin Oncol 25:2122-2126. © 2007 by American Society of Clinical Oncology

INTRODUCTION

Even though overall survival (OS; time from random assignment to death from any cause) is the most error-free, and typically the most clinically relevant end point in cancer clinical trials, it may not always be the most scientifically relevant end point. In particular, the effect of the experimental treatment on progression-free survival (PFS; time from random assignment to progression of disease or death from any cause) may be all that can be reasonably expected for a trial to demonstrate or may be adequate to move the treatment forward for testing in an earlier stage of disease. For example, consider a trial of a new first-line treatment versus a standard treatment for metastatic breast cancer, for which multiple salvage regimens are available. The median OS in the control arm may be more than 2 years, whereas the median PFS in the control arm may be 6 months. After patients progress, they will be offered further treatments (including the possibility of the experimental treatment to progressing control-arm patients). Even if the experimental treatment is more effective than the standard treatment, it may not be reasonable to expect to be able to demonstrate this in terms of OS for any reasonable sample size in this setting because of these salvage therapies. Nevertheless, showing benefit in terms of PFS would suggest interesting biologic activity of the experimental treatment (leading possibly to the development of other treatments) as well as the possibility of moving this treatment forward to be tested in an early-stage breast cancer setting. In general, as more

effective therapies become available for second-line treatment of metastatic disease, it will become harder to evaluate first-line treatments for metastatic disease using OS.¹⁻³

PFS is a “softer” end point than OS, but this by itself does not make it scientifically or clinically irrelevant. A potential problem with PFS, however, occurs when the patients and investigators are not blinded to the treatment assignment. In this situation, it is conceivable that progression events will be declared earlier in one arm than in the other, even when there is really no difference in efficacy between the arms. For example, consider a trial comparing a promising experimental agent with observation, with a cross-over from observation to the agent at progression in the control arm. It is conceivable that a patient with a given clinical course would have his progression declared at a later time if he were in the experimental arm than if he were in the observation arm. There are two possible explanations for this differential reporting of events (assessment or evaluation bias), both related to the desire of the patient and treating physician to get a patient on an active therapy regimen as quickly as possible. In the first case, the patient would have his evaluation for progression performed at the same time regardless of treatment arm, but there would be enough subjectivity in the evaluation that the investigator may declare a control-arm case a progression, whereas he would not have declared it a progression in the experimental arm (perhaps only in a very small group of borderline cases). This type of differential reporting could be ameliorated by having a

From the Biometric Research Branch and the Clinical Investigations Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD; Eastern Cooperative Oncology Group, Harvard School of Public Health, Boston, MA; and the Peace Corps, Washington, DC.

Submitted October 21, 2006; accepted February 26, 2007.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Address reprint requests to Boris Freidlin, PhD, Biometric Research Branch, EPN-8122, National Cancer Institute, Bethesda, MD 20892; e-mail: freidlinb@ctep.nci.nih.gov.

© 2007 by American Society of Clinical Oncology

0732-183X/07/2515-2122/\$20.00

DOI: 10.1200/JCO.2006.09.6198

central blinded review of progression documentation (and a review of a subset of documentation for patients who do not progress at an evaluation time).

A second explanation for the differential reporting of events, and the focus of this article, is that patients may be formally evaluated at a later time for a suspected progression when they are in one treatment arm rather than the other treatment arm.⁴ We denote this as evaluation-time bias. For example, consider a patient reporting a symptom possibly related to progression at a visit that does not coincide with a protocol-specified radiologic progression evaluation time. This patient may be more likely to be formally evaluated for progression with diagnostic scans when he is in the control arm than when in the experimental arm. In general, bias in differential reporting of progression may be present even when the control arm contains active therapy if the experimental arm includes additional promising agents. It is also possible for there to be bias in the opposite direction (ie, progression scans earlier in the experimental arm) if there is greater toxicity with the experimental arm.

In addition to bias resulting from differential reporting of events, there is also the potential bias caused by more patients withdrawing from the trial (without documented progression) from the control arm than the experimental arm. (The original motivation for using placebos in controlled trials was to avoid this.⁵) If OS were the end point, then these patients should be followed for death even though they are no longer receiving protocol treatment and may have received other treatments. With PFS as the end point, however, it may be difficult or impossible to document progression in patients who have left the trial. The statistical analysis of data from these patients, who are effectively lost to follow-up, would have to be censored at the time when the patients left the trial. With differential withdrawal rates between the arms, this can lead to a bias in PFS comparisons (attrition bias⁶). In what follows, we assume that the number of patients lost to follow-up is small, and focus on evaluation-time bias.

How big a potential problem is evaluation-time bias? We evaluate this in the next section through computer simulations. In the third section, we suggest ways to lessen the effects of this potential bias, and offer a proposal for trials for which blinded execution is impractical. We end with some recommendations.

ESTIMATING THE POTENTIAL EFFECTS OF EVALUATION-TIME BIAS

We assume a simplified statistical model to assess the potential effects of evaluation-time bias (Appendix, online only). We assume that each patient has a true progression time. By this, we mean the time that the tumor has grown enough so to satisfy the radiologic criteria for progression (eg, 20% increase in the longest maximum dimension). We refer to this as the “true” progression time because it is the time when the tumor has progressed, whether or not a scan has been performed at that time to document it. In practice, the progression time recorded for analysis is at a time later than the true progression time, either at a protocol-specified scan time or at a scan requested by the treating oncologist because of patient symptoms. We assume that patient visits are roughly monthly, and that scheduled scans for progression are every second or third visit. (We assume that the protocol-specified scheduled scan times are the same in the treatment arms; otherwise, this would be an obvious source of bias.⁷) The sensitivity and specific-

ity of scan examination are specified in terms of a 10% false-positive rate (the probability of declaring progression in progression-free patients) and a 10% false-negative rate (the probability of missing a true progression). If a scheduled scan detects progression, then the patient's documented progression time is the time from random assignment to this scan time. If the patient has truly progressed at the time of a visit that is not a scheduled scan time, we assume that there is a certain probability that a scan will be performed (and detect progression). This probability can be modeled as being higher in the control arm than in the experimental arm.

The differential modeling for recording of progression times between the experimental and control treatment arm will lead to an observed PFS difference between the arms even when there is no true difference between the arms. In Table 1, we consider six simulations for trials with scheduled scan times at every third monthly visit. In simulation 1, the median time to true progression is 1.5 months, and the probability of detecting a true progression at a nonscan monthly visit is 60% in the control arm and 20% in the experimental arm. The differential detection probabilities lead to a difference in the recorded median PFS (2.6 v 3.2 months), and an estimated hazard ratio different from 1 (simulated median = 1.34, simulated quartiles = 1.21 and 1.48). In 52% of the simulated data sets, the one-sided *P* value is less than .025; if there were no bias, this should happen in 2.5% of the data sets.

Simulation 2 is identical to simulation 1, except that the probability of detecting a progression at a nonscan visit for the control arm has been lowered to 40% from 60%. This lessens the bias in the recorded PFS distributions, although the *P* values are still less than .025 an unacceptably large 19% of the time. Simulations 3 and 4, and simulations 5 and 6, are identical to simulations 1 and 2, except that the median time to true progression has been set at 2.5 and 5.5 months, respectively. The bias is lower in these simulations; what affects the bias is the scheduled scanning frequency relative to the median time to true progression. This is also seen in the simulations 7 through 12 (Table 2), which are identical to simulations 1 through 6 except that the scheduled scanning frequency has been taken as every second instead of every third monthly visit.

If the probability of progression detected at a nonscan visit is different between the two treatment arms, and even if the null hypothesis is true, one would expect the proportion of progressions recorded at nonscan visits to be different in the two arms; this can be observed in Tables 1 and 2. Thus, observing between-arm difference in proportions of nonscan-visit progressions is an indication of potential evaluation-time bias. Determining whether a particular scan for progression was at a scheduled scan time may not be straightforward, but it would be important to attempt to obtain this information. We note that the proportion of progressions recorded at nonscan visits can be different between the treatment arms even without evaluation-time bias if the experimental treatment is truly better than the control treatment. However, this effect is quite small (results not shown).

LESSENING THE POTENTIAL EFFECTS OF DIFFERENTIAL REPORTING

If the trial can be performed as a placebo-controlled double-blind trial and the blinding is effective, there is no possibility of evaluation-time bias. In some cases (eg, an experimental treatment involving

Table 1. Simulation of Trial Results When the Null Hypothesis of No True PFS Differences Is True and Progression Scans Are Scheduled for Every Third Monthly Patient Visit

Simulation	Median Time to True Progression (months)	Treatment Arm	%		Simulated Median of Recorded Median PFS (months)	Percentiles of Hazard Ratio			Proportion of One-Sided $P < .025^*$
			Probability True Progression Detected at Nonscan Visit	Proportion of Progressions Recorded at Nonscan Visits		25th	50th	75th	
1	1.5	Control	60	61	2.6	1.21	1.34	1.48	.52
		Exp	20	26	3.2				
2	1.5	Control	40	46	2.9	1.06	1.17	1.29	.19
		Exp	20	26	3.2				
3	2.5	Control	60	55	3.3	1.08	1.19	1.32	.23
		Exp	20	23	3.6				
4	2.5	Control	40	41	3.4	1.00	1.10	1.21	.10
		Exp	20	23	3.6				
5	5.5	Control	60	45	5.5	0.99	1.09	1.21	.09
		Exp	20	19	6.0				
6	5.5	Control	40	34	5.8	0.95	1.05	1.16	.05
		Exp	20	19	6.0				

NOTE. A 100-patient per arm design was simulated; similar results apply across typical phase III sample sizes.

Abbreviations: PFS, progression-free survival; Exp, experimental.

*Log-rank test using recorded progression times.

hospitalization *v* a nontoxic standard therapy), blinding is not possible. In other cases, blinding is possible but may not be practical because of patient inconvenience. If PFS is only a secondary end point, it may not be considered appropriate to ask patients to return to the clinic for an intravenous placebo, especially if this must be done for an extended period of time. We note that it is sometimes said that blinding should not be used in a trial because a frequent toxicity in one of the treatment arms will effectively unblind the trial. However, unless the toxicity is seen uniformly

with one treatment and not the other, we do not agree. For example, even if 50% of the patients in the experimental arm have a treatment-specific toxicity, there will be enough doubt about the treatment assignment for the other patients to substantially lessen any potential evaluation-time bias.

For trials that are not blinded, more frequently scheduled progression evaluations will lessen any evaluation-time bias. For example, scheduling monthly progression scans would make the bias clinically irrelevant. However, formal monthly evaluations are infeasible in

Table 2. Simulation of Trial Results When the Null Hypothesis of No True PFS Differences Is True and Progression Scans Are Scheduled for Every Second Monthly Patient Visit

Simulation	Median Time to True Progression (months)	Treatment Arm	%		Simulated Median of Recorded Median PFS (months)	Percentiles of Hazard Ratio			Proportion of One-Sided $P < .025^*$
			Probability True Progression Detected at Nonscan Visit	Proportion of Progressions Recorded at Nonscan Visits		25th	50th	75th	
7	1.5	Control	60	38	2.3	1.05	1.15	1.27	.17
		Exp	20	13	2.4				
8	1.5	Control	40	26	2.4	0.98	1.08	1.18	.07
		Exp	20	13	2.4				
9	2.5	Control	60	34	3.0	0.99	1.09	1.20	.08
		Exp	20	12	3.4				
10	2.5	Control	40	23	3.2	0.95	1.05	1.15	.05
		Exp	20	12	3.4				
11	5.5	Control	60	26	4.8	0.94	1.04	1.15	.05
		Exp	20	9	5.0				
12	5.5	Control	40	18	4.9	0.92	1.02	1.13	.03
		Exp	20	9	5.0				

NOTE. A 100-patient per arm design was simulated; similar results apply across typical phase III sample sizes.

Abbreviations: PFS, progression-free survival; Exp, experimental.

*Log-rank test using the recorded progression times.

most situations. The relationship between the evaluation frequency and the potential bias needs to be considered relative to the median PFS; a frequency of every 3 months is potentially more problematic in terms of evaluation-time bias when the median PFS is 4 months than when it is 9 months.

Another strategy is to compare PFS rates at a single time point. For example, one can compare 6-month PFS rates between the treatment arms. Instead of using the reported progression time of each patient, one uses only the binary indicator of whether the patient has progressed by 6 months. It is important, if one uses this strategy, that patients who have not progressed by 6 months have a formal evaluation for progression at 6 months. (We assume throughout this discussion that symptomatic progressions need to be formally documented by scans.) This single-point comparison essentially eliminates the evaluation-time bias, but may entail a loss of statistical power. In particular, if the PFS curves satisfy a proportional hazards assumption, then larger sample sizes are required with the single-point procedure to have the same power as a trial that uses the actual reported progression times.⁸⁻¹⁰ However, the investment in larger sample sizes may be worthwhile if the reported progression times lead to questionable results because of potential evaluation-time bias. Without a proportional hazards assumption, use of a single-point analysis can result in very little power in some cases, or in higher power in other cases, depending on the relative shape of the PFS curves and where the time point is chosen.¹¹ For example, if the PFS curves separate to a maximum distance apart at 4 months and come back together at 8 months, then comparison of PFS rates at 4 months (8 months) has greater (no) power compared with the use of the actual recorded progression times.

To improve on the performance of comparing PFS rates at a single time point, we propose the following generalization: (a)

schedule two scan times, (b) ensure that patients are formally evaluated at these two scan times, and (c) for progressions that are documented at nonscheduled scan times and for deaths, use for the statistical analysis the next scheduled scan time as the event time. That is, if scheduled scan times are 3 months and 6 months, then a documented progression at 4 months would be changed to 6 months for the statistical analysis. Of course, because patients will not be seen at exactly a specified date, the protocol should specify the range of dates acceptable for a scheduled scan time (eg, 6 months \pm 2 weeks after the randomization date). The analysis consists of presenting the PFS rates at the two scan times, and calculating a *P* value based on the grouped data (Appendix). This two-point procedure, like the procedure that compares PFS rates at single time point, essentially eliminates any evaluation-time bias. However, unlike the single-point procedure, if the two scan times are chosen well, there is little risk of major power loss compared with using the actual reported progression times. In particular, we recommend choosing the two scan times to be approximately the median PFS and twice the median PFS of the control arm. In addition, all nonprogressing patients should have a follow-up time of at least the second scan time to ensure that they have their second formal progression evaluation. For example, if the median PFS in the observation arm is expected to be 4 months, then the evaluation scans would be scheduled at 4 and 8 months, and all nonprogressing patients should be followed for at least 8 months.

The relative performance of the procedures is illustrated in Table 3. Under proportional hazards (simulations 13 through 18) the two-point procedure has only a marginal loss in power relative to the log-rank test, while providing considerable improvement over single-point analyses. When treatment effect dissipates over time (simulation 19) or when treatment effect is delayed (simulation 20)

Table 3. Power Comparison

Simulation	Type of Treatment Effect	Median Observed Control	Power			
			Single-Point Procedure		Two-Point Procedure	Log-Rank (recorded progression times)
			At 1 Median	At 2 Medians		
Proportional hazards,* hazard ratio 1.66, scan visit every 2 months						
13	Median PFS 1.5 v 2.5 months	2.4	0.57	0.53	0.73	0.80
14	Median PFS 2.5 v 4.15 months	3.4	0.54	0.57	0.71	0.77
15	Median PFS 5.5 v 9.13 months	5.0	0.42	0.58	0.62	0.59
Proportional hazards,* hazard ratio 1.66, scan visit every 3 months						
16	Median PFS 1.5 v 2.5 months	3.2	0.58	0.40	0.73	0.80
17	Median PFS 2.5 v 4.15 months	3.6	0.55	0.55	0.71	0.81
18	Median PFS 5.5 v 9.13 months	6.0	0.48	0.57	0.67	0.69
Departures from proportional hazards,† scan visit every 2 months						
19	Early effect dissipates over time‡	3.4	0.77	0.07	0.50	0.34
20	Delayed (lag) effect§	3.4	0.39	0.76	0.75	0.78

NOTE. The single-point procedure at 1 median compares PFS rates at the median PFS on the control arm. The single-point procedure at 2 medians compares PFS rate at twice the median PFS on the control arm. The two-point procedure is based on the median PFS and twice the median PFS on the control arm. The log-rank test is using the recorded progression times.

Abbreviation: PFS, progression-free survival.

*The true progression distributions have proportional hazards; because of the false positive and negatives errors, recorded times to progression do not have proportional hazards.

†Simulated using piecewise exponential distribution on the experimental arm and exponential with median of 2.5 months on the control arm.

‡PFS curves separate and then come together: hazard ratio = 2 in the first 3.5 months, hazard ratio = .75 after the first 3.5 months.

§Lag in effect: hazard ratio = 1 in the first 1.5 months, hazard ratio = 2.5 after the first 1.5 months.

the two-point analysis preserves most of the power of the optimal single-point analysis.

In theory, more than two scan times could be specified (eg, scans every 3 months for the first year and every 6 months thereafter), with progressions again being carried forward to the next scheduled scan time for analysis purposes.^{10,12} This would lessen the possibility of power loss of using two scheduled scans (compared with actual reported progression times). However, it may not be practical to ensure that patients are formally evaluated with scans at more than two exactly specified time points (eg, patients' clinic visits may not exactly follow a regular schedule).

In addition to essentially eliminating evaluation-time bias, the proposal to use two scheduled scan times also makes it easier to perform central review of progression scans (positive and negative), because there will be at most two scan times for each patient. The proposal should, therefore, also help eliminate potential bias resulting from subjective reading of the scans. When calculating the sample size for a trial using this proposal, the fact that the grouped data will be used for the analysis should be taken into account using either asymptotic results¹¹ or simulations.

Finally, the potential bias caused by control-arm withdrawal (as discussed in the Introduction) can be lessened by including in the trial design a cross-over to the experimental treatment at the time of documented progression. However, the inclusion of such a cross-over may increase the potential for evaluation-time bias, which could be mitigated with the methods discussed previously herein.

RECOMMENDATIONS

When PFS is the primary end point or an important secondary end point, use placebo-controlled double-blinded trials when practical (with log-rank analyses). When this is not practical, then either space the progression evaluation times relatively close together or restrict the data analysis to two scheduled evaluation times at approximately the

median PFS and twice the median PFS of the control arm (as described in the section Lessening the Potential Effects of Differential Reporting). If it is possible to ensure that patients will be evaluated with scans at additional times, then more than two scheduled evaluation times could be considered. To minimize the impact of patients' withdrawing from the trial before documentation of progression, every attempt should be made to establish their progression status at the scheduled evaluation times (even though they may have received off-study treatments). If this is not possible, then there is a chance of significant attrition bias in the trial results.

For help in assessing the magnitude of any potential evaluation-time bias, record for each progressing patient whether the reported progression was at a protocol-specified evaluation time or whether it resulted from an evaluation instigated by patient symptoms. For help in assessing the magnitude of any potential attrition bias, the proportion of patients withdrawing from the trial before documented progression (before the last scheduled evaluation time) should be recorded for both treatment arms.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The authors indicated no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Conception and design: Boris Freidlin, Edward L. Korn, Sally Hunsberger, Robert Gray, Jo Anne Zujewski, Scott Saxman
Collection and assembly of data: Boris Freidlin, Edward L. Korn, Sally Hunsberger, Robert Gray, Jo Anne Zujewski, Scott Saxman
Data analysis and interpretation: Boris Freidlin, Edward L. Korn, Sally Hunsberger, Robert Gray, Jo Anne Zujewski, Scott Saxman
Manuscript writing: Boris Freidlin, Edward L. Korn, Sally Hunsberger, Robert Gray, Jo Anne Zujewski, Scott Saxman
Final approval of manuscript: Boris Freidlin, Sally Hunsberger, Robert Gray, Jo Anne Zujewski, Scott Saxman

REFERENCES

- De Gramont A, Figer A, Seymour M: Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. *J Clin Oncol* 18:2938-2947, 2000
- Di Leo A, Bleiberg H, Buyse M: Overall survival is not a realistic end point for clinical trials of new drugs in advanced solid tumors: A critical assessment based on recently reported phase III trials in colorectal and breast cancer. *J Clin Oncol* 21:2045-2047, 2003
- Goldberg P: FDA approves Gemzar for ovarian cancer, contradicting ODAC recommendation. *Cancer Lett* 32:1-3, 2006
- Buyse M: Cornerstones of a well-designed phase III trial. *Eur J Cancer Suppl* 1:67-75, 2003
- Kaptchuk TJ: Intentional ignorance: A history of blind assessment and placebo controls in medicine. *Bull Hist Med* 72:389-433, 1998
- Jüni P, Altman DG, Egger M: Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 323:42-46, 2001
- Williams G, He K, Chen G, et al: Operational bias in assessing time to progression (TTP). *Pros Am Soc Clin Oncol* 20:244a, 2002 (abstr 975)
- Cuzick J: The efficiency of the proportions test and the logrank test for censored survival data. *Biometrics* 38:1033-1039, 1982
- Gail MH: Applicability of sample size calculations based on a comparison of proportions for use with the logrank test. *Control Clin Trials* 6:112-119, 1985
- Stone A, Wheeler C, Carroll K, et al: Optimizing randomized phase II trials assessing tumor progression. *Contemp Clin Trials* 28:146-152, 2007
- Berger A, Wallenstein S: The effect of grouping on the power of the Mantel-Haenszel test for the comparison of survival rates. *Stat Prob Lett* 16:19-25, 1993
- U.S. Food and Drug Administration: Guidance for Industry, Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics (DRAFT GUIDANCE), April 2005. www.fda.gov/cder/Guidance/6592dft.pdf
- Mantel N, Haenszel W: Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 22:719-748, 1959
- SAS Institute Inc. SAS/STAT User's Guide, Version 8. Cary, NC, SAS Institute Inc, 2000

Appendix

The Appendix is included in the full-text version of this article, available online at www.jco.org. It is not included in the PDF version (via Adobe® Reader®).

Appendix

Simulation Details

For [Tables 1](#) and [2](#), we simulate true progression times as having an exponential distribution with the same median for the control and experimental arms; that is, the treatment is ineffective in improving the true PFS. In practice, a small number of deaths are recorded without documented progression, with the death time being taken as the event time. We have not modeled these cases in the simulations.

Each monthly visit follows the previous visit by a time U , where U has a uniform distribution on 2 to 6 weeks. Thus, on average, the visits are 4 weeks apart. The probability that a true progression is detected and recorded at a nonscan monthly visit is a certain probability for the control arm and a possibly different probability for the experimental arm.

For [Tables 1](#) and [2](#), each simulated data set contains 100 patients in each treatment arm accrued uniformly over 24 months with additional 6 months of follow-up. For [Table 3](#), each simulated data set contains 100 patients in each treatment arm accrued uniformly over 24 months with additional follow-up of $2 \times M$ months, where M is approximately the median PFS observed in the control treatment arm. For [Table 3](#), the probability of detecting a true progression at a nonscan monthly visit was 20% for both arms. Each simulation is based on 20,000 simulated data sets.

Analysis of Grouped Data

For two time points, the data can be summarized as in Table A1. If there are no missing scheduled scans, there would only be category A, B, and C patients. Category D includes patients who are progression free at their time point 1 scan but who do not have a documented scan at time point 2. Note that this category would include both patients who withdrew before time point 2 and are not re-evaluated and patients documented to be progression free at a later time than time point 2 (eg, a patient missing his scheduled [time point 2] scan at 12 months but with a scan at 14 months that documents his lack of progression). Even though one might believe that it is highly likely that this patient would have documented nonprogression at 12 months if his scan had been done at this time, it could lead to a potential bias between the treatment arms if this patient was classified in category A instead of category D (eg, if more patients in the experimental arm missed their time point 2 evaluation than did patients in the control arm). Categories E and F also represent patients with missing scheduled scan information. Ideally, the proportion of patients in categories D, E, and F combined should be very small (eg, less than 10%).

Based on the numbers in Table A1, one can calculate the outcome data for time point 1, and the outcome data for time point 2 after removing patients who are known to have the event at time point 1 (Table A2). The P value for testing the null hypothesis that the PFS distributions are the same for the two treatments is then calculated using a Mantel-Haenszel statistic¹³ applied to these

two tables. Many statistical software packages will perform this calculation (eg, Proc Freq in SAS [SAS Institute, Cary, NC]).¹⁴ If the design uses more than two scheduled scans, then there will still be a table for the outcome data for each time point and the Mantel-Haenszel statistic can still be applied.

Table A1. Table for Summary Progression-Free Survival Data Grouped at Two Time Points

Category	Patient Condition	No. of Patients	
		Control Arm	Experimental Arm
A	Alive and progression free at time point 2	x_A	y_A
B	Alive and progression free at time point 1, and having progressed or died by time point 2	x_B	y_B
C	Progressed or died by time point 1	x_C	y_C
D	Alive and progression free at time point 1, with status unknown at time point 2	x_D	y_D
E	Status unknown at time point 1, and having progressed or died by time point 2	x_E	y_E
F	Status unknown at time points 1 and 2	x_F	y_F

Table A2. Summary of Outcome Data for Time Points 1 and 2 (derived from Table A1)

Treatment Arm	Event	
	No	Yes
Time point 1		
Control	$x_A + x_B + x_D$	x_C
Experimental	$y_A + y_B + y_D$	y_C
Time point 2		
Control	x_A	$x_B + x_E$
Experimental	y_A	$y_B + y_E$

*Outcome data for time point 2 after removing patients who are known to have the event at time point 1.