Missing Data and Measurement Variability in Assessing Progression-Free Survival Endpoint in Randomized Clinical Trials

Rajeshwari Sridhara, Sumithra J. Mandrekar, and Lori E. Dodd

Abstract

Progression-free survival (PFS) is frequently used as the primary efficacy endpoint in the evaluation of cancer treatment that is considered for marketing approval. Missing or incomplete data problems become more acute with a PFS endpoint (compared with overall survival). In a given clinical trial, it is common to observe incomplete data due to premature treatment discontinuation, missed or flawed assessments, change of treatment, lack of follow-up, and unevaluable data. When incomplete data issues are substantial, interpretation of the data becomes tenuous. Plans to prevent, minimize, or properly analyze incomplete data are critical for generalizability of results from the clinical trial. Variability in progressive disease measurement between radiologists further contributes to data problems with a PFS endpoint. The repercussions of this on phase III clinical trials are complex and depend on several factors, including the magnitude of the variability and whether there is a systematic reader evaluation bias favoring one treatment arm particularly in open-label trials. Clin Cancer Res; 19(10); 2613–20. ©2013 AACR.

Introduction

Randomized controlled trials are the gold standard for providing conclusive evidence about treatment efficacy and for guiding patient management. Selection of an appropriate primary endpoint is fundamental to the validity and interpretation of a trial. Progression-free survival (PFS) is an important endpoint used in the evaluation of oncologic drug products in recent times for consideration of marketing authorization by the U.S. Food and Drug Administration (FDA). Although improving overall survival remains the ultimate goal of new cancer therapy, an intermediate endpoint such as PFS, a composite endpoint, defined as time from randomization to disease progression or death, is commonly used to evaluate the treatment effect of new oncologic products studied in randomized clinical trials (1).

In clinical practice, determination of disease progression is a complex process that includes clinical, radiologic, and other laboratory measurements for a given patient. Discussion of whether improving PFS represents a true patient benefit is beyond the scope of this article. In clinical trials conducted to establish efficacy and safety of cancer therapy, disease progression in nonhematologic tumors is most commonly determined using radiologic scans with specific disease progression criteria such as the Response Evaluation Criteria in Solid Tumors (RECIST) (2). These assessments are made at prespecified intervals corresponding to regularly scheduled visits. The frequency of these visits and scans is dictated by logistical issues such as cost and patient convenience as well as health issues such as exposure to radiation. As a result, the exact date of disease progression is never known, which is in clear contrast with survival data. Missing or incomplete data problems are common in a trial with a primary PFS endpoint arising from premature treatment discontinuation, missed or flawed assessments, change of treatment, and lack of follow-up.

Another concern relates to the variability in the disease evaluation assessments. Radiologic measurements have a variety of inadequacies, including measurement error, reader variability, and unevaluable scans, as discussed elsewhere in this CCR Focus section (3). Whether variability in progression determinations has a meaningful impact on interpretations of treatment effect and whether subjective bias favoring a treatment might be present and hence invalidate estimates of treatment effect have been evaluated (4–6). In open-label randomized clinical trials, there is further concern that the investigator- or site-radiologist progression assessments might be biased. For example, an investigator preference for the experimental treatment arm (compared with the control treatment arm) might result in progression calls that are consistently later in the experimental arm.

Authors' Affiliations: 1CDER/OTS/OB/DBV, U.S. Food and Drug Administration, Silver Spring; 2Biostatistics Research Branch, National Institute of Allergy and Infectious Disease, NIH, Bethesda, Maryland; and 3Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota

Corresponding Authors: Rajeshwari Sridhara, U.S. Food and Drug Administration, WO21, 10903 New Hampshire Avenue, Silver Spring, MD 20993, Phone: 301-796-1759; Fax: 301-796-9733; E-mail: rajeshwari.sridhara@fda.hhs.gov; and Lori E. Dodd, E-mail: doddl@mail.nih.gov.

doi: 10.1158/1078-0432.CCR-12-2938

©2013 American Association for Cancer Research.
compared with the control arm. This would lead to requiring confirmation of disease progression by an independent radiologic committee (IRC) or the blinded independent review committee (BICR), particularly in clinical trials that are likely to be submitted for regulatory consideration; however, these practices have several limitations.

In this article, we review the issues surrounding the use of PFS as an endpoint in clinical trials, particularly the missing or incomplete data problems, variability in the progression assessment, and limitations of BICR, offer some potential solutions, and discuss the potential impact on estimates of treatment effect.

**Missing or Incomplete Data for PFS Assessment**

In most clinical trials, patients enter the study at different points in time. Consequently, some patients may not experience the event of interest, such as progression or death, at the end of the study or at the time of analysis. Also, event times may not be observed for patients who are lost to follow-up or who withdraw from the study. For such patients, the time-to-event data is censored when conducting the standard time-to-event analysis. Such methods assume that the censoring mechanism is independent of the outcome. Censoring an observation has 3 forms, namely, right, left, and interval censoring:

- **Right censoring** occurs when the event of interest has not occurred at the time of the analysis and actual event time is greater than the observed time. In general, all time-to-event observations are subject to potential right censoring.

- **Left censoring** occurs when the disease state (event) is observed but when it began is unknown (e.g., time from seropositivity to AIDS).

- **Interval censoring** occurs when the event takes place between 2 time points. In this case, an imputed value is assigned to the time of the event.

Informative censoring is a separate issue that occurs when censoring is not independent of the outcome, and this will be discussed in a later section.

**Interval-censored data**

Interval censored data occurs when the exact time of an event is not directly observed but instead is known only to have occurred within a specific time interval, as is the case with progression assessment evaluated at fixed prespecified visit times. There are 2 primary approaches to dealing with interval-censored data: (i) parametric modeling (accelerated failure time); and (ii) nonparametric maximum likelihood estimation (NPMLE) methods (7). Standard survival methods (e.g., Kaplan–Meier curves, log-rank tests, and Cox proportional hazards regression models) must be modified to properly account for the interval censoring. Midpoint imputation (where the midpoint of the assessment interval is assigned as the failure time) or right censored imputation are commonly used to deal with interval-censored data; however, both of these approaches can lead to inflation of type I errors or false-positive conclusions. More recently, complex imputation methods have been developed for the analysis of interval-censored data. Several iterative procedures are used to compute the NPMLE, for example, an expectation-maximization algorithm that imputes the missing data using the conditional expectation given the observed data (8).

Qi and colleagues (9) conducted simulation studies to understand the impact of the length of time elapsed between the last progression-free scan and the progression date on time-to-progression estimates in advanced non–small cell lung cancer trials. PFS estimates using reported progression date (method 1) were highest due to the length of the assessment interval. Method 2 (1 day after the last progression-free scan) was the most conservative. Method 3 (midpoint between the last progression-free scan and reported progression date) and method 4 (nonparametric interval censoring) yielded similar results, lying between the estimates using method 1 and 2, as the majority of the disease progression occurred during treatment when frequent disease assessments (as is common in the setting of advanced lung cancer) were conducted (every 4, 6, 8, or 10 weeks). Analysis of randomized trials revealed that the trial conclusions remained unaffected by method of progression date determination.

The PhRMA working group also compared different censoring approaches and conducted simulations comparing log-rank test and interval censoring methods (10,11). These studies showed that there are no significant shifts in the estimated relative treatment effect sizes using different censoring approaches.

**Missed Assessments**

This type of incomplete data can occur due to missed visits, all target lesions not being measured, or unevaluable scans. For example, consider scenario 1, in which the fourth visit of tumor evaluation is missed and in the next visit progression is documented. In Scenario 2, assume both third and fourth visits are missed and progression is documented in the fifth visit. In both of these scenarios, the true progression could have occurred during the missed time period. In the PFS analysis, when incomplete data are due to missed visits, several options can be considered. The observation for a given patient with documented progression after multiple missed assessments can be (i) recorded as an event when progression was documented, (ii) recorded as an event at the first missed assessment date, (iii) recorded as a progression event at the time point midway between progression assessments, (iv) censored at the last documented assessment with no progression, or (v) interval censored as disease progression at prespecified intervals when the exact time of progression cannot be ascertained (9). Patients who are lost to follow-up and whose disease status is unknown at the end of study or at the final assessment, are censored at the time of last tumor assessment with no documented progression as these instances are generally considered unrelated to treatment.
**Informative censoring**

When the censoring mechanism is not independent of the outcome, as in the case of censoring when the treatment is changed, we have informative censoring and such censoring can lead to biased estimates of the treatment effect. In this case, we will be making an assumption that 2 patients whose times to progression are censored at the same time have the same risk of an event, although one of them was censored due to discontinuation of treatment or change of protocol treatment. In addition, we make the assumption that both of these patients have the same risk of progression as those who remain on study. Such strong assumptions may not be valid.

There are different reasons for incomplete data with informative censoring (12) in PFS assessment that can commonly occur in clinical trials. The first reason can be broadly categorized as arising from missed assessments that can be due to acceptable/intolerable toxicity, premature treatment discontinuation/change of treatment, outcome not measured at the scheduled visit, or the radiologic scan not being evaluable. The second reason for informative censoring is atypical and occurs when blinded independent review is used to mitigate potential bias introduced by the investigator or local site reader.

**Incomplete Data Due to Treatment Discontinuation**

This type of incomplete data could be generally anticipated and associated problems minimized by continuing to assess tumor progression in patients who discontinue protocol treatment due to toxicity, change of treatment, or investigator-determined progression. However, it is recognized that this may not be practical in patients who enroll in other clinical trials, receive life-extending treatment such as surgery, or move to hospice care due to progression. This would also require the tumor assessments to be continued at the same frequency as specified in the protocol, which may not be convenient or cost effective. In oncology clinical trials, physicians generally continue to follow patients despite discontinuation although not necessarily at the same protocol-specified frequency. However, if the protocol treatment is stopped, then the biologic activity of the treatment ceases to exist after a certain period of time, and the protocol treatment effect may be confounded if progression is documented after change of protocol treatment. The question to be answered is how observation time will be counted for a given patient in the time-to-event analysis that has no documented progression due to protocol treatment discontinuation. Several sensitivity analyses can be conducted, for example, right censor the observation at the time of last tumor assessment with no documented progression, count observation as an event at the time of treatment discontinuation, count observation as event whenever tumor progression is documented after treatment discontinuation, or use the interval censoring approach for the time between last complete tumor evaluation with no documented progression and document progression after treatment discontinuation. Although none of these analyses are optimal, consistency among different sensitivity analyses may provide support for the treatment effect.

Incomplete data due to treatment discontinuation in oncology clinical trials are unavoidable because of the expected and unexpected toxicities of cancer therapy. If missing data are not substantial and the missing data pattern is similar in both the experimental and control treatment arms, the influence of informative censoring may not affect the inference on the relative treatment effect as measured by an HR. The lapatinib example described in more detail below illustrates this phenomenon of informative censoring (Figs. 1 and 2; ref. 13). Although the investigators or the local site readers had assessed that the disease had progressed, either the IRC reviewers did not agree that the disease had progressed or they had insufficient information to make the determination.

**Disagreement between Investigator and IRC**

The necessity of a radiologist's interpretation introduces an element of subjectivity to the process and opens the door for differences in time-of-progression calls. These differences may be due to selection of divergent target lesions, failure to identify new lesions (or misclassification of a new abnormality as a new tumor), and variability in target lesion measurement. Discrepancy rates between 2 radiologists have been found to be around 30% (4,14) and may be even higher for cancers that are more difficult to quantify. Few data incorporating the time dimension have been provided in the literature. Dodd and colleagues (4) report discrepancies between 2 readers from a randomized phase II study of bevacizumab in metastatic renal cancer.

In the assessment of radiologic progression in open-label clinical trials, when the investigator or site radiologist and the IRC evaluate disease progression, discrepancy is common because change of treatment or treatment discontinuation could occur at the time of investigator-determined disease progression, and patients typically are not followed further with radiologic scans at the same frequency. Hence, when the IRC does not concur with the investigator-
determined progression and no further radiologic scans are available, missing data occur for IRC evaluation, which is unique to oncology. Missing data could occur, if at the time of protocol-specified interim analysis, not all the radiologic scans have been reviewed by the IRC due to scheduled IRC review of radiologic scans in batches. Additional discrepancies could also occur due to differences in choice of target lesions by the investigator and independent reviewers. For example, as presented in the lapatinib product label (13) for the treatment of HER2-positive metastatic breast cancer, the HR favoring lapatinib was 0.57 as assessed by IRC with a median difference in time-to-progression of 8.5 weeks based on a total of 184 progression events versus an HR of 0.72 as assessed by the investigator with a median difference in time-to-progression of 5.6 weeks based on a total of 247 progression events. At the time of analysis, the observed difference in the information regarding number of progression events and hence the difference in estimated HRs was largely due to missing data in the IRC evaluation. Important distinctions between a local read and IRC that may influence variability are described in Table 1.

Recently, it has been shown that despite discrepancy in the progression evaluation between the investigator and IRC at the individual patient level, the relative treatment effect as measured by an HR using investigator-determined progression and IRC-determined progression is highly correlated (refs. 4,15,16; correlation coefficient, r = 0.954). The objective of an IRC is to audit an investigator’s evaluation to mitigate potential subjective bias that the investigator or site readers may introduce in the evaluation of progression. Whether a complete case review or a sample-based audit should be conducted was discussed in the July 2012 Oncologic Advisory Committee meeting (17). Currently, 2 sample-based audit methods (6,15) have been proposed in the literature, and these 2 methods were evaluated by the FDA and presented at this advisory meeting (17,18). The committee opined that given the strong correlation between the investigator- and IRC-assessed relative PFS treatment effect, the investigator-assessed PFS treatment effect may be used in evaluating a new treatment with confirmation of no systematic bias based on a sample-based IRC audit. This approach can potentially assure concerns about missing data for IRC-based PFS evaluation due to discrepancy between investigator and IRC assessments.

In a properly designed and conducted randomized clinical trial, randomization balances known and unknown effect modifiers, provides unbiased estimates, and permits generalization of results. However, when there are missing data, no single analysis can ensure an unbiased estimate of the treatment effect. Under these circumstances, conducting sensitivity analyses (19) using different censoring approaches in the time-to-event analysis are needed to determine how sensitive the results are to missing data assumptions. For example, in a sensitivity analysis, the discontinuation of change of treatment could be considered as a disease progression event. On the other hand, observations censored at the time of treatment discontinuation or change of treatment could lead to informative censoring, which might lead to biased estimates.

**Variability in Disease Evaluation Assessments**

Ideally, the precise timing at which a patient’s therapy fails to suppress tumor growth would be known. However, in practice, determination of the occurrence and timing of progression is inexact due to the shortcomings of radiographic imaging and lesion measurement. Many factors influence the ability to correctly characterize disease progression, including tumor size, margination and conspicuity, rate of tumor growth, frequency of measurement, and skill of the image interpreter. Lesions of sizes near the limit of radiographic resolution will be subject to greater reader variability, as will lesions with poorly defined margins and low conspicuity. With regard to timing of imaging, a greater frequency of imaging is advocated to more closely characterize the true progression time. Radiologists tend to agree more when evaluating an image with a dramatic increase in tumor burden, whereas disagreements are more common when the increase is relatively small. As greater growth might be observed from images that are spaced several months apart relative to images spaced a few weeks apart, scheduling can affect discrepancy rates. We do not advocate reductions in imaging frequency based on the above arguments, but the scheduling of imaging is important. An imaging schedule that coincides with knowledge of tumor growth is advised (faster-growing tumors require more frequent imaging), but this optimal schedule may be difficult to establish and image frequency is usually determined by feasibility and timing of treatment cycles.

Making a distinction between variability that occurs more often in a given treatment arm and variability that is not differential by treatment arm is important. Table 2 provides a brief summary of the 2 distinctions, discussed in detail below. If there is a reader evaluation bias, one would expect the pattern of variability to depend on the treatment arm. For example, if there is a tendency for progression to be called earlier in one treatment arm, it is easy to see how this would cause bias in treatment effect estimates. A trend for earlier progression times will make the PFS data for that...
### Table 1. Summary of important areas of distinctions between local and BICR progression determinations

<table>
<thead>
<tr>
<th>Area of distinction</th>
<th>Investigator or local evaluations</th>
<th>BICR or IRC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Difference</td>
<td>Advantage</td>
</tr>
<tr>
<td>Patient information</td>
<td>More complete information about patient status</td>
<td>More accurate reflection of a patient’s true status</td>
</tr>
<tr>
<td>Patient management</td>
<td>Reads will be used directly for patient management</td>
<td>May incentivize better reads</td>
</tr>
<tr>
<td>Reader skill/training</td>
<td>May or may not have extensive training in RECIST guidelines; differences in skill and training will vary by site</td>
<td>None noted, except that greater patient information may offset this</td>
</tr>
<tr>
<td>Blinding to treatment assignment</td>
<td>Larger cancer centers often have RECIST services, which facilitate blinded local reviews; extent of knowledge of treatment assignment unclear</td>
<td>None</td>
</tr>
<tr>
<td>Informative censoring</td>
<td>Not a concern</td>
<td>Potential for bias from informative censoring not a concern</td>
</tr>
<tr>
<td>Loss of events</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Abbreviation: NA, not applicable.
treatment arm seem worse than it is in truth, and its performance relative to the arm without this bias will look worse. We refer to this as "reader evaluation bias." Trials that are double-blind are protected from reader evaluation bias. Two important and extensive meta-analyses generally found that reader evaluation bias is not a concern, even in open-label studies (15,16).

The impact of variability that is non-differential (i.e., equally likely to occur in either treatment arm) is not as immediately apparent. Before proceeding, we must first note that this problem is different from the classic measurement error setup, in which measurement variability in an important explanatory variable (e.g., caloric intake in nutritional studies, which is measured imprecisely) attenuates estimates of its effect on an outcome variable (e.g., body mass index). With PFS, the measurement variability is fundamentally different, as it occurs in the outcome variable, which is the timing of disease progression. Far less is known about the impact of this. Computer simulation studies by Korn and colleagues (5) evaluated the potential impact on HR estimates. The simulation models presented by them varied the proportion of patients whose progression times were subject to error (a proportion of patients' progression times were assigned with certainty) as well as the amount of error. The studies showed that large (and likely unreasonable) variability was required to attenuate HR estimates. These authors conclude that a large impact on the HR was unlikely despite observed variability. Alternative simulation studies based on a tumor growth model were reported by Hong and colleagues (20). These studies showed some attenuation of the hazard ratio—the extent of which depended on the magnitude of the simulated variability. The greatest attenuation was a 17% reduction in the HR with an associated reduction in power from 88% to 70%.

**Summary**

Incomplete data are a reality in clinical trials and not unique to oncology clinical trials. There is no one process or statistical methodology to prevent or ensure unbiased interpretation of results when missing data occur, and it is difficult to interpret the results when there are substantial missing data. Regulators have published documents recognizing missing or incomplete data problems in clinical trials (21). Although missing data are common in clinical trials in all disease areas, there are unique considerations in oncology clinical trials. Imbalances of incomplete data across treatment arms are of particular concern, and placebo-controlled oncology studies may not be possible for ethical reasons. Placebo-controlled studies are only possible in oncology clinical trials that are designed to compare the standard of care to the standard of care plus a new therapy. It is also understood that patients are offered alternative treatment options when their disease progresses.

The use of PFS as the primary efficacy outcome for evaluating a new treatment poses unique challenges that are not present when overall survival is used. First, it is to be expected that a percentage of patients will discontinue treatment due to intolerable toxicity, as identification of baseline characteristics that predispose patients to toxicities is often impractical. In some circumstances, patients experiencing toxicities may continue to be followed as per the trial protocol to document progression. However, if such patients receive nonprotocol therapy after

**Table 2. Sources of variability and their impact on estimates of treatment effect**

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
<th>Potential Impact</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader-evaluation bias</td>
<td>Reader preference for a given therapy that influences timing of progression calls; for example, a tendency to call progressions earlier for patients on a control therapy in order to switch treatment earlier (relative to those on the experimental treatment)</td>
<td>If extreme, could bias estimates; in the case of a tendency to call progressions earlier in the control arm the experimental arm will appear better relative to the control arm</td>
<td>This would require consistently over- or under-calling progression in a given treatment arm, which is unlikely; furthermore, meta-analyses suggest this is not a concern</td>
</tr>
<tr>
<td>Nondifferential variability</td>
<td>A result of the lack of precision in progression determinations; this occurs when differences in progression times (as assessed by more than one reader) have the same pattern of variability across treatment arms</td>
<td>Simulation studies show small impact on estimates of treatment effect, which are typically attenuated</td>
<td>This is not generally a concern as supported by simulation studies and meta-analyses</td>
</tr>
</tbody>
</table>
discontinuation of protocol therapy, then the PFS effect cannot solely be attributed to the randomized treatment as the time-to-progression may be confounded by the effect of the nonprotocol therapy, although some have argued this is not a concern in a phase III trial (22). On the other hand, if such observations are censored at the time of discontinuation of protocol-specific therapy, the resulting estimate of treatment effect could be biased due to informative censoring. Because this situation is unavoidable, sensitivity analyses using different censoring approaches to assess the impact of missing data assumptions on the treatment effect (23), report the rate of dropouts due to toxicity, and compare missing rates across treatment groups are helpful.

Second, progression assessments can be missed due to schedule conflicts, prolonged toxicity, or ungradable scans. Although statistical methods can address this type of incomplete data, no one method can be endorsed. In general, if there is a large treatment effect almost all approaches lead to the same conclusion.

Third, although the same prespecified criteria (e.g., RECIST) may be used to assess disease progression by different radiologists (independent or site readers), subjectivity (24) is involved in the evaluation due to variations in tumor measurement, target lesion selection, identification of new lesions, and other factors. Thus, in registration trials, IRC or BICR plays an important role in providing confidence in the results. In using BICR assessments for the primary PFS analysis, missing data could occur due to no further follow-up on patients whose disease was assessed as progressed by the investigator. However, a complete-case BICR may not be necessary, and the use of a random sample–based BICR audit (6, 15, 20) of radiologic scans may be sufficient.

We have highlighted many of the major measurement concerns and have reviewed research indicating that the associated problems are not likely to be large when it comes to interpreting trial results with respect to relative treatment effect. More confidence should be given to treatments that have large effects on PFS for 2 reasons. First, such treatments are more likely to have an impact on a true patient benefit endpoint. Second, they are more likely to be robust to the concerns about measurement variability described here. Definitive trials should be powered to provide strong evidence that the treatment effect is large. This means that rather than designing a study to prove that the PFS HR is considerably better than just "any improvement," we need substantial evidence that the effect is of some clinically meaningful minimum threshold and sample sizes should be determined accordingly.

Although missing or incomplete data are inevitable in oncology clinical trials, every effort should be made to reduce and prevent missing data. Judicious use of sensitivity analyses can provide some reassurance of the observed results and the sensitivity of the results to missing data assumptions. Clinical trials should be prospectively planned with progression assessments occurring at the same frequency in both the experimental and control treatment arms. The protocols should also clearly prespecify sensitivity analyses and methods to address incomplete data. Random sample–based audits by an IRC or BICR may be useful to ensure lack of investigator bias in the determination of progression and mitigate incomplete data issues for IRC evaluation (18). However, further evaluation of such random sample–based audit strategies are needed in prospectively planned randomized clinical trials.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Conception and design: R. Sridhara, S.J. Mandrekar, L.E. Dodd
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): S.J. Mandrekar
Writing, review, and/or revision of the manuscript: R. Sridhara, S.J. Mandrekar, L.E. Dodd
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): R. Sridhara
Study supervision: R. Sridhara

Received December 21, 2012; revised March 20, 2013; accepted March 20, 2013; published online May 13, 2013.

References


Missing Data and Measurement Variability in Assessing Progression-Free Survival Endpoint in Randomized Clinical Trials

Rajeshwari Sridhara, Sumithra J. Mandrekar and Lori E. Dodd