Right-sizing Adjuvant and Neoadjuvant Clinical Trials in Breast Cancer

Donald A. Berry

Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, Texas

Summary

I describe a simple and intuitive method for finding sample size for clinical trials in primary breast cancer based on neoadjuvant trial results and using the FDA's patient-level meta-analysis. Then I explain a problem with this method and how the problem can be remedied.

In this issue of *Clinical Cancer Research*, Hatzis and colleagues (1) propose a method for designing clinical trials in early breast cancer with a survival end point based on an assumed benefit in pathologic complete response (pCR) rate from a neoadjuvant trial. In a similar vein in a recent article (2) I indicated how improvements in pCR rate from the FDA's meta-analysis of neoadjuvant trials in breast cancer (3) should be used to design such trials. I will demonstrate the latter method here by addressing the following statement in the authors' abstract: "For example, if the pCR rates are 30% and 60% (OR=3.5) and the 10-year RFS of the control arm is 0.74, the trial would require 3550 patients per arm, whereas if the RFS is 0.54 the trial would only require 425 patients per arm to detect significant survival benefit." In the process I will address considerations that pharmaceutical companies and cancer clinical trials consortia should consider when designing adjuvant and neoadjuvant trials that have survival end points.

I will neither emulate nor critique the Hatzis analyses but instead I will say how I design trials that have a survival endpoint such as event-free survival (EFS) based on evidence that the experimental therapy improves on the pCR rate of control. The method is intuitive and easy to develop and use. For example, in comparison with Hatzis and colleagues (1) there are no simulations or odds ratios.

Odds ratios comparing pCR rates are not relevant for inferring the effect of pCR on survival end points. Converting an additional 5% of patients to pCR implies an OR of 2.11 if the change is from 5% to 10% but only 1.23 if the change is from 40% to 45%. Moreover, to have a pCR OR of 2.11 in the latter setting would require an improvement in pCR from 40% to 58.5%. So the same OR in the latter cases affects 3.7 times as many women.

Corresponding Author: Donald A. Berry, University of Texas MD Anderson Cancer Center, Department of Biostatistics 1515 Holcombe Boulevard, Houston, TX 77030. dberry@mdanderson.org.

The relevant measure for predicting a survival benefit from pCR improvement is the proportion of patients who move from the no-pCR survival curve to the pCR curve, and that is simply the delta in pCR rates (2). Since the methodology is easy to apply to any biomarker subtypes using the FDA patient-level meta-analysis I will show the results for both HER2$^+$ and triple-negative breast cancer (TNBC).

Figure 1 shows the expected EFS HR depending on pCR improvement for these two biomarker subtypes. These curves assume exponential EFS distributions for patients experiencing and not experiencing pCR. (For TNBC this assumption does not fit the Kaplan-Meier curves very well but it has little impact on quantitative conclusions and no impact on my overall conclusions.) The blue dots on Figure 1 show the expected HRs 0.759 and 0.676, respectively, at the authors' assumed delta in pCR of 0.30. So, for example, if the true improvement in pCR rate for TNBC is 0.30 then the most likely value of EFS HR is 0.676.

Table 1 shows the resulting sample sizes for the two assumptions regarding RFS (EFS) considered in the abstract by Hatzis and colleagues (1) (where "med" means median assuming exponential time to event). Calculations in Table 1 are based on the standard method of George and Desu (4) assuming 90% power, 2-sided type I error rate 0.05, 2 years for accruing patients and 2 versus 5 years of additional follow-up (FU). I chose these quantities as characterizing modern trials more generally and I have not attempted to make assumptions analogous to those of Hatzis and colleagues (1). The fact that our answers differ is not surprising. What is surprising is that Hatzis and colleagues conclude a sample size increase when assuming a control 10-year RFS of 0.74 versus 0.54 of 3550/425=8.35 times whereas this ratio is approximately 2 for all scenarios considered in Table 1.

As pointed out by Berry and Hudis (2), there are two major problems when designing trials using either Table 1 or the method by Hatzis and colleagues (1). One is pervasive in clinical research and helps explain why Phase 3 trials fail much more often than their statistical power suggests. The problem is illustrated in Figure 1 by comparing the histogram of Phase 2 results with the power curve for the planned study. The observed delta for pCR was a statistically significant 0.30 but the true delta is uncertain. This uncertainty is usually communicated by a confidence interval but the histogram in the figure is more informative in the sense that it shows regions of higher versus lower likelihood. If the true delta is greater than 0.30 then the trial will be overpowered and if it is less than 0.30 the trial will be underpowered, with the latter decreasing much more dramatically than the former is increasing. The average power (called "predictive power") is the probability the trial will be positive given the prior data and its associated uncertainty. It is less than the nominal 90% and in the example of Figure 1 the predictive power is only 75%.

The other major problem with the approach is that the relationship between pCR and EFS may be different from that in the FDA meta-analysis (3) and the data from Hatzis and colleagues (1). Drugs' mechanisms of action differ, with some drugs substantially decreasing measurable tumor burden but they have little effect on slowing the disease process. The opposite scenario is also possible. Indeed, these same authors (5) have emphasized that pCR is a dichotomous partitioning of tumor burden at surgery which

reduces the information in assessing drug effect. Better to focus on degree of response. A therapy might improve the rate of "partial response," for example, which can be assessed using Residual Cancer Burden class 1 (5). And converting a patient to RCB1 may be associated with prolonging EFS quite apart from the positive consequences of experiencing a pCR.

Berry and Hudis (2) offer a solution to both problems: adaptive trial design, including re-estimating sample size based on the accruing trial results for pCR rates and also on the observed relationship between pCR and EFS in the context of the investigational therapy in the trial. For example, Figure 1 shows that a sample size with 90% power when delta is 0.30 has only about 50% power when delta is 0.20. If interim data are suggesting that the true delta is only about 0.20 then the sample size can be increased, depending on what the available evidence about the relationships between pCR and EFS. On the other hand if the true delta seems closer to 0.40 then the sample size can be decreased. (Both of these possibilities affect the type I error rate which can be easily adjusted and controlled.) And if the true delta seems to be less than 0.10 then the trial can be stopped for futility rather than increasing the sample size to an impossibly large number.

In summary, I propose using the FDA meta-analysis (3) for designing neoadjuvant breast cancer clinical trials that have both pCR and EFS endpoints. It applies as well for adjuvant trials based on information about pCR from phase 2 neoadjuvant trials. My approach allows for calculating statistical power as usual by assuming particular values of pCR rates for the experimental and control arms. It also enables finding the probability that the trial will be successful allowing for the variability in treatment-specific pCR rates from previous neoadjuvant clinical trials. This approach—and every standard approach for determining sample size—is fraught with such uncertainty as to discourage investigators from taking a neoadjuvant approach. The uncertainty due to having to make questionable assumptions about treatment-specific pCR rates and treatment-specific relationships between pCR and EFS can be mitigated using an adaptive design in which the sample size is re-estimated periodically based on data accruing in the trial.

Acknowledgments

References

1. Hatzis C, Symmans WF, Zhang Y, Gould RE, Moulder SL, Hunt KK, et al. Relationship between complete pathologic response to neoadjuvant chemotherapy and survival in triple negative breast cancer. Clin Cancer Res. 2015 Aug 15. [Epub ahead of print].

2. Berry DA, Hudis CA. Neoadjuvant therapy in breast cancer as a basis for drug approval. JAMA Oncology. 2015; 1:875–6. [PubMed: 26181139]

3. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet. 2014; 384:164–72. [PubMed: 24529560]

4. George SL, Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. J Chron Dis. 1974; 27:15–24. [PubMed: 4592596]

5. Symmans WF, Peintinger F, Hatzis C, Rajan R, Kuerer H, Valero V, et al. J Clin Oncol. 2007; 25:4414–22. [PubMed: 17785706]
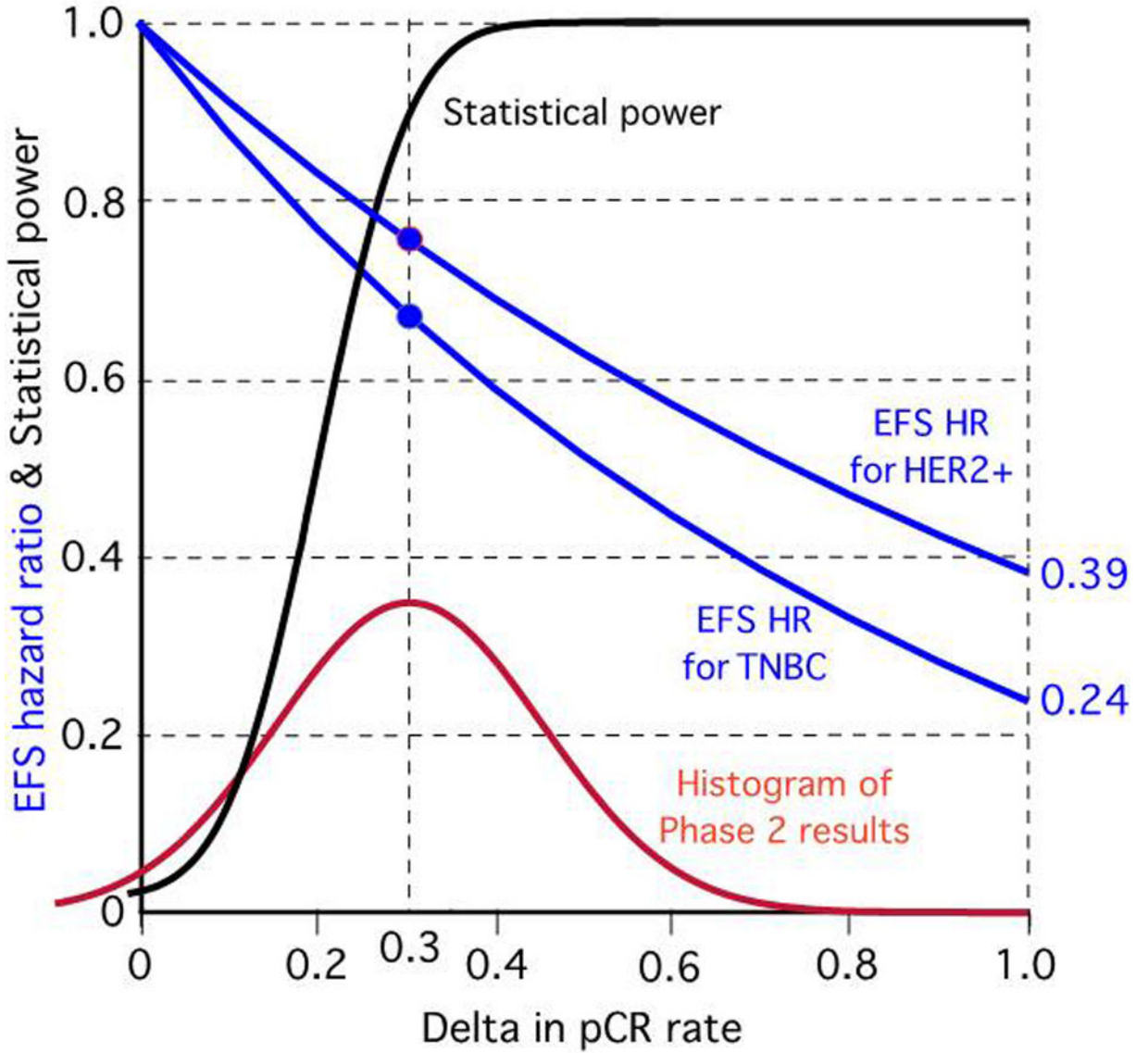
Figure 1.
Estimated relationship between EFS hazard ratio (HR) and improvement in pCR rate for
HER2[+] and TNBC based on FDA meta-analysis shown in blue (3). The dots at delta equals
0.30 occur at EFS hazard ratio 0.759 and 0.676, respectively. The respective values 0.39 and
0.24 at delta equals 1.0 are from the FDA meta-analysis and are the HRs for comparing EFS
for patients with pCRs with EFS for patients with other than pCRs. The red histogram shows
the uncertainty in the true delta based on a Phase 2 clinical trial with observed delta 0.30 that
is statistically significant with 2-sided p-value 0.05. The black curve shows power as it
depends on the true delta when the Phase 3 trial has 90% power to detect a difference of
90%. The reason this curve appears on this figure is described in the text. Adapted from
Berry and Hudis (2). Copyright © 2015 American Medical Association. All rights reserved.

Table 1

EFS Trial Sample Size Derived from FDA Meta-analysis

Control 10-yr RFS (med)	Subtype	Annual hazards		N per arm	
		Control	Exper	2 yrs FU	5 yrs FU
0.74 (23.0 yrs)	HER2[+]	0.0301	0.0229	3710	1930
	TNBC	0.0301	0.0205	2010	1040
0.54 (11.3 yrs)	HER2[+]	0.0616	0.0468	1900	1030
	TNBC	0.0616	0.0416	996	536