ISSN: 2469-5831

Cheng et al. Int J Clin Biostat Biom 2017, 3:010
DOI: 10.23937/2469-5831/1510010
Volume 3 | Issue 1
Open Access

International Journal of
Clinical Biostatistics and Biometrics

ORIGINAL ARTICLE

Inference Based on Small Randomized Oncology Clinical Trials: Is the Observed Treatment Effect True?

Joyce Cheng*, Hui Zhang, Shenghui Tang, and Rajeshwari Sridhara

FDA, Division of Biometrics 5, Center for Drug Evaluation and Research, Office of Translational Sciences, Office of Biostatistics, USA

***Corresponding author:** *Joyce Cheng, FDA, Division of Biometrics 5, Center for Drug Evaluation and Research, Office of Translational Sciences, Office of Biostatistics, FDA/CDER, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA, E-mail: Joyce.Cheng@fda.hhs.gov*

Abstract

The drug development paradigm in oncology has changed in recent times as developments in science and technology have led to more targeted therapies. Drug products are receiving marketing approvals based on single randomized studies enrolling 100-200 patients, including early phase (phase II) clinical trials. In this paper, we examine the likelihood of observing a significant treatment effect when in fact the true treatment effect is modest to null by exploring a range of sample sizes via simulation studies. Results showed that the Cox model performed appropriately in studies as small as n = 50 and extreme treatment effect estimates were very rarely observed when the true treatment effect was modest to null, at least for the examples considered under our assumed conditions. It appears that a hazard ratio of large magnitude observed in a small study is likely to be indicative of a true treatment effect, although of uncertain magnitude. However, the simulation assumes a well-conducted study with minimal to no amendments or adaptations and a non-ambiguous endpoint, which unfortunately is not often the case in the early phases of drug development. As the paradigm in oncology changes, randomized phase II studies can no longer be seen as simply supporting go/no-go decisions. When promising drugs are evaluated in trials with overall survival as an endpoint, companies may want to consider providing a pre-specified contingency statistical analysis plan in anticipation of unexpectedly promising survival results.

Keywords

Clinical trials, Small sample, Log-rank test, Cox model, Hazard ratio

Introduction

Oncology drug development has seen a paradigm shift in recent years as innovations in science and technology have led to the development of new and effective targeted therapies. The perceived effectiveness of these new therapies has led to questions regarding the ethical appropriateness of traditional drug development when early phase studies show large treatment effects with limited toxicity [1-4]. The increased desire for earlier access to new therapies, along with the subsequent passage of the FDA Safety and Innovation Act of 2012, has allowed for the FDA's conscious efforts to implement expedited programs such as Fast Track, Breakthrough Therapy, Accelerated Approval, and Priority Review for serious conditions [3-5]. As a direct result of this, there have been an increasing number of oncology drug approvals based on single randomized studies enrolling small populations of 100-200 patients, often in early phase clinical trials. Early phase studies by nature are often plagued by conduct issues, so expediting effective therapies to market while maintaining statistical rigor and regulatory scrutiny is an important issue the FDA is facing that must be explored further [6].

Early phase trials are not designed for registration purposes, so their design and conduct open up great potential for uncertainty. First of all, these trials are often conducted in small populations. Redman and Crowley warned that small randomized studies often result in unstable estimates of efficacy and may not be large

Citation: Cheng J, Zhang H, Tang S, Sridhara R (2017) Inference Based on Small Randomized Oncology Clinical Trials: Is the Observed Treatment Effect True. Int J Clin Biostat Biom 3:010. doi.org/10.23937/2469-5831/1510010

Received: March 01, 2017: **Accepted:** May 18, 2017: **Published:** May 20, 2017

Copyright: © 2017 Cheng J, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

enough to balance potential prognostic and predictive factors between arms [7]. Tuma emphasized the issue of heterogeneity in phase II populations [8].

Secondly, there are concerns related to study conduct in early phase development as it is generally set with more liberal operating characteristics. Lara and Redman cited concerns by Redman and Crowley and Tuma [7-9] regarding the limitations of randomized phase II designs and further pointed out that positive phase II studies do not necessarily result in positive phase III studies noting "the predictive value of a phase II study is a function of the quality and design of the study and not whether it involved a randomized comparison or not." It should be noted that a dampened effect in phase III following a promising phase II result is neither uncommon nor unexpected. The FDA (2017) published a report [10] of 22 case studies where phase III results were divergent from their phase II counterparts in safety, efficacy, or both. The breast cancer drug iniparib was one of the cases described in the report and represents an extreme case in which a promising phase II trial led to a failed phase III trial. Proposed explanations for iniparib's failure in phase III included poor study design in phase II in which crossover may have given it an inadvertent advantage, the possibility of a false positive statistical result and the heterogeneity of triple-negative breast cancer, and the fact that iniparib likely did not work as promoted [11]. These are all examples of conduct and trial design issues common to early phase studies. Other examples of such issues include unplanned interim looks as well as data-driven changes and modifications.

Lastly, the endpoints assessed in early phase are often more ambiguous. For example, progression-free survival is susceptible to ambiguity with respect to the frequency of assessment and progression determination.

Assuming a well-conducted study with an unambiguous endpoint such as overall survival, the major cause for concern that remains with approving indications based on randomized phase II studies is their small sample size. In fact, the FDA report [10] on divergent results between phase II and phase III was intended to show how "controlled trials of appropriate size and duration contribute to the scientific understanding of medical products." In general, small sample sizes are not problematic since studies designed to detect a large treatment effect are expected to be small. The performance of Cox models, which are typically used to estimate the hazard ratio, also should not be problematic in small sample settings. Johnson, et al. [12] has assessed the small sample performance of Cox models in estimating regression parameters in a two-covariate hazard function model, and they cited an earlier report that considered the simpler single variable case. The general conclusion was that in ideal conditions of balanced covariates and no censoring, results were reasonable for samples of size 40 or greater in terms of bias, asymptotic versus finite-sample variance, and power [12]. But, bias increases when the treatment and control groups are unbalanced [12].

However, the situation emerging in oncology drug development is that more and more phase II trials designed for go/no-go decisions are observing treatment effects much larger than what they were initially designed to detect. In these cases, the trials have small sample sizes because they were not meant for registration and were thus designed to allow for high type-1 error. Since the trials were designed to detect a smaller treatment effect than what was observed, and heeding Redman and Crowley's [7] warning, there is concern that the promising result could have occurred by chance alone.

Simulation studies were conducted, under the assumption of a well-conducted study with an overall survival endpoint, to explore the likelihood for results to be inflated in a small sample study and to what extent they can be believed. Results from these simulations will help provide general insight as to how these situations maybe addressed in the future.

Simulation Method

Consider an example study scenario where the accrual period was 12 months with an 18 month follow-up period. Patient start times were generated from a uniform (0,12) distribution, and the failure time for each patient was generated from a distribution with hazard function.

$$\lambda(t) = \lambda_0(t)e^{\beta z}$$

Where Z = 1 for the experimental treatment arm and Z = 0 for the control arm, with a constant baseline hazard function over time, $\lambda(t) = \lambda_0$ for all time t.

A series of trials were simulated with various true hazard ratios representing effect sizes from moderate to null (i.e. HR = 0.7, 0.8, 0.9, and 1), sample sizes, and control medians for overall survival, as summarized in Table 1. These settings were chosen to explore the effect of sample size on hazard ratio estimates, when the true hazard ratio shows moderate to no treatment effect, and whether that effect is further affected by the length of median survival (i.e. the percentage of censored observations).

Table 1: Simulation scenarios for each assumed true hazard ratio.

Sample size	Control median		
800	6	12	24
600	6	12	24
400	6	12	24
200	6	12	24
100	6	12	24
50	6	12	24

Table 2: Simulation results when true hazard ratio = 1.

Control median	Sample size	Percent censored	Bias	ASE	ESD	Proportion of HR		
						< 0.5	< 0.4	< 0.3
6	800	6.8	0.0009	0.0733	0.0733	0.0000	0.0000	0.0000
	600	6.7	0.0011	0.0847	0.0864	0.0000	0.0000	0.0000
	400	6.8	-0.0021	0.1038	0.1067	0.0000	0.0000	0.0000
	200	6.8	0.0009	0.1473	0.1484	0.0000	0.0000	0.0000
	100	6.8	-0.0007	0.2092	0.2124	0.0004	0.0000	0.0000
	50	6.9	0.0031	0.2991	0.3049	0.0112	0.0018	0.0000
12	800	25.5	0.0018	0.0820	0.0835	0.0000	0.0000	0.0000
	600	25.6	-0.0015	0.0948	0.0949	0.0000	0.0000	0.0000
	400	25.5	0.0018	0.1160	0.1165	0.0000	0.0000	0.0000
	200	25.6	-0.0053	0.1646	0.1661	0.0002	0.0000	0.0000
	100	25.7	0.0028	0.2338	0.2360	0.0022	0.0002	0.0000
	50	26.0	-0.0099	0.3341	0.3357	0.0198	0.0036	0.0002
24	800	50.3	0.0013	0.1005	0.1017	0.0000	0.0000	0.0000
	600	50.3	0.0012	0.1160	0.1168	0.0000	0.0000	0.0000
	400	50.3	0.0031	0.1422	0.1421	0.0000	0.0000	0.0000
	200	50.4	-0.0015	0.2019	0.2006	0.0008	0.0000	0.0000
	100	50.4	-0.0035	0.2875	0.2902	0.0112	0.0020	0.0000
	50	50.8	-0.0062	0.4136	0.4207	0.0520	0.0162	0.0030

Median HR rounded to two decimal places was 1.00 in all cases except for two cases when it was 0.99 (Control median = 12, Sample size = 200 and Control median = 12, Sample size = 50).

Table 3: Simulation results when true hazard ratio = 0.9.

Control median	Sample size	Percent censored	Bias	ASE	ESD	Proportion of HR		
						< 0.5	< 0.4	< 0.3
6	800	7.8	0.0019	0.0738	0.0750	0.0000	0.0000	0.0000
	600	7.8	-0.0012	0.0852	0.0851	0.0000	0.0000	0.0000
	400	7.8	-0.0029	0.1045	0.1057	0.0000	0.0000	0.0000
	200	7.8	0.0017	0.1481	0.1506	0.0000	0.0000	0.0000
	100	7.8	-0.0014	0.2105	0.2124	0.0040	0.0000	0.0000
	50	8.1	-0.0113	0.3008	0.2987	0.0290	0.0050	0.0008
12	800	27.4	-0.0002	0.0831	0.0818	0.0000	0.0000	0.0000
	600	27.4	-0.0023	0.0960	0.0967	0.0000	0.0000	0.0000
	400	27.4	0.0007	0.1176	0.1181	0.0000	0.0000	0.0000
	200	27.5	-0.0060	0.1668	0.1638	0.0002	0.0000	0.0000
	100	27.5	0.0013	0.2369	0.2431	0.0096	0.0006	0.0000
	50	27.7	0.0062	0.3381	0.3437	0.0432	0.0102	0.0006
24	800	52.1	-0.0007	0.1024	0.1029	0.0000	0.0000	0.0000
	600	52.0	0.0010	0.1182	0.1165	0.0000	0.0000	0.0000
	400	52.2	0.0006	0.1451	0.1456	0.0000	0.0000	0.0000
	200	52.1	0.0022	0.2056	0.2063	0.0018	0.0002	0.0000
	100	52.1	-0.0028	0.2927	0.2927	0.0238	0.0028	0.0000
	50	52.3	-0.0103	0.4210	0.4344	0.0890	0.0346	0.0080

Median HR rounded to two decimal places was 0.90 in all cases except for one case when it was 0.91 (Control median = 24, Sample size = 200) and one case when it was 0.89 (Control median = 24, Sample size = 50).

Hazard ratios were estimated using the Cox model which is usually specified in oncology trial protocols. Treatment magnitude was quantified under the following assumptions: HR = 1 represents no effect, HR = 0.9, 0.8, and 0.7 represent moderate effects with HR = 0.7 serving as the threshold between moderately effective and 'clinically' effective and any HR < 0.5 is considered an extreme effect.

Each data configuration shown in Table 1 was replicated 5,000 times, and the following values were recorded (1) The average percentage censored; (2) The empirical bias of β, calculated as the mean of the β estimates from the Cox model minus the true value; (3) The Empirical Standard Deviation (ESD) of β, calculated as the standard deviation of the β estimates from the Cox model; (4) The average Asymptotic Standard Error (ASE) of β, calculated as the mean of the standard errors of β given by the Cox model; (5) Hazard ratio estimates from the Cox model; and (6) The proportion of HR estimates from the Cox model that were less than 0.3, 0.4, and 0.5.

Results

Simulation results are summarized in Table 2, Table 3, Table 4, and Table 5 by true hazard ratio, and Figure 1 shows density plots for the hazard ratio estimates by

Table 4: Simulation results when true hazard ratio = 0.8.

Control median	Sample size	Percent censored	Bias	ASE	ESD	Proportion of HR		
						< 0.5	< 0.4	< 0.3
6	800	9.1	0.0000	0.0745	0.0742	0.0000	0.0000	0.0000
	600	9.1	-0.0010	0.0860	0.0856	0.0000	0.0000	0.0000
	400	9.1	-0.0014	0.1055	0.1070	0.0000	0.0000	0.0000
	200	9.1	0.0006	0.1495	0.1508	0.0010	0.0000	0.0000
	100	9.2	-0.0012	0.2124	0.2104	0.0152	0.0014	0.0000
	50	9.3	-0.0035	0.3037	0.3175	0.0740	0.0156	0.0018
12	800	29.5	-0.0014	0.0845	0.0850	0.0000	0.0000	0.0000
	600	29.5	-0.0029	0.0976	0.0986	0.0000	0.0000	0.0000
	400	29.5	0.0000	0.1197	0.1176	0.0000	0.0000	0.0000
	200	29.5	0.0042	0.1695	0.1693	0.0036	0.0000	0.0000
	100	29.6	-0.0008	0.2409	0.2389	0.0280	0.0028	0.0002
	50	29.7	-0.0042	0.3436	0.3462	0.0878	0.0270	0.0048
24	800	53.9	-0.0004	0.1047	0.1045	0.0000	0.0000	0.0000
	600	53.9	0.0007	0.1210	0.1207	0.0000	0.0000	0.0000
	400	54.0	-0.0035	0.1485	0.1500	0.0004	0.0000	0.0000
	200	54.1	-0.0044	0.2109	0.2123	0.0164	0.0006	0.0000
	100	54.1	0.0031	0.3002	0.3025	0.0576	0.0128	0.0006
	50	54.2	-0.0037	0.4311	0.4364	0.1382	0.0592	0.0152

Median HR rounded to two decimal places was 0.80 in all cases except for one case when it was 0.81 (Control median = 24, Sample size = 50).

Table 5: Simulation results when true hazard ratio = 0.7.

Control median	Sample size	Percent censored	Bias	ASE	ESD	Proportion of HR		
						< 0.5	< 0.4	< 0.3
6	800	10.9	0.0007	0.0755	0.0746	0.0000	0.0000	0.0000
	600	10.9	-0.0010	0.0872	0.0880	0.0002	0.0000	0.0000
	400	10.9	-0.0006	0.1069	0.1050	0.0006	0.0000	0.0000
	200	10.9	0.0008	0.1516	0.1515	0.0152	0.0000	0.0000
	100	11.0	-0.0057	0.2155	0.2189	0.0656	0.0054	0.0002
	50	11.0	-0.0143	0.3078	0.3170	0.1514	0.0422	0.0062
12	800	31.9	0.0004	0.0863	0.0880	0.0000	0.0000	0.0000
	600	31.9	-0.0003	0.0997	0.1004	0.0002	0.0000	0.0000
	400	32.0	0.0015	0.1222	0.1200	0.0022	0.0000	0.0000
	200	31.9	-0.0004	0.1732	0.1722	0.0244	0.0012	0.0000
	100	32.0	-0.0053	0.2461	0.2436	0.0874	0.0136	0.0006
	50	32.2	-0.0100	0.3518	0.3592	0.1784	0.0644	0.0088
24	800	56.0	-0.0018	0.1078	0.1068	0.0018	0.0000	0.0000
	600	56.0	-0.0035	0.1245	0.1249	0.0042	0.0000	0.0000
	400	56.1	-0.0023	0.1528	0.1539	0.0166	0.0002	0.0000
	200	56.0	-0.0016	0.2166	0.2151	0.0622	0.0058	0.0000
	100	56.2	-0.0105	0.3095	0.3084	0.1436	0.0408	0.0056
	50	56.4	-0.0138	0.4466	0.4621	0.2320	0.1110	0.0402

Median HR rounded to two decimal places was 0.70 in all cases except for two cases when it was 0.69 (Control median = 6, Sample size = 100 and Control median = 24, Sample size = 0).

true hazard ratio and control median for select sample sizes. The simulation results in Table 2 assume no treatment effect (HR = 1). When the sample size was large (n ≥ 200), the results showed no instances in which an extreme treatment effect (HR < 0.5, 0.4, or 0.3) was detected. Even at smaller sample sizes (n < 200), extreme treatment effects occurred very rarely. The proportion of hazard ratios less than 0.5 under the most extreme scenario, where control median was 24 months and sample size was 50, was less than 6%. Simulation results in Table 3, Table 4, and Table 5 assume modest treatment effects of HR = 0.9, 0.8, and 0.7, respectively. Here incidences of extreme treatment effects (HR < 0.5, 0.4, or 0.3) increased as sample size decreased, but overall there was still a fairly low number of such occurrences with the maximum proportion being < 25% in the most extreme scenario (true HR = 0.7, control median = 4 months and n = 50). Overall, the highest number of extreme effects occurred when the true hazard ratio was 0.7 in which case a hazard ratio of 0.5 would likely not be considered extreme. Even then, the proportion of hazard ratios less than 0.3 remained small (< 5%).

Examining the operating characteristics of the Cox model in Table 2, Table 3, Table 4, and Table 5, it appears the conclusions of Johnson, et al. [12] were upheld

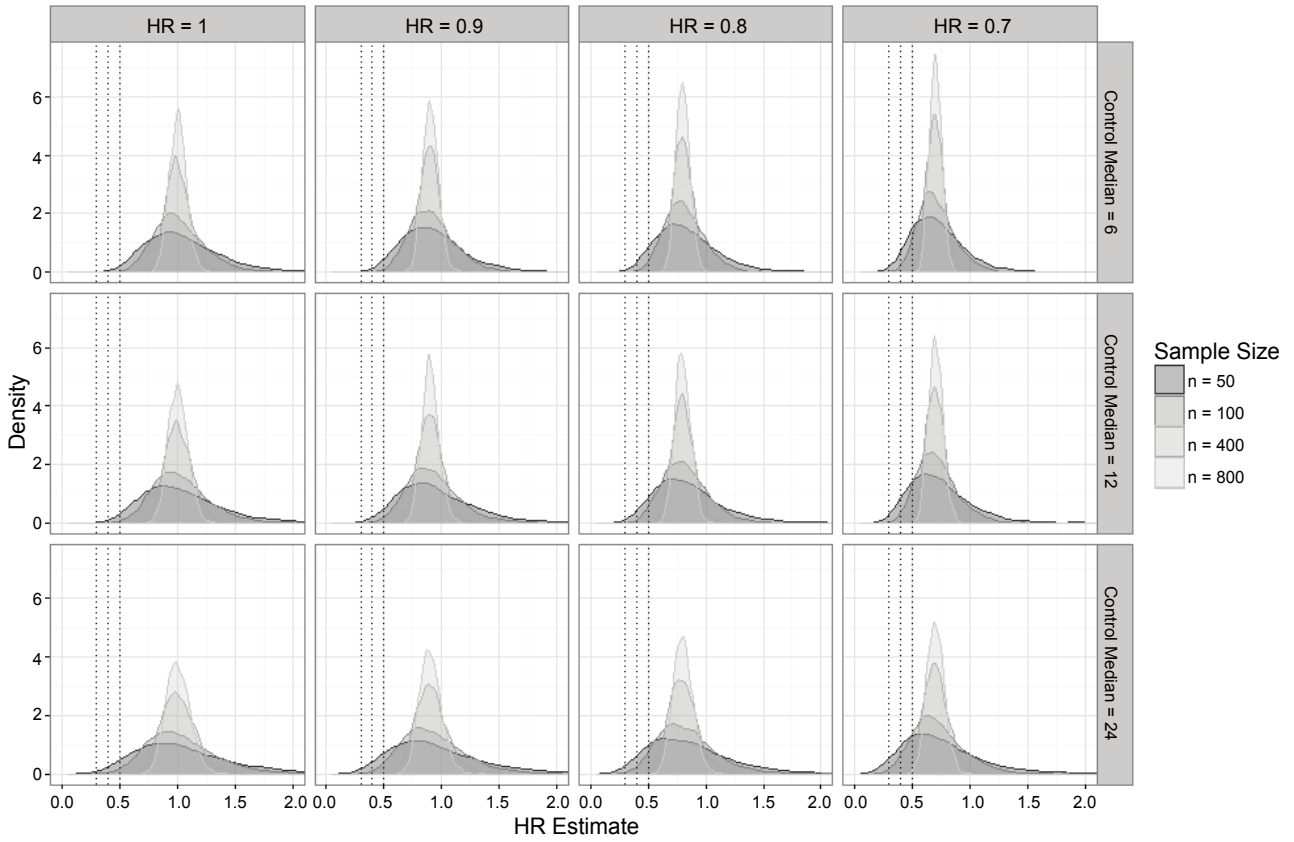

Figure 1: Density plots of hazard ratio estimates as sample size changes across all control median and true hazard ratio specifications.

under the settings for this simulation. While a formal test for non-proportionality was not conducted, it is assumed that the proportional hazards assumption holds, as the simulated data were generated from identical exponential distributions. Biases remained close to zero with no noticeable pattern as sample size decreased, even as control median and true hazard ratio vary. The ASE and ESD were very similar across all settings, and note that ASE and ESD both appeared to increase as the sample size decreased or the control median increased. The median hazard ratio estimate was consistent with the true hazard ratio across all simulation scenarios.

The density plots in Figure 1 show that the spread of estimates increased as sample sizes decreased, so it does appear that small studies are prone to overestimate the effect size, which has been noted in the literature [13]. It should be noted that the effect size can be underestimated as well, but this is likely not a concern in terms of regulatory decision-making. Recall that incidence of extreme treatment effects was generally low, as seen in Table 2, Table 3, Table 4, and Table 5, with the highest number of observed extreme effects understandably occurring when the true hazard ratio was 0.7. These two observations taken together show that while an observed extreme treatment effect is most likely inflated, there is evidence the underlying true hazard ratio is likely to be 0.7 or better, which would represent a slightly diluted but still meaningful effect.

The impact of an unequal allocation ratio of 2:1 (treatment to control), a commonly used unequal allocation ratio in oncology clinical trials, was explored on select data configurations. Table 6 shows some of these results, which were fairly consistent with what was seen with equal 1:1 allocation. Biases remained small, ASE and ESD values were similar, and the median hazard ratio estimates were consistent with the true hazard ratios. The increased bias with unequal allocation ratio (1:4) as pointed out in Johnson, et al. [11] was not observed under the settings considered (2:1) for this simulation.

Olaratumab Case Study

Olaratumab was approved in October 2016 on the basis of an early phase randomized trial and can be used to illustrate the various challenges, regulatory and otherwise, involved with observing unexpected results in randomized phase II trials originally designed for go/no-go decision-making [14]. The olaratumab trial was a phase 1b/2 trial in 133 patients with advanced soft-tissue sarcoma with investigator-assessed Progression-Free Survival (PFS) as the primary endpoint and Overall Survival (OS) as one of the secondary endpoints. The trial was designed to detect a difference for PFS at a two-sided significance level of 0.2 with a power of 0.8, and multiplicity adjustment was not planned for all secondary endpoints including OS. Results of the trial showed a moderate 2.5 month improvement in

Table 6: Select simulation results with unequal allocation.

True HR	SS	Ctrl med	Alloc. ratio	Percent censored	Bias	ASE	ESD	Med HR	Proportion of HR		
									< 0.5	< 0.4	< 0.3
1	100	12	1:1	25.7	0.0028	0.2338	0.2360	1.00	0.0022	0.0002	0.0000
			2:1	25.7	0.0007	0.2494	0.2495	1.00	0.0020	0.0000	0.0000
		24	1:1	50.4	-0.0035	0.2875	0.2902	1.00	0.0112	0.0020	0.0000
			2:1	49.4	0.0056	0.3070	0.3092	1.00	0.0116	0.0008	0.0000
	50	12	1:1	26.0	-0.0099	0.3341	0.3357	0.99	0.0198	0.0036	0.0002
			2:1	25.7	0.0027	0.3506	0.3582	1.00	0.0246	0.0056	0.0008
		24	1:1	50.8	-0.0062	0.4136	0.4207	1.00	0.0520	0.0162	0.0030
			2:1	50.6	0.0218	0.4354	0.4420	1.01	0.0482	0.0146	0.0018
0.7	100	12	1:1	32.0	-0.0053	0.2461	0.2436	0.70	0.0874	0.0136	0.0006
			2:1	34.1	-0.0052	0.2583	0.2627	0.70	0.0976	0.0174	0.0012
		24	1:1	56.2	-0.0105	0.3095	0.3084	0.69	0.1436	0.0408	0.0056
			2:1	58.1	0.0042	0.3230	0.3220	0.70	0.1458	0.0366	0.0036
	50	12	1:1	32.2	-0.0100	0.3518	0.3592	0.70	0.1784	0.0644	0.0088
			2:1	34.5	-0.0109	0.3638	0.3613	0.69	0.1834	0.0622	0.0108
		24	1:1	56.4	-0.0138	0.4466	0.4621	0.70	0.2320	0.1110	0.0402
			2:1	58.0	0.0037	0.4569	0.4640	0.70	0.2268	0.1084	0.0320

For allocation ratio 2:1, sample sizes considered were n=99 and n=51 respectively.

Table 7: Olaratumab simulation results.

True HR	Treat events	Control events	Bias	ASE	ESD	Median HR	Proportion of HR		
							< 0.5	< 0.4	< 0.3
1	57	57	-0.0012	0.1876	0.1876	1.00	0.0000	0.0000	0.0000
0.9	56	57	0.0029	0.1892	0.1919	0.90	0.0018	0.0004	0.0000
0.8	53	57	-0.0003	0.1915	0.1910	0.80	0.0070	0.0000	0.0000
0.7	51	57	-0.0023	0.1946	0.1953	0.70	0.0424	0.0048	0.0000

estimated median PFS with a stratified HR of 0.67 that was statistically significant at the two-sided 0.2 level. However, an unexpectedly large improvement of 11.8 months in estimated median OS with an unstratified HR of 0.52 was observed. While the PFS improvement on its own would likely be unremarkable in such a trial, olaratumab was granted accelerated approval largely on the basis of the OS benefit seen, despite the limitations in the trial design with respect to sample size and the set operating characteristics.

To further assess the unexpected survival benefit seen in this trial, we performed the following simulation. Data settings were chosen to mimic the olaratumab trial. The sample size was set at n = 130 patients, the accrual period was 12 months with approximately 40 months of follow up, and median survival was set as 14.7 months on the control arm. Under these specifications, data were generated to explore scenarios where the true treatment effect was modest to null as quantified by a series of hazard ratios from 0.7 to 1. Each scenario was replicated 5,000 times and the same values as in the previous simulation were recorded.

Results are shown in Table 7. The average number of events on the control arm in each case (57 events) is comparable to the actual number of events observed in the control arm of the olaratumab study (52 events), indicating that the simulation settings adequately mimicked those of the original study. As with the previous simulation, operating characteristics of the Cox model showed its small sample performance to be sound. Biases are small, ASE and ESD values are similar, and the median hazard ratio estimates align with their true values.

Figure 2 shows density plots of the hazard ratio estimates under the different true hazard ratios considered. The spread of the estimates is similar across settings as the sample size in the simulation remained constant at n = 130. Although the spread is quite wide due to the small sample size, it still barely covers any extreme hazard ratios of magnitude less than 0.5 until the truth is around HR = 0.7. Thus these results are supportive of the claim that it is unlikely the extreme result seen in the olaratumab trial was completely due to chance.

Conclusion

Under the new oncology drug development paradigm, it is clear that randomized phase II trials can no longer be seen as simply supporting go/no-go decisions. It is becoming more and more common for products to receive accelerated approval based on the results of small randomized studies not initially designed for registration when a large magnitude of benefit is observed. Thus, it is important to be confident that these

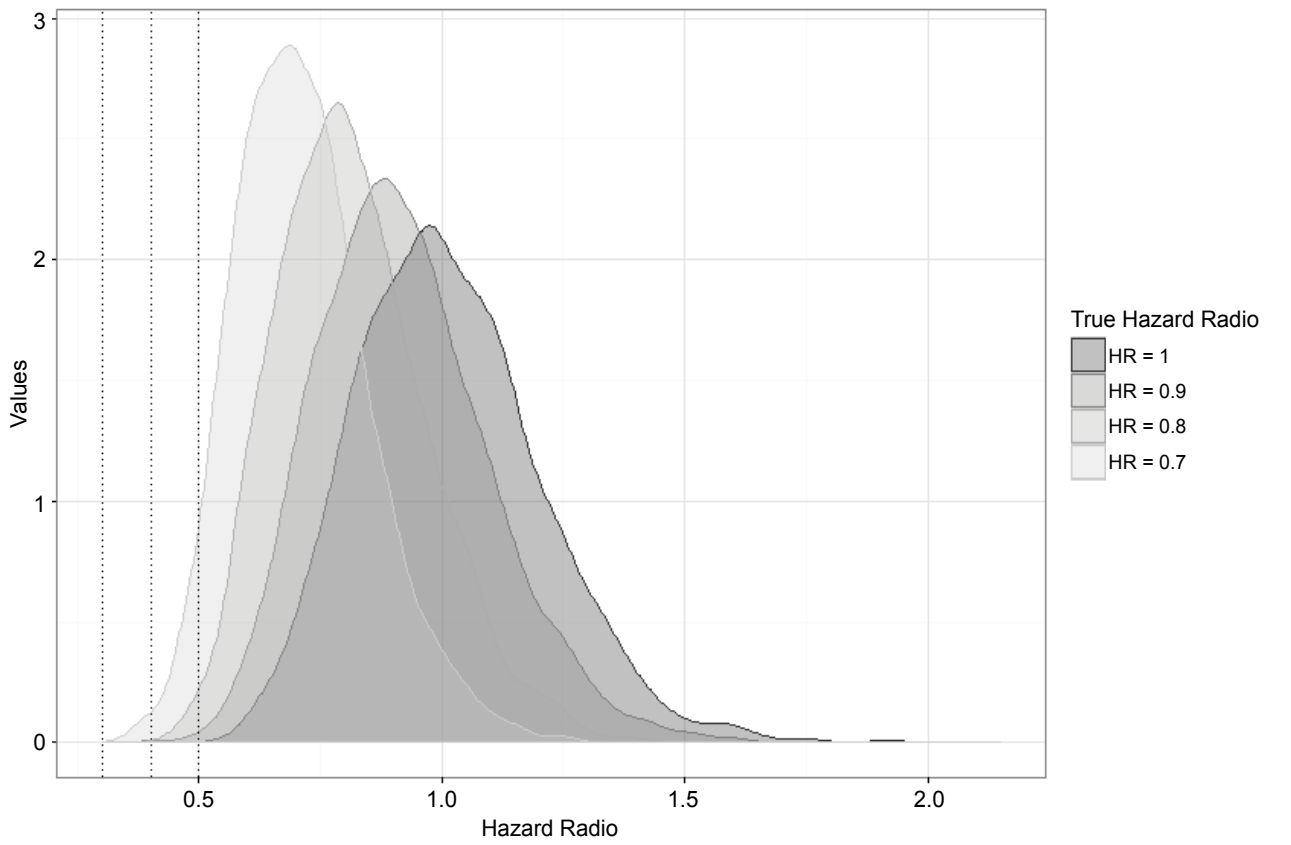

Figure 2: Density plots of hazard ratio estimates under each true hazard ratio.

promising results are due to truly effective innovative therapies without hesitation that they could be chance findings. The simulation studies described above were conducted to address concerns with small sample studies by assessing the likelihood of observing a large treatment effect when the true effect was actually modest to null.

Results from the simulation studies conducted have helped improve confidence that the effects observed are not likely due to chance, although they are most likely of smaller magnitude. They have also helped reinforce that observing a moderate effect in a small study is questionable, as the true treatment effect could be even more modest to null. In both cases, there is still potential for the confidence intervals associated with these hazard ratio estimates to be wide. However, limitations to the simulation included assuming proportional hazards and ignoring the effect of switchover of patients in the control arm to receive experimental treatment after disease progression. Thus, these results are limited solely to the simulation settings considered here.

Concerns also still remain for early phase studies as the simulation does not address issues brought up by Lara and Redman [9], regarding study quality and design in phase II, as well as Redman and Crowley [7] and Tuma [8], regarding imbalance in prognostic and predictive factors in small studies and heterogeneity in phase II populations. It was previously noted that it is not uncommon for larger confirmatory studies to find smaller magnitudes of benefit compared with smaller early phase studies. It has become clear that the reason for this difference goes beyond the small sample size issue. In fact, the reason may have more to do with poor study conduct in the form of unplanned interim looks, data-driven changes, ambiguous endpoints, as well as population heterogeneity. The FDA review for olaratumab [14] noted that there are still concerns about the heterogeneous population of the small study and a further randomized trial will be needed to generalize to other patient subgroups.

At the 2016 Friends of Cancer Research Annual Meeting [6], a panel consisting of regulatory, industry, and academic representatives discussed the optimization of exploratory randomized trials. It was noted that, moving forward, efforts need to be made to prospectively design trials that can potentially support both go/no-go decisions as well as registration. As one potential option, one panelist [6] presented a Bayesian analysis of unexpected survival “significance” in a randomized phase II trial. He proposed a method that combines prior beliefs about a hazard ratio with the results seen in the clinical trial to compute a posterior probability distribution of the hazard ratio. For example, this can then be used to get the posterior probability that the hazard ratio is ≤ 0.75 or 0.70, thresholds defined as being at least minimally clinically significant.

The olaratumab approval was an example case in

which a randomized phase II trial for a promising drug with breakthrough therapy designation planned on evaluating overall survival. For cases that meet similar criteria, it may be worthwhile for pharmaceutical companies to have a pre-specified contingency statistical analysis plan in anticipation of unexpectedly promising survival results. There are no set guidelines for how such a plan should be implemented. However, it should be understood that a well-conducted study of this sort would include a planned hypothesis for each of the endpoints considered and planned analysis timelines, absent from any ad-hoc or exploratory analyses until the final analyses are completed. In addition, clear pre-specified rules for increasing sample size should be included. For example, if the observed hazard ratio is < 0.5, then no sample size adjustment is needed. If the observed hazard ratio is between 0.5 and 0.75, a sample size increase might be needed. If the observed hazard ratio is greater than 0.75 then perhaps a phase 3 trial should be considered. The thresholds used here are for illustration purposes only. They are subject to change and should be dependent on disease setting, the comparative control treatment, and available therapies.

It goes without saying that Phase III trials still remain the standard for determining clinical benefit for the majority of products. Even when phase II trials are designed with our proposed considerations in place, approval should remain in the setting in which the trial was conducted and results should not be extrapolated to earlier settings or particular subgroups. Any such claims need to be studied on their own in a separate phase II or phase III study as appropriate. As the oncology drug development paradigm continues to shift, it will be increasingly important for FDA and industry to work together to find innovative design solutions in order to confidently have effective drugs available for patients in need in a timely manner.

References

1. Chabner BA (2011) Early Accelerated Approval for Highly Targeted Cancer Drugs. N Engl J Med 364: 1087-1089.
2. Sharma MR, Schilsky RL (2012) Role of randomized phase III trials in an era of effective targeted therapies. Nat Rev Clin Oncol 9: 208-214.
3. Horning SJ, Haber DA, Selig WKD, Ivy SP, Roberts SA, et al. (2013) Developing Standards for Breakthrough Therapy Designation in Oncology. Clin Cancer Res 19: 4297-4304.
4. FDA (2014) Guidance for Industry: Expedited Programs for Serious Conditions-Drugs and Biologics.
5. Sherman RE, Li J, Shapley S, Robb M, Woodcock J (2013) Expediting Drug Development - The FDA's New "Breakthrough Therapy" Designation. N Engl J Med 369: 1877-1880.
6. Friends of Cancer Research (2016) Annual Meeting Panel Three: Optimization of Exploratory Randomized Trials.
7. Redman M, Crowley J (2007) Small randomized trials. J Thorac Oncol 2: 1-2.
8. Tuma RS (2008) Examining heterogeneity in phase II trial designs may improve success in phase III. J Natl Cancer Inst 100: 164-166.
9. Lara PN, Redman MW (2012) The hazards of randomized phase II trials. Ann Oncol 23: 7-9.
10. FDA (2017) 22 Case Studies Where Phase 2 and Phase 3 Trials had Divergent Results.
11. Sinha G (2014) Downfall of Iniparib: A PARP inhibitor that doesn't inhibit PARP after all. J Natl Cancer Inst 106.
12. Johnson ME, Tolley HD, Bryson MC, Goldman AS (1982) Covariate Analysis of Survival Data: A Small-Sample Study of Cox's Model. Biometrics 38: 685-698.
13. Zhang JJ, Blumenthal G, He K, Tang S, Cortazar P, et al. (2012) Overestimation of the effect size in group sequential trials. Clin Cancer Res 18: 4872-4876.
14. FDA (2016) Olaratumab Review.

CLINMED INTERNATIONAL LIBRARY