JNCI Cancer Spectrum (2018) 2(4): pky037

doi: 10.1093/jncics/pky037
Article

ARTICLE

Effect Sizes Hypothesized and Observed in Contemporary Phase III Trials of Targeted and Immunological Therapies for Advanced Cancer

Nicola Jane Lawrence, Felicia Roncolato, Andrew Martin, Robert John Simes, Martin R. Stockler

See the Notes section for the full list of authors' affiliations.
Correspondence to: Dr Nicola Lawrence, BHB, MBCHB, FRACP, NHMRC Clinical Trials Centre, University of Sydney, Chris O'Brien Lifehouse Building, 119–143 Missenden Road, Camperdown, NSW 2050, Australia (e-mail: nicola.mitchell@gmail.com).

Abstract

Background: We sought to compare the effect sizes hypothesized in the trial design, observed in the trial results, and considered clinically meaningful by the American Society of Clinical Oncology (ASCO) 2014 recommendations, in phase III trials of targeted and immunological therapies.

Methods: We studied phase III, superiority trials of targeted and immunological therapies in advanced cancers published from 2005 to 2015. We recorded the characteristics, design parameters, and observed results for the primary endpoint of each trial. The effect sizes hypothesized in the trial design were compared with the ASCO 2014 recommendation that phase III trials be designed to detect overall survival (OS) benefits that are clinically meaningful (hazard ratio $\leq$0.8).

Results: All critical elements of the trial design (effect sizes hypothesized, estimated survival in the control group, power, and significance level) were identified in 165 of 213 included trials (77%). Of trials with a statistically significant result for the primary endpoint, 16 of 30 (53%) with a primary endpoint of OS and 20 of 53 (38%) with a primary endpoint of progression free survival (PFS) had an observed effect size less extreme than hypothesized; and 7 of 30 trials (23%) reported an observed effect size for OS that was statistically significant but not clinically meaningful (HR $>$ 0.80) according to the ASCO 2014 recommendations.

Conclusion: Many trials were designed such that an observed benefit in OS or PFS that was not clinically meaningful would be statistically significant. Phase III trials should be designed to provide results that are statistically significant for observed effects that are clinically meaningful but not for observed results that are of dubious clinical importance.

Well-designed clinical trials are pivotal for medical progress and the development of anti-cancer drugs. Randomized phase III trials are the gold standard for determining efficacy and safety in comparison with a valid control treatment (1). The aims of phase III trials are to provide valid and precise estimates of efficacy, safety, and net clinical benefit. They are consequently designed to have sufficient statistical power to reliably identify clinically important effects on survival and/or aspects of health-related quality of life (2,3) and have sample size targets that are inversely proportional to the (square root of the) hypothesized effect size.

There is increasing recognition that a clinical trial can yield a result that is statistically significant (indicative of a true effect) but of such a modest magnitude that is not considered clinically significant (benefit of marginal importance) (4,5). Although much has been written on the statistical versus clinical significance of trial evidence and its implications for trial design and practice (6–10), little attention has been given to the

Received: December 14, 2017; **Revised:** May 5, 2018; **Accepted:** July 13, 2018

contributing role of the effect size hypothesized in sample size calculations of oncology phase III trials. Targeted and immunological therapies have produced substantial improvements in outcomes for some patients with cancer and led to the hypothesis that clinical trials should be designed to look for larger treatment benefits (11).

We reviewed phase III trials of targeted and immunological therapies in advanced cancers to determine the magnitudes and predictors of the effect sizes hypothesized in trial designs; to compare the effect sizes hypothesized with the recommendations of the American Society of Clinical Oncology (ASCO) 2014 for the design of phase III trials; to compare the effect sizes hypothesized with two frameworks for rating the value of treatments from the observed results of cancer clinical trials; and to determine the relationship between the effect sizes hypothesized in trial designs with the effect sizes observed in trial results.

Methods

We used Medline to identify randomized controlled trials published from January 2005 to December 2015. The search included an extensive list of terms and synonyms for randomized controlled trial, advanced or metastatic cancer, and an outcome measure that included overall survival (OS), progression-free survival (PFS), or quality of life.

Two reviewers (NL and FR) independently selected phase III randomized controlled trials that tested a targeted therapy, including immunotherapies and hormone therapies, in nonhaematologic, metastatic malignancies. Eligibility was limited to superiority trials in adults with a primary endpoint of OS, PFS, or time to progression (TTP), and with 100 or more participants per treatment group. Trials were excluded that tested vaccines or radioisotopes or were not given systemically. Disagreements were resolved by discussion or adjudication by a third reviewer (MS).

Data were extracted from the primary publication of each trial along with the Supplementary Materials, including the protocol if publicly available. Trial characteristics included year of publication, journal name, country of corresponding author, sponsor, funding source, tumor type, treatment class, line of therapy, number of participating countries and sites, recruitment duration, median follow-up, trial design, sample size, and biomarker or tissue requirement for eligibility. We recorded the specifications used in the sample size calculation including the expected outcome in the control group (eg, median OS), the relative effect size hypothesized (typically a hazard ratio [HR]), the absolute effect size hypothesized (typically a difference in survival times or survival rates), the power, and the significance level. If only some of the information was explicitly provided, then the other elements were calculated where possible and a constant hazard over time was assumed. The relative effect size observed (in the results of the trial) for the primary endpoint was extracted, along with the *P* value and confidence intervals (CIs) for this endpoint. A trial was considered positive if the *P* value for the primary endpoint was less than the level specified for statistical significance in the protocol (typically a two-sided *P* value $<.05$) and/or the appropriately sized CI excluded a null effect.

We used simple descriptive statistics to summarize the distributions of effect sizes hypothesized and observed. Trials with PFS and TTP endpoints were grouped together for all analyses. The relationship between effect sizes hypothesized and observed were assessed with Spearman's rank correlation coefficient (r_s), differences were assessed with Mann-Whitney U tests, and predictors were assessed with multiple linear regression.

The effect sizes hypothesized for the design of each trial were compared with the ASCO 2014 recommendations for the design of phase III trials, in particular that a HR of 0.80 or less (more extreme, HR further from 1.0) should be the "minimum incremental improvement over standard therapy that would define a clinically meaningful outcome," with additional recommendations for clinically meaningful benefits in specific tumor types and settings (11).

The effect sizes hypothesized in each trial were also assigned clinical benefit scores according to two frameworks designed to evaluate the value of the observed results of cancer clinical trials, one proposed by ASCO in 2015 (12) and the other proposed by the European Society for Medical Oncology (ESMO) in 2015 (13). Clinical benefit scores were calculated as recommended by each group according to the effect size hypothesized in the trial design, not the observed results after the trial was completed. There were no adjustments for toxicity or symptom benefit. Our clinical benefit score based on the 2015 ASCO framework was calculated using the advanced disease framework that has discrete groupings for "OS score" and "PFS score" based on percentage change in median survival. This is based on the hypothesized relative percentage improvement in median OS or PFS. Our clinical benefit score based on the 2015 ESMO Magnitude of Clinical Benefit Scale (ESMO-MCBS v1.0) score was calculated using the "preliminary magnitude of clinical benefit grade" categorical scoring. This is based on the lower bound of the 95% CI for the HR, the absolute gain in survival time, and the absolute difference in 2- or 3-year survival rates. If the expected survival in the control group was not specified or calculable, we used the observed survival in the control group. If the hypothesized survival rates were not specified, they were calculated assuming exponential survival distributions and proportional hazards. The lower bound of the 95% CI for a hypothesized HR was calculated in accordance with the statistical power specified in the trial design. If the power and significance levels were not stated, we assumed they were 80% and 0.05 (two-sided), respectively.

Results

Study Cohort

We included the 213 phase III trials that met our selection criteria from a total of 5708 publications identified by our search (Supplementary Appendix Figure 1, available online). Absolute agreement between the two reviewers was 97%.

The most common tumor types were lung, breast, colorectal, and melanoma. The experimental agents evaluated included tyrosine kinase inhibitors (n = 84, 39%), monoclonal antibodies other than immunotherapeutics (n = 61, 29%), immunotherapeutics (n = 20, 9%), and hormonal agents (n = 16, 8%). The majority of trials (n = 159, 75%) were commercially sponsored and/or received commercial funding (n = 186, 87%). First authors were predominantly from North America (n = 97, 46%) and Europe (n = 72, 34%). The primary endpoint was OS in 118 trials (55%), PFS in 86 (40%), and TTP in 9 (4%). Tumor types with longer expected survival times, such as breast and ovarian, used PFS as the primary endpoint more commonly than trials with shorter expected survival times, such as gastric and pancreatic

Table 1. Characteristics of phase III trials*

Characteristic	Number of trials by primary endpoint		Total number of trials	
	OS	PFS	n	%
Tumor Type				
Lung	33	22	55	26
Breast	4	33	37	17
Colorectal	15	13	28	13
Melanoma	17	4	21	10
Renal	7	10	17	8
Prostate	12	0	12	6
Gastric / GOJ	8	1	9	4
HCC	9	0	9	4
Pancreas	6	0	6	3
Head and neck	4	1	5	2
Ovarian / peritoneal	1	4	5	2
Other	2	7	9	4
Experimental agent				
Tyrosine kinase inhibitor	37	47	84	39
Monoclonal antibody	32	29	61	29
Immunotherapy	20	0	20	9
Hormonal therapy	8	8	16	8
mTOR inhibitors	3	9	12	6
Other	18	2	20	9
Sponsor				
Industry	83	76	159	75
Academic	31	18	49	23
Unspecified	4	1	5	2

*HCC = hepatocellular carcinoma; GOJ = gastro-oesophageal junction; mTOR = mammalian target of rapamycin; OS = overall survival, PFS = progression free survival.

(Table 1). The median number of participants per treatment group was 298 (interquartile range [IQR] = 194 to 412). The majority of trials (n = 191, 90%) had two treatment groups, 16 (8%) had three, and 6 (3%) had four. Cross-over to experimental treatment after progression in the control group was allowed in 13 of 118 trials (11%) with a primary endpoint of OS, although in 7 of 13 trials this occurred after reporting an interim analysis for the primary endpoint of OS. A biomarker result (eg, hormone receptor status) was required for participant eligibility in 54 trials (25%). Tissue for analysis at a central laboratory was required for participant eligibility in 37 trials (17%).

Statistical Design and Hypothesized Effect Size

The proportion of trials with sufficient information to determine all critical elements of the design (relative and absolute effect sizes hypothesized, estimated survival in the control group, power, and significance level) was 165 of 213 (77%). Sufficient information was available to determine the relative and the absolute effect sizes hypothesized in 209 trials (98%). The hypothesized survival in the control group was able to be determined in 170 trials (80%). In these trials, the observed survival in the control group was somewhat longer than hypothesized (absolute difference in medians: 0.5 months, IQR = 0.4 to 2.3; ratio of medians = 1.1, IQR = 0.9–1.3). The power and significance levels were both specified in 96% of trials.

The distributions of relative (n = 209) and absolute (n = 208) effect sizes hypothesized are shown in Figure 1. Trials with a primary outcome of PFS hypothesized more extreme relative effect sizes (median HR = 0.71, IQR = 0.67–0.75) than trials with a primary outcome of OS (median HR = 0.75, IQR = 0.73–0.78), *P* values less than .0001. Trials with a primary outcome of PFS hypothesized smaller absolute effect sizes (median increase of 2.4 months, IQR = 1.5–3.0) than trials with a primary outcome of OS (median increase of 3.0 months, IQR = 2.2–4.0), *P* = .0002.

Trial characteristics associated with the relative effect size hypothesized were: trial endpoint (PFS associated with more extreme HR than OS), tumor type (trials in melanoma associated with a more extreme HR than trials in other tumor types [see footnote in Table 2]), class of treatment, power (higher power associated with more extreme HR), expected survival in the control group (accounting for the primary endpoint, shorter expected survival in the control group was associated with a more extreme HR), number of study sites (less sites associated with more extreme HR), and treatment line (first line associated with a less extreme HR). There were no apparent associations with type of sponsor, publication year, statistical significance proposed in the trial design, or geographic region of the corresponding author (Table 2).

The ASCO 2014 recommendation that phase III trials evaluating OS should be designed to detect a clinically meaningful HR of 0.80 or lower (more extreme) (11) was met by 98% (113/115) of the trials with a primary endpoint of OS. The ASCO recommendations for first-line treatment of non-small-cell lung and pancreatic cancer were more likely to be met for the relative than the absolute effect sizes hypothesized (Table 3). There were insufficient trials for comparisons with the recommendations for the remaining two scenarios.

A grading of clinical benefit from the ASCO 2015 framework (12) was applicable to the effect size hypothesized in the design of 209 of 213 trials (98%). Approximately three-quarters of the trials (172/209, 81%) had effect sizes hypothesized that represented an improvement in median OS or PFS of greater than 25% to 49% as per the ASCO framework (Table 4). A grading of clinical benefit from the ESMO-MCBS v1.0 2015 framework (13) was applicable to the effect size hypothesized in the design of 208 of 213 trials (98%). Approximately three-quarters of the trials (161/208, 77%) had effect sizes hypothesized that reflected the potential to attain a "high level of proven clinical benefit" (a score of 3 or 4 that could be upgraded to a score of 4 or 5 if additional criteria were met for quality of life or toxicity, Table 4). A comparison of the grading according to the ASCO and ESMO frameworks is shown in the Supplementary Appendix, Table 1 (available online).

Observed Effect Size

The number (proportion) of trials with observed results that were statistically significant for the primary endpoint was 83 (39%) for all trials, 53 (56%) for those based on PFS, and 30 (25%) for those based on OS. The relationships between the relative effect size hypothesized in the trial design and the relative effect size observed in the trial results are shown in Figure 2, A and B. The correlation between the effect size hypothesized vs the effect size observed was moderate for trials based on PFS ($r_s = 0.38$) and minimal for trials based on OS ($r_s = 0.14$). Effect sizes hypothesized were more extreme (HR closer to 0) than effect sizes observed ($P < .0001$). Of the trials with a statistically significant primary endpoint, 20 of 53 (38%) based on PFS and 16 of 30 (53%) based on OS had an observed effect size less extreme (HR closer to 1) than hypothesized.

The ASCO 2014 recommendation (11) for a clinically meaningful benefit in OS (HR = 0.80 or lower) was met by 23 of 30

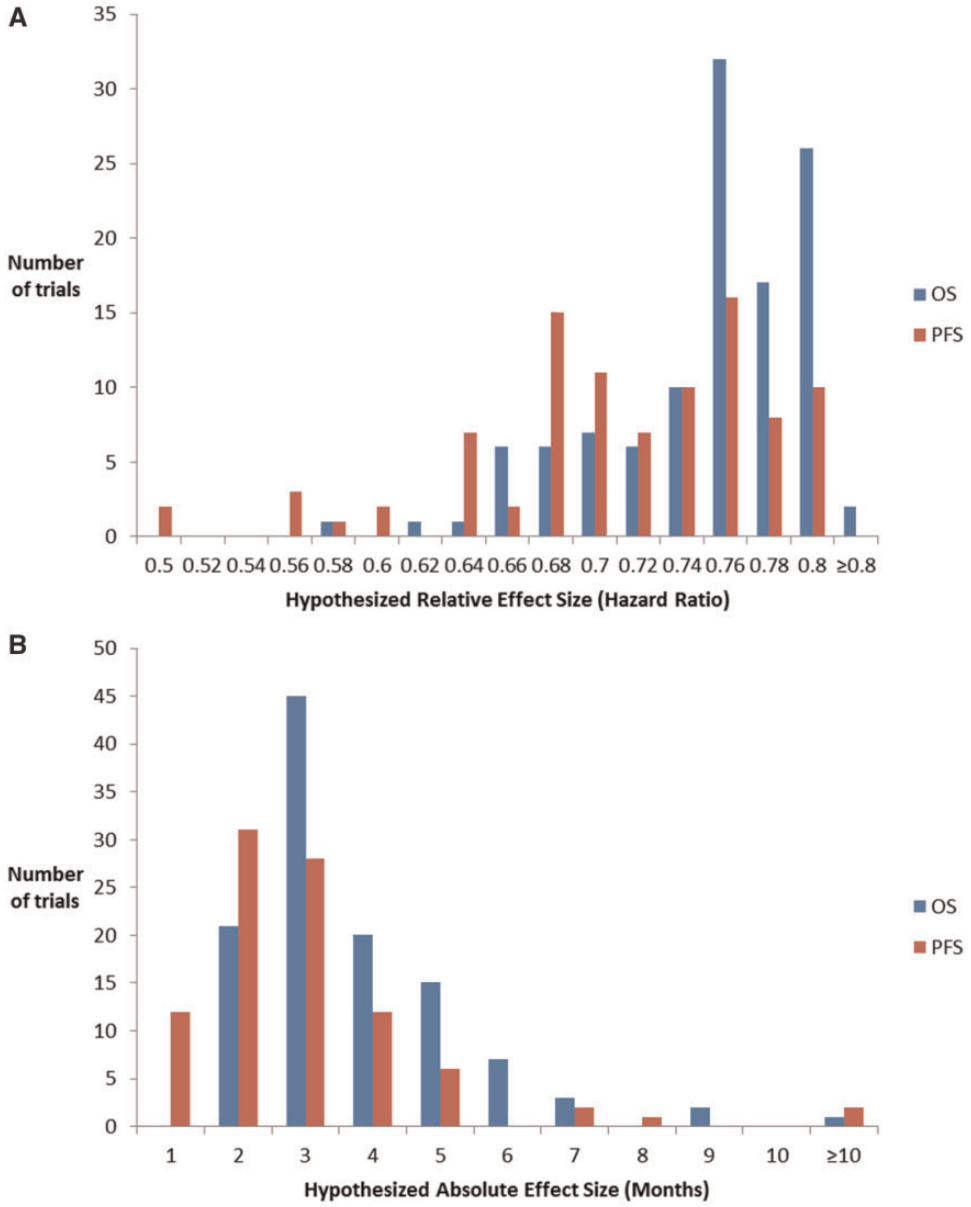

Figure 1. Distribution of hypothesized effect size. **A**) Distribution of relative effect size hypothesized by primary endpoint. **B**) Distribution of absolute effect size hypothesized by primary endpoint. OS = overall survival; PFS = progression free survival.

trials (77%) with a statistically significant HR for the primary endpoint of OS observed in the results. The remaining 7 trials (23%) reported an observed HR for OS that was statistically significant but not clinically meaningful (HR >0.80, see Figure 2A). Cross-over was not allowed in six of these seven trials and not specified in the remaining trial.

Discussion

Sufficient information to determine all critical elements of the trial design were available in 77% of contemporary phase III trials of targeted agents in advanced solid malignancies. The median effect size hypothesized in trials with a primary endpoint of OS was an HR of 0.75 (25% reduction in the hazard for death and 33% prolongation of median survival time) or an absolute benefit of 3 months, and for PFS was an HR of 0.71 (29% reduction in the hazard for death and 41% prolongation of median survival time), or an absolute benefit of 2.4 months. The effect size hypothesized was moderately correlated with the effect size observed for trials with a primary endpoint of PFS, but not for trials with a primary endpoint of OS. Twenty-three percent of trials with a statistically significant result for the primary endpoint of OS had point estimates for an observed benefit that did not meet the ASCO 2014 recommendation for the "minimum incremental improvement over standard therapy that would define a clinically meaningful outcome" (11).

The absolute benefits hypothesized for trials with PFS as the primary endpoint were smaller than for OS, whereas the relative benefits hypothesized for trials based on PFS were larger (more extreme, HR closer to 0) than for OS. This may simply reflect the shorter duration of PFS than OS, meaning that a given absolute benefit translates into a larger relative benefit in PFS than in OS. Trials in advanced lung and pancreatic cancer were more likely to have effect sizes hypothesized that met the ASCO

2014 targets for relative effects than for absolute effects (94% vs 17%). We propose that expressing effects in relative terms may provide a better depiction of a treatment's activity for researchers because the relative benefit of the experimental intervention vs the control accounts for differences in outcomes over time and accurately represents the biological effect of the experimental agent. However, expressing effects in absolute terms may provide a better depiction of a treatment's benefit for individual patients and doctors making clinical decisions because the expected survival time with the comparator and the absolute benefit of the experimental treatment are crucial considerations and easier to understand and appreciate (14).

Table 2. Characteristics associated with the relative effect size hypothesized

Characteristic	Mean HR	*P*
Trial endpoint		<.0001
OS	0.75	
PFS	0.71	
Tumor type*		<.0001
Breast	0.72	
Colorectal cancer	0.74	
Lung	0.75	
Melanoma	0.67	
Other	0.72	
Prostate	0.77	
Renal	0.72	
Class of treatment		.03
Hormonal therapies	0.75	
Immunotherapeutics	0.72	
Monoclonal antibodies	0.74	
mTOR inhibitors	0.73	
Other	0.75	
Tyrosine kinase inhibitors	0.71	
Power ↑	↓HR	.0013
Expected control survival ↓(adjusted for primary endpoint variable)	↓HR	.0021
Number of participating sites ↓	↓HR	<.0001
Treatment line		.0013
First line	0.74	
Second line	0.73	
Other	0.70	
Sponsor		.35
Academic	0.74	
Commercial	0.73	
Publication year	–	.47
Statistical significance proposed in trial design	–	.81
Region of corresponding author		.08
Canada	0.76	
Europe	0.73	
ROW	0.71	
UK	0.72	
US	0.72	

*The pairwise comparisons revealed significant differences between melanoma and breast ($P = .04$), melanoma and colorectal ($P = .0004$), melanoma and lung ($P < .0001$), melanoma and prostate ($P < .0001$), and melanoma and other tumor types ($P = .01$), and breast and prostate ($P = .04$). There were no significant differences between melanoma and renal ($P = .10$). mTOR = mammalian target of rapamycin; OS = overall survival; PFS = progression free survival.

An ASCO 2014 working group recommended that clinical trials of targeted drugs should be designed to seek larger benefits than those sought and achieved in trials of cytotoxic chemotherapy, with a recommended HR of 0.80 or lower chosen to represent a clinically meaningful improvement in OS (11). Any criterion for clinical significance is arbitrary. Drugs with substantial toxicity may require larger benefits to warrant the side effects of treatment, while for drugs with minimal toxicity smaller benefits may be justified. Expensive drugs may require large benefits to justify the initial cost of treatment, but the longer term cost-effectiveness potential (eg, when treatment is off-patent or is used in an earlier disease setting) should also be considered. The ideal effect size hypothesized should depend on all these factors as well as the disease setting. A single criterion will not be applicable to all circumstances; nevertheless, we agree that an HR of 0.80 or lower for OS is a reasonable, useful, and appropriate criterion of clinical benefit for an expensive and/or toxic treatment.

Interpretation of the absolute benefits of a treatment based on observed differences in median survival times requires caution. The difference in observed median survival times represents two single points on the survival curves, not a comparison over the entire time course of follow-up and is therefore susceptible to random fluctuations that may not accurately reflect the effect of a treatment throughout follow-up. This is not a problem for the hypothesized absolute benefit in the trial design, because one typically assumes a proportional hazards model that defines the relationship between medians, and in fact any time point, and so the same treatment effect is represented throughout follow-up. The proposed absolute benefits of a treatment should be considered carefully and specified in the statistical hypothesis to ensure that the potential treatment benefit is clinically meaningful in that particular disease and setting.

We used the ASCO 2014 criterion (HR of 0.80 or lower) to assess the clinical importance of the effect sizes hypothesized in the design of contemporary oncology phase III trials of targeted therapies. The majority (98%) of trials in our analysis with a primary endpoint of OS were designed with a target HR of 0.80 or lower and therefore met this criterion. However, an adequately powered clinical trial will yield statistically significant results for benefits that are smaller than those hypothesized. For example, the 64 trials (54%) with a target HR of 0.72 to 0.79 for OS

Table 3. Number of trials with hypothesized treatment effects that met ASCO's 2014 disease-specific recommendations for phase III trials*

	Non-small-cell lung cancer trials (n = 30)		Pancreatic cancer trials (n = 6)	
ASCO recommendation	Relative effect size (HR ≤ 0.8)	Absolute effect size (≥2.5 months squamous, ≥3.25 months nonsquamous)	Relative effect size (HR ≤ 0.75)	Absolute effect size (≥ 3 months)
Met	29 (97%)	6 (20%)	5 (83%)	0 (0%)
Not met	0 (0%)	23 (73%)	1 (17%)	6 (100%)
Indeterminate	1 (3%)	1 (3%)	0 (0%)	0 (0%)

*ASCO = American Society of Clinical Oncology.

Table 4. The effect size hypothesized in phase III trials according to the ESMO and ASCO frameworks for clinical benefit*

	Primary outcome of OS	Primary outcome of PFS or TTP	Total
	n = 118	n = 95	n = 213
ESMO 2015 clinical benefit grade			
1 Least benefit	1 (1%)	16 (17%)	17 (8%)
2	6 (5%)	24 (25%)	30 (14%)
3	63 (53%)	54 (57%)	117 (55%)
4 Most benefit	44 (37%)	NA	44 (21%)
Indeterminate	4 (3%)	1 (1%)	5 (2%)
ASCO 2015 clinical benefit grade			
Improvement in median OS or PFS			
0> to 24%	2 (2%)	0 (0%)	2 (1%)
25 to 49%	102 (86%)	70 (74%)	172 (81%)
50 to 75%	11 (9%)	19 (20%)	30 (14%)
76 to 100%	0 (0%)	4 (4%)	4 (2%)
>100%	0 (0%)	1 (1%)	1 (0.5%)
Indeterminate	3 (3%)	1 (1%)	4 (2%)

*ASCO = American Society of Clinical Oncology; ESMO = European Society for Medical Oncology; OS = overall survival; PFS = progression free survival.

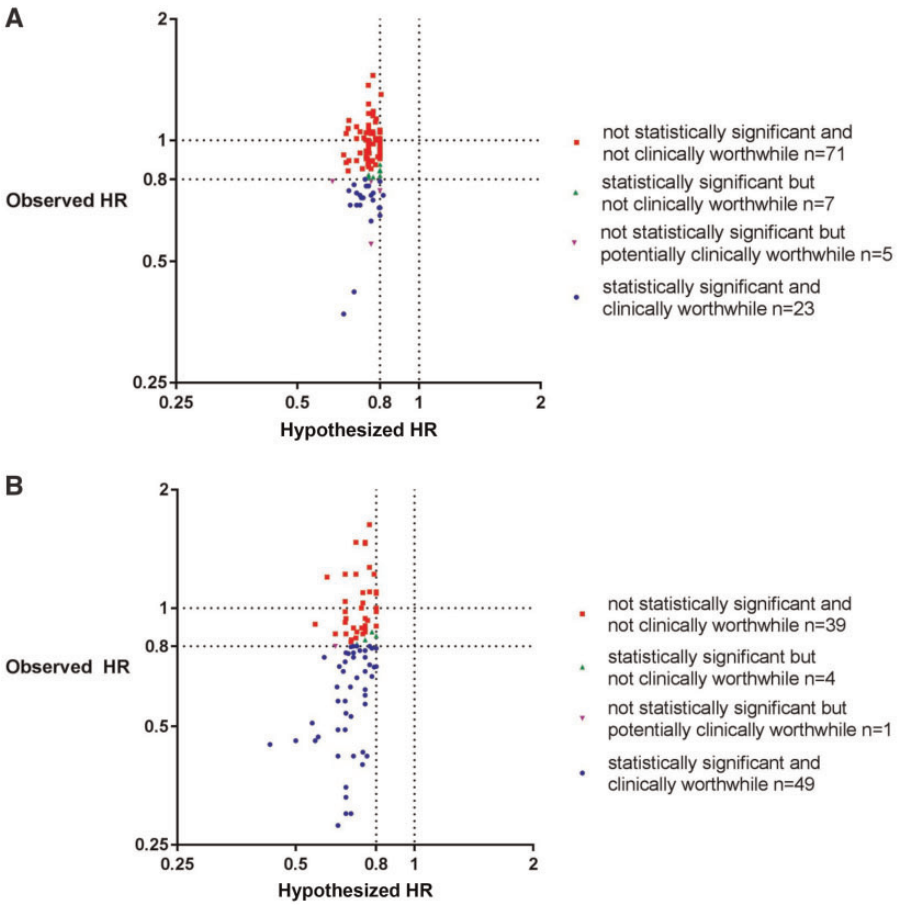

Figure 2. Effect sizes hypothesized versus observed. **A**) Overall survival. **B**) Progression-free survival.

would be expected to yield statistically significant results with an observed HR around 0.80 to 0.85 if their power was 80%, or with an observed HR around 0.82 to 0.875 if their power was 90%. The ASCO criterion implies that a phase III trial producing an observed HR of 0.80 or greater (HR closer to 1), but with an associated 95% CI indicating that a HR of less than 0.80 is plausible, may not be sufficiently compelling to change practice. This calls into question the merit of designing phase III trials that have high statistical power to detect hypothesized HRs of 0.75 to 0.80 and suggests it is unwise to pursue development to the phase III setting without a strong rationale for hypothesizing a substantial effect size (eg, HR of 0.75 or lower).

In our study, the effect sizes hypothesized in the trial design were assigned greater value with the ESMO framework (13) than

with the ASCO framework (12). This partly reflects the ESMO framework's use of the lower bound of the 95% CI for the HR and assigning value for gains in survival rates at 2 or 3 years, whereas the ASCO framework only considered improvements in median survival time. The updated ASCO 2016 framework (15) uses the HR rather than median survival and has assigned value for benefits in the tails of survival curves; this may reduce some of the discrepancies between the frameworks. The updated ESMO-MCBS v1.1 2017 framework (16) removed the value assigned for small gains in survival rates and revised the required absolute survival gains for patients with prolonged expected survival. However, retaining the use of the lower bound of the 95% CI for the HR, and the minimal changes to the scoring of comparative trials in field testing mean the discrepancies between the ASCO and ESMO frameworks are likely to remain.

Our assessment of the effect sizes hypothesized in phase III trials according to the ESMO-MCBS v1.0 framework differs from the recent publication by Del Paggio et al. (17), which found 31% of their included trials were designed to detect an effect size that could attain a "high level of proven clinical benefit" compared with 77% in our study. Likely explanations for the differing results is their use of the point estimate of the HR in the statistical hypothesis, whereas we used the lower bound of the 95% CI and the survival rate differences as recommended. We limited trials to targeted therapies in solid organ tumors only, whereas they included trials of chemotherapy, but only in breast, lung, colorectal, and pancreatic malignancies.

The main limitation of our study is our use of the ASCO and ESMO frameworks to grade the effect sizes hypothesized in the trial design, whereas they were developed to grade effect sizes observed in the trial findings. A number of assumptions were necessary to grade the effect sizes hypothesized for the ESMO framework, including calculations of the lower bound of the 95% CI of the HR and estimated differences in survival rates.

Escalating healthcare costs are a global problem, particularly in oncology because of the increasing incidence of cancer due to aging of the population and the expense of new anticancer treatments. Many have recommended increasing attention on the value of anticancer treatments to prioritize resource allocation (18,19), and this has primarily focused on the interpretation of the results from completed clinical trials (20,21). We found that many contemporary trials were designed to detect small effects. As a consequence, approximately one in four trials with a significant *P* value for OS observed a point estimate for benefit that was not clinically meaningful. Randomized phase III trials should be designed to provide statistically significant results for observed effects, relative and absolute, that are clinically meaningful, and trial reports should discuss the potential value of new treatments in terms of their net clinical benefit.

Acknowledgments

NL received support from an NHMRC Postgraduate Scholarship (Australia), Goldman Sachs New Zealand Fellowship (New Zealand), and a Top-Up PhD Scholarship from Sydney Catalyst (University of Sydney and Cancer Institute NSW, Australia).

Declaration of interests

MS reports his institution received research funding from Astellas Pharma, Celgene, Bayer, Bionomics, Medivation, Sanofi, Pfizer, AstraZeneca, Bristol-Myers Squibb, and Roche. NL, FR, AM, and RS report no conflicts of interest.

Notes

Affiliations of authors: NHMRC Clinical Trials Centre, University of Sydney, Camperdown, New South Wales, Australia (NJL, FR, AM, RJS, MRS); Macarthur Cancer Therapy Centre, Campbelltown, New South Wales, Australia (FR); Concord Cancer Centre, Concord Repatriation General Hospital, Concord, New South Wales, Australia (MRS); Chris O'Brien Lifehouse, Camperdown, New South Wales, Australia (MRS).

References

1. Markman M. Recent reminders of why the gold standard for clinical research in oncology is the well-designed and conducted randomized phase III trial. *Curr Oncol Rep*. 2004;6(6):421–422.
2. Booth CM, Tannock I. Reflections on medical oncology: 25 years of clinical trials—where have we come and where are we going? *J Clin Oncol*. 2008;26(1):6–8.
3. Wilson MK, Karakasis K, Oza AM. Outcomes and endpoints in trials of cancer treatment: the past, present, and future. *Lancet Oncol*. 2015;16(1):e32–e42.
4. Saltz LB. Progress in cancer care: the hope, the hype, and the gap between reality and perception. *J Clin Oncol*. 2008;26(31):5020–5021.
5. Ocana A, Tannock IF. When are "positive" clinical trials in oncology truly positive? *J Natl Cancer Inst*. 2011;103(1):16–20.
6. Case LD, Kimmick G, Paskett ED, et al. Interpreting measures of treatment effect in cancer clinical trials. *Oncologist*. 2002;7(3):181–187.
7. Jardim DL, Schwaederle M, Wei C, et al. Impact of a biomarker-based strategy on oncology drug development: a meta-analysis of clinical trials leading to FDA approval. *J Natl Cancer Inst*. 2015;107(11): pii: djv253.
8. Sacher AG, Le LW, Leighl NB. Shifting patterns in the interpretation of phase iii clinical trial outcomes in advanced non–small-cell lung cancer: the bar is dropping. *J Clin Oncol*. 2014;32(14):1407–1411.
9. Mehta C, Schafer H, Daniel H, et al. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Stat Med*. 2014;33(26):4515–4531.
10. Escudier B, Heng DY, Smyth-Medina A, et al. Considerations for the design of future clinical trials in metastatic renal cell carcinoma. *Clin Genitourin Cancer*. 2014;12(1):1–12.
11. Ellis LM, Bernstein DS, Voest EE, et al. American Society of Clinical Oncology perspective: raising the bar for clinical trials by defining clinically meaningful outcomes. *J Clin Oncol*. 2014;32(12):1277–1280.
12. Schnipper LE, Davidson NE, Wollins DS, et al. American Society of Clinical Oncology statement: a conceptual framework to assess the value of cancer treatment options. *J Clin Oncol*. 2015;33(23):2563–2577.
13. Cherny NI, Sullivan R, Dafni U, et al. A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS). *Ann Oncol*. 2015;26(8):1547–1573.
14. Studts JL, Abell TD, Roetzer LM, et al. Preferences for different methods of communicating information regarding adjuvant chemotherapy for breast cancer. *Psychooncology*. 2005;14(8):647–660.
15. Schnipper LE, Davidson NE, Wollins DS, et al. Updating the American Society of Clinical Oncology value framework: revisions and reflections in response to comments received. *J Clin Oncol*. 2016;34(24):2925–2934.
16. Cherny NI, Dafni U, Bogaerts J, et al. ESMO-magnitude of clinical benefit scale version 1.1. *Ann Oncol*. 2017; doi:10.1093/annonc/mdx310.
17. Del Paggio JC, Azariah B, Sullivan R, et al. Do contemporary randomized controlled trials meet ESMO thresholds for meaningful clinical benefit? *Ann Oncol*. 2016; doi:10.1093/annonc/mdw538.
18. Kleinrock M. Oncology innovation and challenging choices: balancing value and funding priorities in light of an abundance of new treatment options. *Am Health Drug Benefits*. 2015;8(4):196–199.
19. Schnipper LE, Schilsky RL. Converging on the value of value frameworks. *J Clin Oncol*. 2017;35(24):2732–2734.
20. Grossmann N, Del Paggio JC, Wolf S, et al. Five years of EMA-approved systemic cancer therapies for solid tumours-a comparison of two thresholds for meaningful clinical benefit. *Eur J Cancer*. 2017;82:66–71.
21. Bentley TGK, Cohen JT, Elkin EB, et al. Measuring the value of new drugs: validity and reliability of 4 value assessment frameworks in the oncology setting. *J Manag Care Spec Pharm*. 2017;23(6-a suppl):S34–S48.