Published in final edited form as:
Clin Cancer Res. 2012 April 15; 18(8): 2309–2315. doi:10.1158/1078-0432.CCR-11-1815.

Resampling phase III data to assess phase II trial designs and endpoints

Manish R. Sharma, **Theodore G. Karrison**, **Yuyan Jin**, **Robert R. Bies**, **Michael L. Maitland**, **Walter M. Stadler**, and **Mark J. Ratain**
Department of Medicine (MRS, MLM, WMS, MJR), Department of Health Studies (TGK), Comprehensive Cancer Center (TGK, MLM, WMS, MJR), and Committee on Clinical Pharmacology and Pharmacogenomics (MRS, MLM, MJR), University of Chicago, Chicago, IL, Clinical Pharmacology Research Unit, Pfizer Inc, Groton, CT (YJ), and Departments of Medicine and Medical Genetics, Indiana University School of Medicine, Indianapolis, IN (RRB)

Abstract

Purpose—The best phase II design and endpoint for growth inhibitory agents is controversial. We simulated phase II trials by resampling patients from a positive (sorafenib vs. placebo; TARGET) and a negative (AE941 vs. placebo) phase III trial in metastatic renal cancer to compare the ability of various designs and endpoints to predict the known results.

Experimental design—770 and 259 patients from TARGET and the AE 941 trial, respectively, were resampled (5,000 replicates) to simulate phase II trials with alpha = 0.10 (one-sided). Designs/endpoints: single arm, two-stage with response rate (RR) by RECIST (37 patients); and randomized, two arm (20–35 patients/arm) with RR by RECIST, mean log ratio of tumor sizes (log ratio), PFS rate at 90 days (PFS-90), and overall PFS.

Results—Single arm trials were positive with RR by RECIST in 55% and 1% of replications for sorafenib and AE 941, respectively. Randomized trials versus placebo with 20 patients per arm were positive with RR by RECIST in 55% and 7%, log ratio in 88% and 25%, PFS-90 in 64% and 15%, and overall PFS in 69% and 9% of replications for sorafenib and AE 941, respectively.

Conclusions—Compared with the single arm design and the randomized design comparing PFS, the randomized phase II design with the log ratio endpoint has greater power to predict the positive phase III result of sorafenib in renal cancer, but a higher false positive rate for the negative phase III result of AE 941.

Keywords

resampling; phase II; randomized; single arm; tumor size

Introduction

Phase II trials in oncology assess whether a novel agent or combination has enough antitumor activity to warrant further study. While there are many definitions of antitumor activity, the most widely adopted is an “objective response”, defined as a reduction in the sum of unidimensional target lesions by 30% or more according to the Response Evaluation Criteria in Solid Tumors (RECIST).[1] Single arm phase II trials in oncology typically use the

Corresponding author: Mark J. Ratain, MD, mratain@medicine.bsd.uchicago.edu, 5841 S. Maryland Avenue, MC 2115, Chicago, IL 60637-1470, Phone: (773) 702-4400, Fax: (773) 702-3969.

Previous presentation:
Presented, in part, at the 46th Annual Meeting of the American Society of Clinical Oncology, June 2–5, 2010, Chicago, IL.

proportion of objective responders (i.e., response rate) as the primary measure of efficacy.[2] Although there is increasing support for randomized phase II trials in oncology[3–6], oncology phase II trials are still significantly less likely to use control subjects than comparable trials in other specialties.[7]

Unfortunately, the high rate of failure in phase III oncology trials emphasizes that current phase II trials are not sufficiently informative. Only 57% of oncology drugs succeed in phase III trials, compared with 68% of non-oncology drugs.[8] Furthermore, less than 5% of positive phase II combination therapy trials have eventually resulted in improved standards of care.[9] While it is possible that the prevalence of truly active agents is lower in oncology than in non-oncology, it is reasonable to question whether single arm designs with response rate endpoints are contributing to the low success rate. Ratain and Eckhardt pointed out that many drugs, particularly those developed based on activity against a molecular target, may be active without causing a high degree of tumor regression.[10] El Maraghi and Eisenhauer found that four such drugs that were eventually approved on the basis of clinical benefit had RECIST response rates less than 10%, two of them less than 5%.[11] These data call into question the wisdom of using RECIST response rate as the endpoint for phase II trials.

Another major shortcoming of response rate as an endpoint is the loss of statistical information that occurs when a continuous change in tumor size is dichotomized.[12] To avoid this, Karrison and colleagues proposed using the mean log ratio of tumor size at eight weeks versus baseline as the endpoint in a randomized phase II trial[13], and subsequent modeling studies have supported the potential utility of this endpoint.[14,15] In the setting of measurable disease, calculating the log ratio does not require any additional data compared with RECIST response rate or PFS.

To explore these issues further, we evaluated a variety of phase II designs and endpoints utilizing data from two completed phase III trials in renal cancer. The first is a phase III trial of sorafenib versus placebo (TARGET), a “positive” trial that demonstrated a clinically meaningful improvement in progression-free survival (PFS) and led to approval by the U.S. Food and Drug Administration of sorafenib for renal cancer.[16] The second is a phase III trial of AE 941 (shark cartilage extract) versus placebo in the same disease, a “negative” trial that demonstrated no improvement in PFS.[17] In order to compare the operating characteristics of single arm and randomized phase II designs with various endpoints, we randomly selected (hereafter, resampled) individual patients from the actual phase III trials and simulated phase II trials using data from these patients. The ideal design and endpoint would accurately predict both the positive phase III result with sorafenib and the negative phase III result with AE 941.

Methods

Data sets

TARGET randomized 903 patients with advanced renal cancer who had failed prior immunotherapy to sorafenib 400 mg (n = 451) or placebo (n = 452) by mouth twice daily, and investigator assessed data was obtained from the analysis conducted in May 2005. The AE 941 trial randomized 305 patients with metastatic renal cancer who had failed prior immunotherapy to AE 941 120 mL (n = 153) or placebo (n = 152) by mouth twice daily. Both trials were designed with overall survival as the primary endpoint. TARGET was powered to detect a hazard ratio of 1.33, but was stopped early after a pre-specified interim analysis evaluating PFS. The AE 941 trial was powered to detect a hazard ratio of 1.5. The proportional hazards assumption was reasonably well met in TARGET, and the observed PFS curves were virtually the same in the AE 941 trial. Additional methodological details are available in the original reports of these two trials.[16,17]

133 and 46 patients from TARGET and the AE 941 trial, respectively, were excluded from our analyses (Figures 1a, 1b). Data included treatment assignment, unidimensional measurements of target lesions from computed tomography (CT) scans (every six weeks in TARGET; every eight weeks in the AE 941 trial), PFS, and whether or not the event was censored. Tumor measurements were extracted as follows: (1) identified lesions measured consistently across all CT scans; (2) kept the three largest lesions; and (3) excluded lymph nodes if three other lesions were available. The revised target lesions were summed to calculate tumor sizes at baseline and at six or eight weeks. Patients who missed the first CT scan had tumor sizes imputed by assuming a linear change between baseline and the second CT scan. Those who had a PFS event before the first CT scan had tumor sizes imputed by assuming a percentage increase equal to the largest in that arm; the alternative approach of inverse probability weighting could not be used because covariates that correlate with outcomes were not available in the provided data sets.

Resampling simulations

Random sampling with replacement was performed at the level of the individual patient. Sampling with replacement was conducted because this effectively treats the empirical cumulative distribution function as reflective of the population distribution. Five thousand replicates were performed for each simulated trial, yielding 95% confidence interval (CI) widths of less than + 1.5% for the simulation margin of error. Simulated trials were classified as positive or negative according to statistical criteria outlined below, and the percentage of positive trials was compared across designs, endpoints, and sample sizes. Statistical analyses were done using standard software (STATA, version 11.2). A one-sided alpha (Type I error rate) of 0.10 was used in all cases.

Study designs evaluated

The single arm design was an optimal, two-stage design[2], in which patients were resampled from the sorafenib or AE 941 arm only. In order to test the null hypothesis that response rate (RR) by RECIST[1] at six weeks (for TARGET) or eight weeks (for AE 941) was ≤ 5% versus the alternative that RR was ≥ 20% with 90% power, 37 patients were required. Trials stopped early if there were no responses in the first 12 patients, and were positive if there were ≥ four responses in 37 patients. This design was selected because it reflects what investigators would likely choose for a single arm trial based on RR by RECIST for the two drugs in question. Other endpoints, such as median PFS, were not considered for single arm designs, since results using such endpoints are difficult to interpret in the absence of concurrent controls.[18] This is especially true for renal cancer, as the unpredictable natural history of the disease makes historical controls unreliable.[19]

Randomized designs of sorafenib versus placebo or AE 941 versus placebo were performed with 1:1 randomization and sample sizes of 20, 25, 30, and 35 patients per arm. The sample size of 20 patients per arm was selected to correspond approximately to the sample size for the single arm design. The sample sizes of 25, 30 and 35 patients per arm were chosen to test the effect of increasing sample size, while keeping the total number of patients feasible for a phase II trial conducted at a single institution or through a small consortium. Endpoints included RR by RECIST at six or eight weeks, mean log ratio of tumor size at six or eight weeks relative to baseline (log ratio) [13], PFS rate at 90 days (PFS-90), and overall PFS. Arms were compared using a chi-squared test for RR by RECIST, a two sample t-test for log ratio, a chi-squared test for PFS-90, and a log-rank test for overall PFS. Trials were positive if patients on the treatment arm did significantly better than patients on the placebo arm (one-sided $p<0.10$). Trials were stopped early (i.e., halfway) for all endpoints (except overall PFS) if the treatment arm was performing worse than the placebo arm, a rule associated with a very small loss in power.[20]

Results

For patients included in the analyses from TARGET, PFS was significantly longer in the sorafenib arm ($p < 0.0001$ by log-rank test, Supplemental Figure 1a), consistent with the results of TARGET.[15] For patients included in the analyses from the AE 941 trial, there was no significant difference in PFS between the two arms ($p = 0.68$ by log-rank test, Supplemental Figure 1b). Tables 1a and 1b summarize the results for patients included in the analyses from TARGET and the AE 941 trial, respectively, based on the actual trial data.

Table 2a shows the results of resampling simulations for single arm and randomized designs using data from TARGET. For a single arm design with RR by RECIST, 55.2% (95% CI 53.8%–56.6%) of simulated trials were positive. Similarly, for a randomized design with RR by RECIST and 20 patients per arm, 55.0% (95% CI 53.6%–56.4%) of simulated trials showed a significant benefit for sorafenib compared with placebo. Using log ratio increased the percentage of positive trials to 87.7% (95% CI 86.8%–88.6%). Randomized designs with PFS-90 and PFS resulted in a percentage of positive trials that was higher than RR by RECIST but lower than log ratio. Increasing the sample size in randomized designs to 25, 30 and 35 patients per arm gradually increased the percentage of positive trials for all endpoints, but in all cases log ratio outperformed the other endpoints.

Table 2b shows the results of resampling simulations for single arm and randomized designs using data from the AE941 trial. For a single arm design with RR by RECIST, 0.9% (95% CI 0.6%–1.2%) of simulated trials were positive. For a randomized design with RR by RECIST and 20 patients per arm, 6.9% (95% CI 6.1%–7.6%) of simulated trials showed a significant benefit for AE 941 compared with placebo. Using log ratio increased the percentage of positive trials to 24.7% (95% CI 23.5%–26.0%). Again, randomized designs with PFS-90 and PFS resulted in a percentage of positive trials that was higher than RR by RECIST but lower than log ratio.

Discussion

Our study demonstrates the potential utility of a database of phase III trials for subsequent investigation, particularly for clinical trial simulations. We compared phase II designs and endpoints in renal cancer, including an early endpoint obtained after the first CT scan at six or eight weeks and recorded on a continuous scale. Early endpoints were emphasized because of their value for making early decisions about drug efficacy, and the continuous scale was incorporated because it increases statistical efficiency. Since we resampled data from a phase III trial that demonstrated the efficacy of sorafenib in renal cancer, the best designs and endpoints were those that most frequently showed that sorafenib is active in renal cancer. We also resampled data from a phase III trial that demonstrated no efficacy of AE 941 in renal cancer, in order to evaluate the false positive rates with these designs and endpoints. The large number of simulated trials allowed precise estimates of the frequency of a positive result. The stipulated one-sided alpha level of 10% is consistent with the more exploratory nature of a phase II study.

The results of our study show that a randomized phase II design with log ratio is preferable to more conventional designs and endpoints for predicting the phase III results of sorafenib in renal cancer. We did not attempt to use endpoints such as log ratio or PFS-90 in a single arm design, as it would be difficult to specify a null hypothesis from historical data and validity would be compromised by potential selection bias. According to expert consensus guidelines, single arm phase II monotherapy trials should generally use RR by RECIST as the endpoint.[18] For sorafenib, this conventional design with 37 patients has a false negative rate of 45% (55% of simulated trials were positive). Furthermore, a positive single arm trial means that the null hypothesis (which could correspond to a low RR) can be rejected, but

does not necessarily establish that the RR is high.[21] Although slightly better with a comparable number of patients, the conventional randomized phase II design (with PFS) with 20 patients per arm still has a false negative rate of 31% for sorafenib in renal cancer. In contrast, the randomized design with log ratio and the same number of patients has a false negative rate of only 12%, and even lower false negative rates with larger sample sizes. In order to achieve a certain threshold of power (e.g., 80% or 90%) with PFS, one would need a larger sample size and longer trial duration than would be required for a similar trial with log ratio. Assuming that all patients have measurable disease, no additional data are required for calculating log ratio compared with RR by RECIST or PFS.

The randomized design with log ratio, as demonstrated by the AE 941 results, had false positive rates ranging between 25% and 30% (increasing with sample size), which are higher than the rates observed for other designs/endpoints and higher than the 10% Type I error rate that would be expected if the drug were truly equivalent to placebo. Patients in the AE 941 arm had slightly less tumor growth at eight weeks compared with those in the placebo arm (Table 1b), suggesting that AE 941 had a small growth inhibitory effect without a corresponding PFS benefit (i.e., this was not a perfectly negative trial). This suggests that AE 941 has some activity in metastatic renal cancer, further supported by the observation of two objective responses in the original phase II trial.[22] Thus, AE 941 may not be a true negative control.

Randomized trials that were simulated by resampling twice from the placebo arm of either trial had the expected false positive rate of less than 10% (data not shown). Log ratio is not a perfect surrogate for PFS and, like any surrogate endpoint, may increase the risk of false positives.[23] For drugs with large treatment effects (e.g., RR by RECIST > 20%), it is likely that any of the designs evaluated will be positive. However, for active drugs with smaller treatment effects (e.g., sorafenib), a randomized design is most appropriate.

There are a number of limitations of our study. First, we excluded a minority (~15%) of patients on each trial from our analyses, primarily due to incomplete tumor size data. We addressed this limitation by showing that PFS of included patients was consistent with published results of the trials. Second, we used a small number of lesions to assess tumor burden, a limitation that was necessary to restrict our analyses to consistently measured lesions. Third, our method of imputing tumor size data for patients who had a PFS event before the first CT scan (more of whom were in the placebo arms) might have exaggerated both drug effects, but excluding these patients would have done the opposite. Fourth, the potential use of log ratio is only applicable to disease settings with measurable disease. Fifth, our negative control (AE 941) may actually have some activity. Finally, the results regarding log ratio apply only to early assessment (at six to eight weeks) during treatment with a growth inhibitory drug for renal cancer, a disease that is known to have a heterogeneous prognosis. These results might not be generalizable to drugs that have different mechanisms of action, or to other disease settings.

In contrast to our results, Fridlyand and colleagues concluded that PFS was a superior phase II endpoint to percentage change in tumor burden by resampling data from six phase III trials in colorectal cancer, breast cancer and non-small cell lung cancer.[24] These results emphasize that further analyses of the continuous log ratio endpoint are necessary before it can be broadly accepted. However, log ratio may be useful as a primary endpoint in settings where an early decision is required or where early crossover is preferred on ethical grounds.

Also in contrast to our results, An and colleagues concluded that continuous tumor measurement-based metrics were no better than categorical metrics (complete/partial response vs. stable disease vs. progressive disease) using data from three phase III trials of

cytotoxic therapy in colorectal cancer and non-small cell lung cancer.[25] However, the efficiency of continuous measurement-based metrics depends on the quality of the tumor measurement data. It is common in large multicenter trials that are not audited by industry sponsors not to collect full tumor measurement data as prescribed by RECIST. It would be expected that, in studies where the average number of lesions assessed was small, the advantage of continuous assessments over categorical assessments would be diminished. To resolve the contradictions in the literature will require assessment of these thresholds in future analyses on a disease-by-disease basis. In addition, An and colleagues[25] did not address power but rather the ability of an early outcome measure to predict subsequent survival using a landmark analysis approach, with predictive ability assessed by the concordance index. The inclusion of patients who progressed or died before the first CT scan in our study is also an important methodologic difference.

In conclusion, this study supports the potential use of an alternative endpoint for randomized phase II trials of growth inhibitory drugs, the log ratio of tumor size at six or eight weeks compared to baseline. This design/endpoint is better than more conventional designs/endpoints for detecting the efficacy of sorafenib, but may increase the risk of false positives with drugs that eventually prove to lack clinical benefit. Additional empirical evidence is needed before definitive conclusions can be reached about whether log ratio is an endpoint that should be utilized in phase II trials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Bayer Pharmaceuticals for providing the data from TARGET that was used in this research. We would like to thank Dr. Bernard Escudier and Dr. Daniel Croteau for providing the data from the AE 941 trial that was used in this research.

Grant support

This work was supported by NIH grant T32GM007019 and a Conquer Cancer Foundation Translational Research Professorship award to MJR; National Cancer Institute Mentored Career Development Award K23CA124802 to MLM. The funders did not have any involvement in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication.

References

1. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009; 45:228–47. [PubMed: 19097774]
2. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989; 10:1–10. [PubMed: 2702835]
3. Sharma MR, Stadler WM, Ratain MJ. Randomized phase II trials: a long-term investment with promising returns. J Natl Cancer Inst. 2011; 103:1093–100. [PubMed: 21709274]
4. Tang H, Foster NR, Grothey A, Ansell SM, Goldberg RM, Sargent DJ. Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. J Clin Oncol. 2010; 28:1936–41. [PubMed: 20212253]
5. Gan HK, Grothey A, Pond GR, Moore MJ, Siu LL, Sargent D. Randomized phase II trials: inevitable or inadvisable? J Clin Oncol. 2010; 28:2641–47. [PubMed: 20406933]
6. Cannistra SA. Phase II trials in journal of clinical oncology. J Clin Oncol. 2009; 27:3073–76. [PubMed: 19451415]

7. Michaelis LC, Ratain MJ. Phase II trials published in 2002: a cross-specialty comparison showing significant design differences between oncology trials and other medical specialties. Clin Cancer Res. 2007; 13:2400–05. [PubMed: 17438099]
8. DiMasi JA, Grabowski HG. Economics of new oncology drug development. J Clin Oncol. 2007; 25:209–16. [PubMed: 17210942]
9. Maitland ML, Hudoba C, Snider KL, Ratain MJ. Analysis of the yield of phase II combination therapy trials in medical oncology. Clin Cancer Res. 2010; 16:5296–302. [PubMed: 20837695]
10. Ratain MJ, Eckhardt SG. Phase II studies of modern drugs directed against new targets: if you are fazed, too, then resist RECIST. J Clin Oncol. 2004; 22:4442–45. [PubMed: 15483011]
11. El-Maraghi RH, Eisenhauer EA. Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. J Clin Oncol. 2008; 26:1346–54. [PubMed: 18285606]
12. Lavin PT. An alternative model for the evaluation of antitumor activity. Cancer Clin Trials. 1981; 4:451–57. [PubMed: 7318127]
13. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non small-cell lung cancer. J Natl Cancer Inst. 2007; 99:1455–61. [PubMed: 17895472]
14. Wang Y, Sung C, Dartois C, Ramchandani R, Booth BP, Rock E, et al. Elucidation of relationship between tumor size and survival in non-small-cell lung cancer patients can aid early decision making in clinical drug development. Clin Pharmacol Ther. 2009; 86:167–74. [PubMed: 19440187]
15. Claret L, Girard P, Hoff PM, Van Cutsem E, Zuideveld KP, Jorga K, et al. Model-based prediction of phase III overall survival in colorectal cancer on the basis of phase II tumor dynamics. J Clin Oncol. 2009; 27:4103–08. [PubMed: 19636014]
16. Escudier B, Eisen T, Stadler WM, Szczylik C, Oudard S, Siebels M, et al. TARGET Study Group. Sorafenib in advanced clear-cell renal-cell carcinoma. N Engl J Med. 2007; 356:125–34. [PubMed: 17215530]
17. Escudier B, Choueiri TK, Oudard S, Szczylik C, Négrier S, Ravaud A, et al. Prognostic factors of metastatic renal cell carcinoma after failure of immunotherapy: new paradigm from a large phase III trial with shark cartilage extract AE 941. J Urol. 2007; 178:1901–05. [PubMed: 17868728]
18. Seymour L, Ivy SP, Sargent D, Spriggs D, Baker L, Rubinstein L, et al. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. Clin Cancer Res. 2010; 16:1764–918. [PubMed: 20215557]
19. Sun M, Shariat SF, Cheng C, Ficarra V, Murai M, Oudard S, et al. Prognostic factors and predictive models in renal cell carcinoma: a contemporary review. Eur Urol. 2011; 60:644–61. [PubMed: 21741163]
20. Wieand S, Therneau T. A two-stage design for randomized trials with binary outcomes. Control Clin Trials. 1987; 8:20–28. [PubMed: 3568693]
21. Ratain MJ, Karrison TG. Testing the wrong hypothesis in phase II oncology trials: there is a better alternative. Clin Cancer Res. 2007; 13:781–82. [PubMed: 17289865]
22. Batist G, Patenaude F, Champagne P, Croteau D, Levinton C, Hariton C, et al. Neovastat (AE-941) in refractory renal cell carcinoma patients: report of a phase II trial with two dose levels. Ann Oncol. 2002; 13:1259–63. [PubMed: 12181250]
23. Rubinstein LV, Dancey JE, Korn EL, Smith MA, Wright JJ. Early average change in tumor size in a phase 2 trial: Efficient endpoint or false promise? J Natl Cancer Inst. 2007; 99:1422–23. [PubMed: 17895470]
24. Fridlyand J, Kaiser LD, Fyfe G. Analysis of tumor burden versus progression-free survival for Phase II decision making. Contemp Clin Trials. 2011; 32:446–52. [PubMed: 21266203]
25. An MW, Mandrekar SJ, Branda ME, Hillman SL, Adjei AA, Pitot HC, et al. Comparison of continuous versus categorical tumor measurement-based metrics to predict overall survival in cancer treatment trials. Clin Cancer Res. 2011; 17:6592–9. [PubMed: 21880789]

Statement of Translational Relevance

The high failure rate of phase III oncology trials emphasizes that current phase II trials are not sufficiently informative regarding the efficacy, or lack thereof, of new experimental treatments. This is especially the case for growth inhibitory (cytostatic) drugs, for which conventional designs and endpoints suitable for cytotoxic agents (e.g., single arm design with response rate by RECIST) may not accurately predict eventual phase III results. In this study, we apply a resampling methodology to two completed trials (one positive, one negative) in metastatic renal cancer in order to simulate and compare various phase II designs and endpoints. We found that a randomized design with an early continuous change in tumor size endpoint was the best predictor of the positive result, although it had a higher false positive rate for the negative result. If validated in future studies, this design/endpoint has the potential to make phase II trials of growth inhibitory drugs more informative with relatively short follow-up and feasible sample sizes.

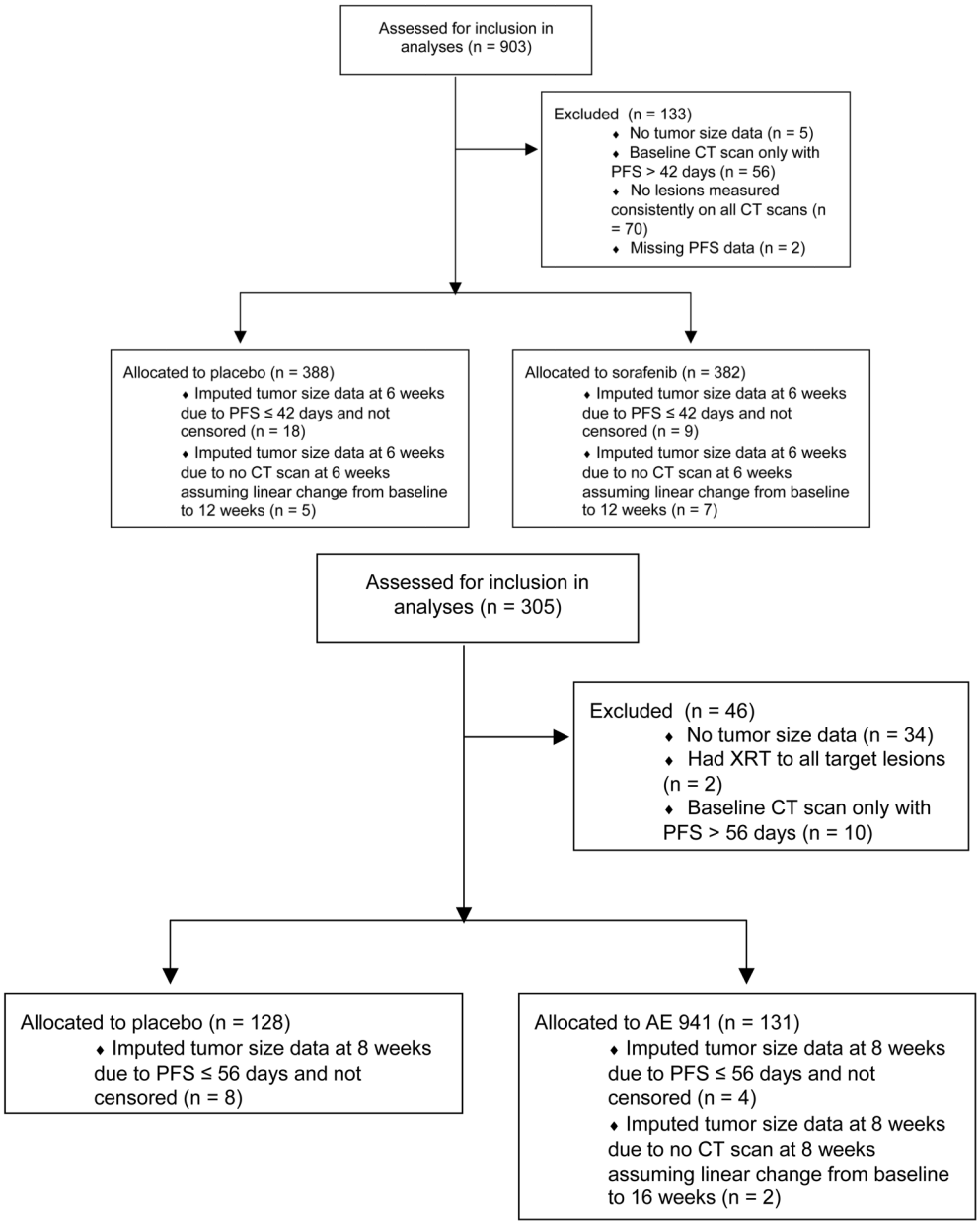

Figure 1.

Figure 1a. Flow diagram for TARGET data, indicating the reasons for exclusion from the analyses and the numbers of patients with imputed data.

Figure 1b. Flow diagram for AE 941 trial data, indicating the reasons for exclusion from the analyses and thenumbers of patients with imputed data.

Table 1a

Summary of results from the patients included in the analyses from the phase III trial of sorafenib vs. placebo (TARGET).

	Sorafenib arm (n = 382)	**Placebo arm (n = 388)**
Mean ± SD number of target lesions used	2.1 ± 0.9	2.1 ± 0.9
Mean ± SD baseline tumor size (mm)	78.0 ± 55.1	79.6 ± 52.1
RR by RECIST at 6 weeks	10.5 %	1.0%
Mean ± SD % change in tumor size at 6 weeks vs. baseline	−5.9% ± 28.6%	15.3% ± 32.6%
Mean log ratio ± SD of tumor size at 6 weeks vs. baseline	−0.10 ± 0.29	0.11 ± 0.25

Table 1b

Summary of results from the patients included in the analyses from the phase III trial of AE 941 vs. placebo.

	AE 941 arm (n = 131)	Placebo arm (n = 128)
Mean ± SD number of target lesions used	2.3 ± 0.8	2.3 ± 0.8
Mean ± SD baseline tumor size (mm)	97.1 ± 57.6	104.1 ± 58.9
RR by RECIST at 8 weeks	2.3 %	0.8%
Mean ± SD % change in tumor size at 8 weeks vs. baseline	13.0% ± 27.9%	16.2% ± 22.1%
Mean log ratio ± SD of tumor size at 8 weeks vs. baseline	0.10 ± 0.22	0.13 ± 0.18

Table 2a

Results of the resampling simulations for single arm (sorafenib only) and randomized (sorafenib vs. placebo) phase II designs with various endpoints. PFS-90: PFS rate at 90 days.

Design	Endpoint	% positive (95%CI)	% stop early
Single arm, n = 37	RR by RECIST at 6 weeks	55.2 (53.8–56.6)	26.3
Randomized, n = 20/arm	RR by RECIST at 6 weeks	55.0 (53.6–56.4)	3.3
	log ratio at 6 weeks	87.7 (86.8–88.6)	3.9
	PFS-90	63.5 (62.2–64.9)	11.1
	PFS	68.5 (67.2–69.8)	--
Randomized, n = 25/arm	RR by RECIST at 6 weeks	63.3 (61.9–64.6)	3.1
	log ratio at 6 weeks	91.9 (91.1–92.6)	2.7
	PFS-90	70.5 (69.2–71.8)	9.8
	PFS	74.9 (73.7–76.1)	--
Randomized, n = 30/arm	RR by RECIST at 6 weeks	70.8 (69.5–72.1)	2.9
	log ratio at 6 weeks	95.1 (94.5–95.7)	1.2
	PFS-90	76.2 (75.0–77.4)	7.2
	PFS	82.6 (81.5–83.7)	--
Randomized, n = 35/arm	RR by RECIST at 6 weeks	76.0 (74.8–77.2)	3.0
	log ratio at 6 weeks	96.5 (96.0–97.0)	1.1
	PFS-90	79.0 (77.8–80.2)	6.8
	PFS	85.2 (84.2–86.2)	--

Table 2b

Results of the resampling simulations for single arm (AE941 only) and randomized (AE941 vs. placebo) phase II designs with various endpoints. PFS-90: PFS rate at 90 days.

Design	Endpoint	% positive (95%CI)	% stop early
Single arm, n = 37	RR by RECIST at 6 weeks	0.9 (0.6–1.2)	75.1
Randomized, n = 20/arm	RR by RECIST at 6 weeks	6.9 (6.1–7.6)	6.2
	log ratio at 6 weeks	24.7 (23.5–26.0)	33.7
	PFS-90	15.2 (14.4–16.3)	33.7
	PFS	9.3 (8.5–10.1)	--
Randomized, n = 25/arm	RR by RECIST at 6 weeks	9.5 (8.7–10.4)	7.9
	log ratio at 6 weeks	25.1 (23.9–26.3)	32.2
	PFS-90	16.9 (15.9–18.0)	36.4
	PFS	8.0 (7.3–8.8)	--
Randomized, n = 30/arm	RR by RECIST at 6 weeks	11.3 (10.4–12.2)	7.6
	log ratio at 6 weeks	27.1 (25.9–28.4)	31.2
	PFS-90	17.6 (16.5–18.6)	33.7
	PFS	8.2 (7.4–9.0)	--
Randomized, n = 35/arm	RR by RECIST at 6 weeks	14.2 (13.2–15.2)	8.3
	log ratio at 6 weeks	29.7 (28.4–31.0)	28.4
	PFS-90	19.6 (18.5–20.7)	35.4
	PFS	7.6 (6.8–8.3)	--