Statistics in Clinical Cancer Research

Clinical Cancer Research

Borrowing Information across Subgroups in Phase II Trials: Is It Useful?

Boris Freidlin and Edward L. Korn

Abstract

Because of the heterogeneity of human tumors, cancer patient populations are usually composed of multiple subgroups with different molecular and/or histologic characteristics. In screening new anticancer agents, there might be a scientific rationale to expect some degree of similarity in clinical activity across the subgroups. This poses a challenge to the design of phase II trials assessing clinical activity: Conducting an independent evaluation in each subgroup requires considerable time and resources, whereas a pooled evaluation that completely ignores patient heterogeneity can miss treatments that are only active in some subgroups. It has been suggested that approaches that borrow information across subgroups can improve efficiency in this setting. In particular, the hierarchical Bayesian approach putatively uses the outcome data to decide whether borrowing of information is appropriate. We evaluated potential benefits of the hierarchical Bayesian approach (using models suggested previously) and a simpler pooling approach by simulations. In the phase II setting, the hierarchical Bayesian approach is shown not to work well in the simulations considered, as there appears to be insufficient information in the outcome data to determine whether borrowing across subgroups is appropriate. When there is strong rationale for expecting a uniform level of activity across the subgroups, approaches using simple pooling of information across subgroups may be useful. *Clin Cancer Res; 19(6); 1326–34. ©2012 AACR.*

Introduction

One of the major challenges in the development of anticancer agents is the heterogeneity of patient populations. In early clinical studies assessing activity of new treatments (phase II trials), patients can often be classified into nonoverlapping subgroups for which it may be reasonable to assume that the activity (or lack thereof) is homogeneous within each subgroup, and for which there is some reason to believe that the activity level may be similar across subgroups. Three scenarios where such subgroups arise are (i) patients expressing a molecular target of interest from multiple cancer histologies (1), (ii) multiple histologic subtypes of a given cancer histology, and (iii) biomarker-defined molecular subtypes of a given cancer histology. An example of the first scenario is a trial (NCT01306045) that evaluates 5 agents with distinct biomarker targets in which each agent is evaluated in patients expressing the corresponding biomarker in 3 different histologic subgroups [non–small cell lung cancer (NSCLC), small cell lung cancer, and thymic cancer] separately. An example of the second scenario is a trial (2) that evaluated imatinib in 10 histologic subtypes of soft-tissue sarcoma. An example of the third scenario is the BATTLE trial (3) that evaluated therapies in 5 biomarker-defined subgroups of NSCLC.

The 2 common approaches to analyzing activity with subgroups are to ignore them (and do one pooled analysis) or to conduct a separate stand-alone analysis in each subgroup. Either approach can be problematic. A pooled analysis can miss agents that are only active in one or a few subgroups. On the other hand, conducting an independent evaluation in each subgroup is time/resource-consuming and is often not feasible because of the large total sample size that would be required.

Rather than stand-alone subgroup analyses or a single pooled analysis, an attractive middle ground would share the outcome results from the different subgroups to improve the inference for each subgroup. This is sometimes referred to as "borrowing information" or "borrowing strength" across the subgroups (4–6). For example, in Table 1, by borrowing information from subgroups 1 to 4, one might find the 30% response rate in subgroup 5 more believable in scenario 1 than in scenario 2. Formal statistical borrowing of information is often done via Bayesian methods. For example, when some preliminary information on the overall response rate is available, a simple Bayesian model summarizes the preliminary data in a prior distribution for the true subgroup response rates. The inference for each subgroup is based on the posterior distribution for its response rate, which uses the preliminary information by shrinking the observed response rates toward the

Authors' Affiliation: Biometric Research Branch, Cancer Therapy Evaluation Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland

Corresponding Author: Boris Freidlin, Biometric Research Branch, EPN-8122, National Cancer Institute, Bethesda, MD 20892. Phone: 301-402-0640; Fax: 301-402-0560; E-mail: freidlinb@ctep.nci.nih.gov

doi: 10.1158/1078-0432.CCR-12-1223

Table 1. Two hypothetical 5-subgroup scenarios in which the 30% response rate in subgroup 5 is more believable in scenario 1 than in scenario 2

	Scenario 1		Scenario 2	
Subgroup	Patients, *n*	Responses (%)	Patients, *n*	Responses (%)
1	25	8 (32)	25	1 (4)
2	25	6 (24)	25	0 (0)
3	25	7 (28)	25	2 (8)
4	25	9 (36)	25	1 (4)
5	10	3 (30)	10	3 (30)

mean of the prior distribution, with more shrinking if the subgroup sample size is small. The amount of shrinking also depends on the spread of the prior distribution around its mean, with more shrinking if the spread is narrow.

The simple Bayesian approach does not allow borrowing information across the subgroups. Furthermore, the amount of shrinking to the mean of the prior distribution does not depend on the observed outcomes, as it is fixed by the spread of the prior distribution and the subgroup sample size. To overcome these deficiencies, a hierarchical Bayesian approach has been suggested (7–9). This approach (10) uses the observed response rates to help to decide whether to borrow the information (by shrinking to roughly the mean of all the observed response rates) and how much to borrow (shrink less if the observed response rates are far apart). In theory, the idea of adapting the degree of borrowing across the subgroups according to the observed data, "outcome-adaptive borrowing," sounds attractive. It has been suggested to "make for more informed decision making and smaller clinical trials" (10, page 35) and to be "an effective method for studying rare diseases and their subtypes" (2, page 3148). However, the degree to which this approach works warrants careful examination. We evaluate the potential benefits of hierarchical Bayesian modeling via computer simulations. We also evaluate a much simpler approach for borrowing information across subgroups in futility/inefficacy interim monitoring (11).

Outcome-adaptive borrowing: hierarchical Bayesian approaches

We assume that activity is measured by a binary outcome (e.g., response). Let p_i be the observed response rate based on n_i patients in the ith subgroup ($i = 1, \ldots, K$). The true response rates in the subgroups, $\pi_1, \pi_2, \ldots, \pi_K$, (or a specified transformation of them) are assumed to come from a specified prior distribution with unknown mean and variance. One could use the observed response rates to estimate this unknown mean and variance and subsequently estimate the π_i; this would be what is known as an empirical Bayes approach (12, chapter 3). Alternatively, a hierarchical Bayesian approach treats the unknown mean and variance as random quantities and specifies distributions for them (9). There are many model specifications for the hierarchical Bayesian approach. We used 2 models specifically developed for phase II clinical trial setting. The first parameterization is from Thall and colleagues (9). For this model (model 1), 2 different sets of hyperpriors were considered, allowing for moderate and strong borrowing, respectively. The second parameterization (model 2) is from Berry and colleagues (ref. 13; see Appendix for details).

Combining pooled and subgroup-specific analyses

LeBlanc and colleagues (11) suggested a simple approach to borrowing information: A futility analysis based on the pooled data of all the subgroups is conducted in addition to futility analyses conducted for each subgroup. The futility analysis on the pooled data is conducted after a specified number of patients (n_{pooled}) are enrolled in the study overall; a test is conducted to see whether the pooled response rate (p_{pooled}) is consistent with an alternative hypothesis, and, if not, the whole trial is stopped. The separate futility analyses are conducted in each subgroup, after a specified number of patients (n_s) are enrolled in that subgroup; a test is conducted in each subgroup to see whether the subgroup response rate is consistent with an alternative hypothesis, and, if not, accrual to the individual subgroup is stopped. The alternative hypothesis for the pooled analysis would typically be taken to be less than the alternative hypotheses for the individual subgroups. For example, LeBlanc and colleagues suggest that for testing an alternative hypothesis of a response rate of 30% (vs. a null response rate of 10%) at significance level α_{0s} (e.g., 0.02), in the individual subgroups, and an alternative hypothesis of 20% (vs. a null response rate of 10%) at significance level α_0 (e.g., 0.02) be used for the pooled futility analysis. That is, accrual to an individual subgroup, i is stopped if $T_i^A < \alpha_{os}$, and accrual to the entire study is stopped if $T^A < a_o$, where $T_i^A = \Pr(X \leq p_i n_s | n_s, \pi = 0.3)$ and $T^A = \Pr(X \leq p_{pooled} n_{pooled} | n_{pooled}, \pi = 0.2)$ are binomial tail probabilities.

Simulations

Following the studies of Thall and colleagues (9) and LeBlanc and colleagues (11), a response rate of 30% was considered promising in a subgroup whereas a response rate of 10% was considered uninteresting. First, we considered a fixed sample size design in which 25 patients are accrued to each subgroup (i.e., no interim monitoring): For each patient in each subgroup, response status was independently generated to be 1 with probability equal to the given response rate, and 0, otherwise. In practice, however, many phase II designs in oncology include interim futility monitoring rules that allow the trial to stop early for disappointing results (to protect patients and resources). Thus, we also simulated designs with interim futility analyses. These later simulations allowed us to assess the benefits of borrowing information across treatment arms when (i) accumulating data

are repeatedly evaluated over time and (ii) the accrual rates differ between the subgroups. In these simulations, patient outcomes were generated sequentially with the subgroup status of each patient generated from a multinomial distribution with prespecified subgroup frequencies. The simulation code was written in R (14), and the hierarchical Bayesian approach was implemented using the WINBUGS package (15).

Fixed sample size (no interim monitoring)

We investigated settings with 5 and 10 subgroups. For an individual evaluation in each subgroup (no borrowing), a design that enrolls 25 patients in each subgroup and rejects the null hypothesis if there are 5 or more responses allows one to distinguish 30% versus 10% response rates with a false-positive rate of 0.1 and a power of 90%. (This design provides so-called "strong" control of marginal false-positive error rates. That is, the false-positive error rate in each subgroup is no greater than 0.1 under the hypothesis that its true response rate is 10% regardless of the response rates of the other subgroups.) In a corresponding hierarchical Bayesian design (with 25 patients in each group), the null hypothesis for the *i*th subgroup is rejected if the posterior probability that $\pi_i > 10\%$ is greater than a certain cutoff value. To facilitate a fair comparison of the 2 analysis strategies, the cutoff value needs to be chosen to provide the same strong control of false-positive error rates (at 0.1 level) as with the subgroup-specific evaluation design (16). To accomplish this for the Thall and colleagues model 1 with moderate borrowing, a cutoff of 0.850 was used in both settings. For the Thall and colleagues model 1 with strong borrowing, cutoffs of .940 and .965 were used for 5- and 10-subgroup settings, respectively (9). For the Berry and colleagues model 2, cutoffs of 0.955 and 0.960 were used for 5- and 10-subgroup settings, respectively (13).

Tables 2 and 3 display the empirical subgroup probabilities of a positive conclusion under a range of scenarios with 5 and 10 subgroups, respectively. In the subgroup-specific analyses, the probabilities of a positive conclusion in a given subgroup are approximately 0.9 (0.1) when that subgroup response rate is 30% (10%), regardless of the response rates in other subgroups. The hierarchical Bayesian strategies are shown to have no better power than the subgroup-specific strategy: The hierarchical Bayesian design model 1 with moderate borrowing takes a conservative approach to borrowing and conducts almost identically to the subgroup-specific approach. The hierarchical Bayesian designs that allow more borrowing (model 1 with strong borrowing and model 2) sometimes have nontrivially lower power than the subgroup-specific approach in cases in which the treatment is active only in a minority of subgroups (cases 3–4 in Tables 2 and 3). Note that under the global null (10% response in all subgroups) model 1 with strong borrowing and model 2 have false-positive error rates that are well below the nominal 0.1 level (case 5 in Tables 2 and 3). This is the price of borrowing across subgroups in a design with strong control of the error rates. Although it might seem tempting to forgo the strong error control by making it easier to reject null hypotheses, this would result in a design with unacceptably high false-positive error rates in some scenarios. For example, for model 2 in Table 3, lowering the bar for rejection so that the case 5 results were 0.1 would increase the false-positive error rate to 0.40 in case 1 (for the first subgroup). We revisit the motivations for requiring strong control of the error rates in the Discussion.

With interim inefficacy/futility monitoring

We assumed a setting with 5 subgroups. For the subgroup-specific analyses approach, we used 2-stage designs for the inefficacy/futility monitoring: After 15 patients are accrued to a subgroup, accrual is stopped to that subgroup if there is one or fewer responses. Otherwise, an additional 10 patients are accrued to that subgroup with the null hypothesis rejected with 5 or more responses. For each subgroup, the false-positive rate and power are 0.1 and 90%, respectively.

For the approach of LeBlanc and colleagues (11), in addition to the first-stage individual subgroup inefficacy/futility monitoring described above (with 15 patients), we pool the data from all the subgroups and test whether the data are consistent with an overall response rate of 20%. If the hypothesis can be rejected at the 0.02 level, accrual to the whole trial is stopped and the treatment is considered inactive for all subgroups. The pooled analyses were conducted when 40 and 80 patients had been accrued on the study overall.

For the hierarchical Bayesian approach, we followed the parameterization and the stopping rule proposed by Thall and colleagues (9), with stopping accrual to the subgroup *i* if the posterior probability that $\pi_i > 30\%$ was <0.005. That probability was calculated for each subgroup when 40 and 80 patients had been accrued to the study overall. Subgroups that are not stopped continue to a sample size of 25. When accrual to the whole trial is over, the posterior probabilities that $\pi_i > 10\%$ are calculated for the remaining subgroups, with the null hypothesis rejected for those subgroups for which this probability is greater than the cutoff of values 0.85 and 0.94 to control the false-positive error at 0.1, for moderate and strong borrowing models, respectively.

Two accrual settings were considered: equal accrual rates (Table 4) and unequal accrual rates (30%, 20%, 20%, 20%, 10%) and relative accrual rates in groups 1 to 5, respectively (Table 5). The ability of a design to stop accrual early to the subgroups with disappointing results is measured by the average sample size. The hierarchical Bayesian and subgroup-specific evaluation approaches conduct similarly in this respect and allow an approximately 20% reduction in accrual in inactive subgroups due to futility stopping. The LeBlanc approach (11) allows for an additional reduction in sample size when most of the subgroups are inactive (cases 1 and 2) albeit at the price of reduced power (case 2). The loss of power can be substantial if the treatment is only active in the subset with a low accrual rate (case 2 of Table 5).

Table 2. Empirical probabilities of rejecting the null hypothesis: 5 subgroups (no interim monitoring, 25 patients per subgroup, 10,000 replications)

Design	True response rate in each subgroup				
Case 1	**0.1**	**0.3**	**0.3**	**0.3**	**0.3**
Subgroup-specific analyses	0.096	0.909	0.912	0.910	0.914
HB model 1 moderate borrowing	0.096	0.909	0.912	0.910	0.914
HB model 1 strong borrowing	0.098	0.895	0.893	0.892	0.891
HB model 2 (Berry et al.; ref. 13)	0.096	0.899	0.898	0.896	0.899
Case 2	**0.1**	**0.1**	**0.3**	**0.3**	**0.3**
Subgroup-specific analyses	0.096	0.096	0.912	0.910	0.914
HB model 1 moderate borrowing	0.096	0.096	0.912	0.910	0.914
HB model 1 strong borrowing	0.085	0.096	0.842	0.842	0.842
HB model 2 (13)	0.091	0.091	0.853	0.855	0.858
Case 3	**0.1**	**0.1**	**0.1**	**0.3**	**0.3**
Subgroup-specific analyses	0.096	0.096	0.097	0.910	0.914
HB model 1 moderate borrowing	0.096	0.096	0.097	0.910	0.914
HB model 1 strong borrowing	0.058	0.061	0.058	0.807	0.809
HB model 2 (13)	0.067	0.065	0.062	0.817	0.820
Case 4	**0.1**	**0.1**	**0.1**	**0.1**	**0.3**
Subgroup-specific analyses	0.096	0.096	0.097	0.096	0.914
HB model 1 moderate borrowing	0.096	0.096	0.097	0.096	0.914
HB model 1 strong borrowing	0.037	0.040	0.038	0.038	0.762
HB model 2 (13)	0.041	0.041	0.036	0.043	0.791
Case 5	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
Subgroup-specific analyses	0.096	0.096	0.097	0.096	0.099
HB model 1 moderate borrowing	0.096	0.096	0.097	0.096	0.099
HB model 1 strong borrowing	0.025	0.030	0.029	0.030	0.025
HB model 2 (13)	0.032	0.033	0.030	0.030	0.033
Case 6	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**
Subgroup-specific analyses	0.912	0.909	0.912	0.910	0.914
HB model 1 moderate borrowing	0.912	0.909	0.912	0.910	0.914
HB model 1 strong borrowing	0.907	0.910	0.907	0.911	0.911
HB model 2 (13)	0.911	0.908	0.912	0.910	0.913

Abbreviation: HB, hierarchical Bayesian.

Discussion

Theoretically, the use of outcome-adaptive borrowing across subgroups is attractive—one could increase the power to find effective treatments in subgroups by borrowing the information from other subgroups only when the data suggest it is reasonable to do so. Unfortunately, our experience with the models proposed for phase II clinical trials by the experts in the hierarchical methodology suggests that this approach does not work for identifying responsive subgroups in the phase II setting with 10 or fewer subgroups. The essential problem is that there is not enough information in the data to determine whether borrowing is appropriate. Therefore, results obtained are sensitive to the details of the hierarchical model specified for the data. With the same observed response rates but with different model parameterizations, it is possible to make the borrowing across subgroups either extremely difficult or extremely easy. For example, using the Thall and colleagues model (9), increasing the mean of the hyperprior for the precision parameter increases the amount of borrowing: Table 6 considers borrowing under the 2 scenarios described in Table 1. For scenario 1, the posterior probability of the response rate in group 5 being greater than 30% is approximately the same regardless of the amount of borrowing. On the other hand, for scenario 2, the posterior probability of the response rate in group 5 being greater than 30% decreases considerably with stronger borrowing. Thus, borrowing makes the observed response rate in subgroup 5 more believable in scenario 1 than in scenario 2. No parameterization, however, is able to use the data to accurately determine whether borrowing is appropriate or not.

It should be noted that we conducted our evaluations under a strong control of the false-positive error rates to avoid unacceptably high false-positive error rates for some subgroups. Our rationale for use of the strong error control framework is as follows: The purpose of conducting phase II trials is to minimize the number of (large) negative phase III trials by screening out clinical settings in which agents do

Table 3. Empirical probabilities of rejecting the null hypothesis: 10 subgroups (no interim monitoring, 25 patients per subgroup, 10,000 replications)

Design	True response rate in each subgroup									
Case 1	**0.1**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**
Subgroup-specific analyses	0.096	0.905	0.908	0.907	0.909	0.908	0.910	0.908	0.909	0.912
HB model 1 moderate borrowing	0.096	0.905	0.908	0.907	0.909	0.908	0.910	0.908	0.909	0.912
HB model 1 strong borrowing	0.099	0.892	0.892	0.894	0.897	0.896	0.903	0.895	0.895	0.898
HB model 2 (Berry et al.; ref. 13)	0.099	0.905	0.908	0.907	0.908	0.912	0.910	0.908	0.909	0.911
Case 2	**0.1**	**0.1**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**
Subgroup-specific analyses	0.096	0.098	0.908	0.907	0.909	0.908	0.910	0.908	0.909	0.912
HB model 1 moderate borrowing	0.096	0.098	0.908	0.907	0.909	0.908	0.910	0.908	0.909	0.912
HB model 1 strong borrowing	0.085	0.087	0.857	0.857	0.861	0.858	0.867	0.859	0.857	0.861
HB model 2 (13)	0.097	0.098	0.904	0.903	0.906	0.905	0.906	0.904	0.905	0.907
Case 3	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.3**	**0.3**
Subgroup-specific analyses	0.096	0.098	0.095	0.095	0.100	0.099	0.108	0.094	0.909	0.912
HB model 1 moderate borrowing	0.096	0.098	0.095	0.095	0.100	0.099	0.108	0.094	0.909	0.912
HB model 1 strong borrowing	0.019	0.022	0.020	0.021	0.023	0.019	0.022	0.022	0.677	0.681
HB model 2 (13)	0.032	0.033	0.031	0.031	0.036	0.032	0.037	0.031	0.783	0.790
Case 4	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.3**
Subgroup-specific analyses	0.096	0.098	0.095	0.095	0.100	0.099	0.108	0.094	0.098	0.914
HB model 1 moderate borrowing	0.096	0.098	0.095	0.095	0.100	0.099	0.108	0.094	0.098	0.914
HB model 1 strong borrowing	0.012	0.014	0.013	0.013	0.016	0.012	0.015	0.015	0.014	0.656
HB model 2 (13)	0.029	0.029	0.028	0.028	0.033	0.030	0.034	0.027	0.031	0.747
Case 5	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
Subgroup-specific analyses	0.096	0.098	0.095	0.095	0.100	0.099	0.108	0.094	0.098	0.094
HB model 1 moderate borrowing	0.096	0.098	0.095	0.095	0.100	0.099	0.108	0.094	0.098	0.094
HB model 1 strong borrowing	0.009	0.010	0.010	0.010	0.012	0.007	0.011	0.010	0.010	0.010
HB model 2 (13)	0.022	0.023	0.022	0.023	0.026	0.024	0.026	0.021	0.024	0.023
Case 6	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**
Subgroup-specific analyses	0.913	0.905	0.908	0.907	0.909	0.908	0.910	0.908	0.909	0.912
HB model 1 moderate borrowing	0.913	0.905	0.908	0.907	0.909	0.908	0.910	0.908	0.909	0.912
HB model 1 strong borrowing	0.908	0.910	0.910	0.910	0.912	0.910	0.917	0.910	0.912	0.911
HB model 2 (13)	0.915	0.906	0.909	0.908	0.910	0.908	0.911	0.909	0.909	0.913

Abbreviation: HB, hierarchical Bayesian.

not work. Therefore, relinquishing the strong control of the false-positive error rates would defeat the very purpose of conducting phase II evaluations. Although response rates measured on 5 to 10 subgroups do not provide enough data to determine whether borrowing is appropriate, in the setting of high-dimensional data where treatment effects on thousands of genes are measured, hierarchical models involving borrowing of information across genes can be quite successful (17).

If outcome-adaptive borrowing does not work in the phase II trial setting, how about simple pooling methods for inefficacy/futility monitoring that are not outcome-adaptive (11)? These methods do not increase the false-positive (type I) errors but potentially reduce power, so their application needs to be considered carefully. One would want to be in the situation in which there is a reasonably strong biologic rationale as to why the subgroups should have similar treatment effects. We now give our recommendations for the 3 scenarios mentioned in the beginning of this article. In the first scenario, patients with multiple cancer histologies expressing a particular molecular target, there often could be a rationale for pooling across histologic subgroups. (Caution is still warranted here as is illustrated by a recent experience with a BRAF inhibitor that is highly active in BRAF-mutant melanoma but shows disappointing activity in BRAF-mutant colorectal cancer; ref. 18). In the second scenario, subgroups defined by multiple histologic subtypes of a given cancer histology, the reasonableness of pooling would depend on the cancer and the subtypes. [Consensus on whether pooling is appropriate in a given setting may still be difficult to reach; for example, in soft-tissue sarcoma, Chugh and colleagues (2) used a design with borrowing whereas Maki and colleagues (19) believed pooling was not appropriate and conducted an independent investigation in each subgroup.] For the third scenario, biomarker-defined molecular subtypes of a given cancer histology, we generally cannot see any clear rationale for borrowing

Table 4. Average sample size and empirical probability of rejecting the null hypothesis, design with interim futility monitoring: Equal subgroup accrual rates (10,000 replications)

Design	True response rate in each subgroup				
Case 1	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
Subgroup-specific analyses					
Average sample size	19.4	19.5	19.5	19.5	19.5
Rejection probability	0.093	0.096	0.094	0.098	0.098
HB model 1 moderate borrowing					
Average sample size	19.6	19.6	19.6	19.6	19.4
Rejection probability	0.091	0.094	0.094	0.097	0.097
Simple borrowing (LeBlanc et al.; ref. 11)					
Average sample size	16.2	16.3	16.3	16.3	16.3
Rejection probability	0.064	0.067	0.069	0.068	0.071
Case 2	**0.1**	**0.1**	**0.1**	**0.1**	**0.3**
Subgroup-specific analyses					
Average sample size	19.5	19.5	19.5	19.5	24.6
Rejection probability	0.092	0.089	0.096	0.087	0.900
HB model 1 moderate borrowing					
Average sample size	20.0	20.0	20.0	19.9	24.4
Rejection probability	0.092	0.090	0.097	0.088	0.892
Simple borrowing (11)					
Average sample size	18.6	18.6	18.7	18.7	22.8
Rejection probability	0.086	0.084	0.091	0.084	0.776
Case 3	**0.1**	**0.1**	**0.1**	**0.3**	**0.3**
Subgroup-specific analyses					
Average sample size	19.4	19.4	19.4	24.6	24.6
Rejection probability	0.093	0.092	0.089	0.897	0.900
HB model 1 moderate borrowing					
Average sample size	20.4	20.4	20.4	24.5	24.6
Rejection probability	0.094	0.093	0.090	0.895	0.894
Simple borrowing (11)					
Average sample size	19.3	19.2	19.3	24.0	24.1
Rejection probability	0.092	0.091	0.089	0.875	0.874
Case 4	**0.1**	**0.1**	**0.3**	**0.3**	**0.3**
Subgroup-specific analyses					
Average sample size	19.5	19.5	24.6	24.7	24.6
Rejection probability	0.097	0.095	0.900	0.901	0.892
HB model 1 moderate borrowing					
Average sample size	21.0	21.0	24.6	24.6	24.6
Rejection probability	0.100	0.098	0.901	0.902	0.891
Simple borrowing (11)					
Average sample size	19.5	19.4	24.5	24.6	24.6
Rejection probability	0.097	0.095	0.898	0.899	0.889

Abbreviation: HB, hierarchical Bayesian.

of information. (This opinion is apparently shared by the BATTLE investigators; ref. 3). It should be noted that we are not considering situations in which patient subgroups are defined by the levels of biomarker expression which are expected to be related to response; in this case, sequential procedures that borrow information according to the ordered nature of the subgroups could be appropriate.

To summarize our results, in the phase II setting, the outcome-adaptive approach does not seem to have enough information to determine whether borrowing is appropriate and is therefore of limited use. Designs that borrow (pool) information across subgroups can reduce power to detect treatments that work only in some subgroups. In specialized settings in which a reasonable biologic rationale exists for expecting similar treatment effects in the

Table 5. Average sample size and empirical probability of rejecting the null hypothesis, design with interim futility monitoring: Unequal subgroup accrual rates (30%, 20%, 20%, 20%, 10%; 10,000 replications)

Design	True response rate in each subgroup				
Case 1	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
Subgroup-specific analyses					
Average sample size	19.5	19.5	19.5	19.4	19.6
Rejection probability	0.095	0.090	0.093	0.097	0.098
HB model 1 moderate borrowing					
Average sample size	20.7	19.8	19.6	19.7	21.2
Rejection probability	0.094	0.091	0.093	0.100	0.096
HB model 1 strong borrowing					
Average sample size	20.0	19.2	19.2	19.0	20.4
Rejection probability	0.026	0.025	0.024	0.027	0.022
Simple borrowing (LeBlanc et al.; ref. 11)					
Average sample size	18.3	16.6	16.6	16.5	13.2
Rejection probability	0.077	0.070	0.069	0.072	0.055
Case 2	**0.1**	**0.1**	**0.1**	**0.1**	**0.3**
Subgroup-specific analyses					
Average sample size	19.5	19.4	19.5	19.5	24.6
Rejection probability	0.093	0.085	0.097	0.094	0.900
HB model 1 moderate borrowing					
Average sample size	20.8	20.2	20.2	20.1	24.7
Rejection probability	0.096	0.087	0.097	0.095	0.898
HB model 1 strong borrowing					
Average sample size	20.8	20.3	20.3	20.0	24.2
Rejection probability	0.033	0.037	0.033	0.037	0.742
Simple borrowing (11)					
Average sample size	18.9	18.0	18.1	18.1	19.6
Rejection probability	0.088	0.077	0.087	0.086	0.645
Case 3	**0.1**	**0.1**	**0.1**	**0.3**	**0.3**
Subgroup-specific analyses					
Average sample size	19.4	19.5	19.5	24.6	24.6
Rejection probability	0.093	0.097	0.097	0.893	0.897
HB model 1 moderate borrowing					
Average sample size	21.0	20.5	20.5	24.5	24.7
Rejection probability	0.096	0.098	0.099	0.891	0.897
HB model 1 strong borrowing					
Average sample size	21.9	21.8	21.8	24.5	24.7
Rejection probability	0.051	0.057	0.062	0.800	0.796
Simple borrowing (11)					
Average sample size	19.3	19.2	19.1	24.0	23.6
Rejection probability	0.093	0.096	0.096	0.857	0.848
Case 4	**0.1**	**0.1**	**0.3**	**0.3**	**0.3**
Subgroup-specific analyses					
Average sample size	19.6	19.5	24.7	24.6	24.6
Rejection probability	0.095	0.099	0.898	0.899	0.901
HB model 1 moderate borrowing					
Average sample size	21.3	21.0	24.7	24.6	24.7
Rejection probability	0.099	0.102	0.898	0.898	0.903
HB Model 1 strong borrowing					
Average sample size	23.1	23.2	24.8	24.8	24.8
Rejection probability	0.085	0.085	0.842	0.845	0.840
Simple borrowing (11)					
Average sample size	19.6	19.4	24.5	24.5	24.5
Rejection probability	0.096	0.099	0.891	0.892	0.892

Abbreviation: HB, hierarchical Bayesian.

Table 6. Examples of different degrees of borrowing for data in Table 1

Model	Probability (response rate in subgroup 5 > 30% | data)	
	Scenario 1	Scenario 2
Subgroup-specific analysis		
No borrowing model: beta-binomial	0.437	0.437
Hierarchical model (Thall et al.; ref. 9)		
No borrowing: mean of precision hyperprior 0.01 [Gamma(2,200)]	0.459	0.446
Moderate borrowing: mean of precision hyperprior 0.1 [Gamma(2,20)]	0.453	0.382
Strong borrowing: mean of precision hyperprior 1 [Gamma(2,2)]	0.464	0.160

subgroups, simple pooling of information across subgroups may be appropriate.

Appendix

Thall and colleagues (9) proposed the following parameterization of the hierarchical Bayesian model. For the response rates π_i, a logistic model was assumed: $\theta_i = \log\left\{ {}^{\pi_i}/_{(1-\pi_i)} \right\}$ for $i = 1, \ldots, K$. The θ_i were assumed to be independent and identically distributed normal variables with mean μ and variance $1/\tau$ (note that parameter τ represents precision in this parameterization). The hyperprior distribution for μ was assumed to be normal with mean equal to $\text{logit}(0.2) = 1.386$ and variance 10. The mean of the hyperprior for μ was set to the logit of 0.20 to represent the prior belief that the average response rate is half-way between the interesting response rate of 0.3 and the uninteresting response rate of 0.1. For the hyperprior distribution for τ, Thall and colleagues used a gamma distribution with parameters 2 and 20. This distribution has a mean of 0.1 corresponding to a relatively low precision in the prior distribution of θ_i and thus results in a relatively modest amount of borrowing. To allow more borrowing, we also considered a hyperprior (for τ) with a larger mean of 1 (gamma distribution with parameters 2 and 2), corresponding to a higher precision in the prior distribution of θ_i.

Berry and colleagues (13) assumed that the π_i are randomly sampled from a beta distribution with parameters a and b. Each of the parameters a and b was assumed to have an independent beta distribution for its hyperprior. In our simulations, for a, we used a uniform distribution on [0, 4], and for b, we used a uniform distribution on [0, 16]. Similarly to the Thall and colleagues model, the hyperprior parameters were selected to reflect the prior belief that the average response rate is 0.20.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: B. Freidlin, E.L. Korn
Development of methodology: B. Freidlin, E.L. Korn
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): B. Freidlin
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): B. Freidlin, E.L. Korn
Writing, review, and/or revision of the manuscript: B. Freidlin, E.L. Korn
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): B. Freidlin
Study supervision: B. Freidlin

Received April 13, 2012; revised November 14, 2012; accepted December 22, 2012; published OnlineFirst January 9, 2013.

References

1. 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;16:1764–9.
2. Chugh R, Wathen JK, Maki RG, Benjamin RS, Patel SR, Meyers PA, et al. Phase II multicenter trial of imatinib in 10 histologic subtypes of sarcoma using a bayesian hierarchical statistical model. J Clin Oncol 2009;27:3148–53.
3. Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer–a step toward personalized medicine. Clin Trials 2008;5:181–93.
4. Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. Clin Trials 2005;2:295–300.
5. Hobbs BP, Carlin BP. Practical Bayesian design and analysis for drug and device clinical trials. J Biopharm Stat 2008;18:54–80.
6. Biswas S, Liu DD, Lee JJ, Berry DA. Bayesian clinical trials at the University of Texas M. D. Anderson Cancer Center. Clin Trials 2009;6:205–16.
7. Lindley DV, Smith AFM. Bayesian estimates for the linear model. J Royal Statist Soc Ser B 1972;34:1–41.
8. Kass RE, Steffey D. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J Am Statist Association 1989;84:717–726.
9. Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LH, Benjamin RS. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. Stat Med 2003;22:763–80.
10. Berry DA. A guide to drug discovery: Bayesian clinical trials. Nat Rev Drug Discov 2006;5:27–36.
11. LeBlanc M, Rankin C, Crowley J. Multiple histology phase II trials. Clin Cancer Res 2009;15;4256–62.
12. Carlin BP, Louis TA. Bayes and empirical Bayes methods for data analysis. 2nd ed. Boca Raton, FL; Chapman & Hall/CRC; 2000.

13. Berry SM, Carlin BP, Lee JJ, Muller P. Bayesian adaptive methods for clinical trials. Boca Raton, FL; Chapman & Hall/CRC; 2010.
14. R Development Core Team. *R*: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
15. Spiegelhalter D, Thomas A, Best N, Bayesian Inference Using Gibbs Sampling Manual (BUGS 0.5) Cambridge, UK: MRC Biostatistics Unit; 1995.
16. United States Food and Drug Administration: 2010 draft guidance for industry on adaptive design clinical trials for drugs and biologics. [cited 2012 Jan 22]. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201790.pdf.
17. Ibrahim JG, Chen M-H, Gray RJ. Bayesian models for gene expression with DNA microarray data. J Am Stat Associat 2002;97:88–99.
18. Kopetz S, Desai J, Chan E, Hecht JR, O'Dwyer PJ, Lee RJ, et al. PLX4032 in metastatic colorectal cancer patients with BRAF tumors. 2010 ASCO Annual Meeting. J Clin Oncol 28:15s, 2010 (suppl; abstr 3534).
19. Maki RG, D'Adamo DR, Keohan ML, Saulle M, Schuetze SM, Undevia SD, et al. Phase II study of sorafenib in patients with metastatic or recurrent sarcomas. J Clin Oncol 2009;27:3133–40.