STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:329–339
Published online 18 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1975

On the efficiency of targeted clinical trials

A. Maitournam and R. Simon*,†

Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892-7434, U.S.A.

SUMMARY

The development of genomics-based technologies is demonstrating that many common diseases are heterogeneous collections of molecularly distinct entities. Molecularly targeted therapeutics is often effective only for some subsets patients with a conventionally defined disease. We consider the problem of design of phase III randomized clinical trials for the evaluation of a molecularly targeted treatment when there is an assay predictive of which patients will be more responsive to the experimental treatment than to the control regimen. We compare the conventional randomized clinical trial design to a design based on randomizing only patients predicted to preferentially benefit from the new treatment. Trial designs are compared based on the required number of randomized patients and the expected number of patients screened for randomization eligibility. Relative efficiency depends upon the distribution of treatment effect across patient subsets, prevalence of the subset of patients who respond preferentially to the experimental treatment, and assay performance. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: genomics; clinical trials; molecularly targeted therapeutics; pharmacogenomics; sample size; normal mixture

1. INTRODUCTION

Patient responses to therapeutics are often heterogeneous. In oncology, for example, response rates of less than 50 per cent are not uncommon. Most drugs have potential side effects and hence the cost to the patient of receiving an ineffective drug can be substantial.

Genomic technologies such as DNA sequencing, mRNA transcript profiling, and comparative genomic hybridization [1] are providing evidence that many diseases are more molecularly heterogeneous than previously recognized. For example, substantial effort is currently placed in developing mutation signatures and gene expression signatures of tumors [2, 3]. Such studies provide insight into the heterogeneity of disease pathogenesis and enable molecular disease taxonomies to be defined. Some genetic profiling studies identify new therapeutic targets. In other cases, genomic profiling of disease tissue has provided accurate predictors of response

*Correspondence to: Richard Simon, Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892-7434, U.S.A.

†E-mail: rsimon@mail.nih.gov

Received September 2003
Accepted June 2004

Copyright © 2004 John Wiley & Sons, Ltd.

to therapy even when the mechanism of action of the therapy is not fully understood and target specific assays are not available [4–6].

An important current challenge is determination of how to integrate genomic technologies and biomarker-based classification systems developed using these technologies into clinical drug development. Current drug development for oncology is focused on drugs that are molecularly targeted to protein products of genes that are disregulated in cancer cells [7–10]. Because of the molecular heterogeneity of many cancers and the diversity of genetic lesions occurring in genomically unstable tumors, molecularly targeted drugs may only be effective for a subset of the patients with a traditionally classified type of cancer. Recent clinical trials of several classes of molecularly targeted drugs support this hypothesis. For example [11], Iressa, an epidermal growth factor receptor drug showed clinically meaningful activity in 10 per cent of patients with advanced lung cancer.

Host genetic factors may also influence the rates and patterns of drug distribution and metabolism leading to divergent responses to administration of the drug [12, 13]. Substantial effort has been placed on developing cost efficient technologies for genotyping patients to identify pharmacokinetic and pharmacodynamic phenotypes that modulate drug response.

Our objective is to evaluate clinical trial strategies for the utilization of predictive models and biomarkers of drug response. We consider the design of a phase III randomized clinical trial for evaluating a test regimen T compared to a control regimen C. T might consist of a new drug and C might consist of placebo or no drug administration. Alternatively, T might consist of a new drug plus standard treatment with C being standard treatment alone. We assume that patients potentially eligible for a particular clinical trial can be partitioned into a group R+ predicted to be responsive to the new drug and a group R− that is not predicted to be responsive. The prediction may be based on an assay that measures the expression of the receptor or ligand of the test drug, or on a multivariate gene expression model derived from transcript profiling [4, 6, 14].

We consider two clinical evaluation strategies. The first is to conduct a clinical trial comparing T to C for all patients in the disease category. This is the traditional approach and we will refer to it as the untargeted design. An alternative strategy is to randomize only R+ patients and compare treatment T to C within that subset. We will refer to this as the targeted design. We will compare these two designs with regard to the number of patients required for randomization, and the number of patients required for screening. A third approach is to randomize all patients and to compare T to C separately within the R+ patients and the R− patients. This would generally require at least twice as many patients as the targeted design, however, and so will not be explicitly included in the comparisons we will make.

Heterogeneity in prognosis has long been recognized in clinical trials and is handled usually by balancing randomizations by prognostic factors and by adjusting analysis for prognostic covariates (see for example Reference [15]). Clinical trials have rarely been sized large enough for statistically adequate analysis of subsets of patients. Occasionally, however, the mechanism of action of a drug has been understood sufficiently for clinical trials to be conducted in a targeted manner, for example with Tamoxifen and estrogen receptors in breast cancer. Betensky *et al.* [16] have demonstrated that substantial loss of power is possible when an unrecognized treatment by covariate interaction exists. There has been relatively little work, however, on quantitatively evaluating the efficiency of targeted versus untargeted designs. Brittain and Wittes [17] evaluated screening in the context of potentially excluding patients who show poor treatment compliance during a run-in period [18]. Elston *et al.* [19] developed

Copyright © 2004 John Wiley & Sons, Ltd.

sample size formulas for testing which candidate genes modify drug response in the context of randomized clinical trials. Fijal et *al.* [20] investigated the importance of protective genotype on the sample size and the duration of prevention trials. Their model assumed that the set of patients not developing the disease was a mixture of those protected by non-susceptible genotype and those protected by treatment. They recommended the genetic screening when the proportion protected was large.

Our paper is organized as follows. We describe the model and derive sample size formulas for targeted and untargeted designs in the next section. In Section 3, we compare the efficiencies of the designs. The designs are compared with regard to the number of patients required for randomization and with regard to the number of patients required for screening. For the untargeted design, the two quantities are the same, but many more patients may be required for screening than for randomization for the targeted design. The designs are compared as a function of the proportion of patients in the referral population who are R+, the treatment effects for R+ and R− patients, and the sensitivity and specificity of the assay for identifying whether a patient is R+ or R−. Finally, the results are discussed in Section 4.

2. DERIVATION OF SAMPLE SIZE FORMULAS

Let μ_0 and μ_1 denote the mean responses for control patients in subsets R− and R+, respectively, and let γ denote the proportion of patients in R−. For treatment T the mean responses are denoted $\mu_{0\mathrm{T}}$ and $\mu_{1\mathrm{T}}$ for R− and R+ subsets, respectively. For the untargeted trial design, patients are not ‘typed’ and so the response of a patient has a mixture distribution with mean $\gamma\mu_0 + (1-\gamma)\mu_1$ for C and $\gamma\mu_{0\mathrm{T}} + (1-\gamma)\mu_{1\mathrm{T}}$ for T. This mixture model has been commonly used in classical genetics [21, 22].

Under the null hypothesis that the distribution of response is the same for C and T, the t-statistic

$$t = \frac{\bar{x}_{\mathrm{T}} - \bar{x}_{\mathrm{C}}}{s\sqrt{2/n}} \tag{1}$$

is asymptotically normally distributed with mean 0 and standard deviation 1 where s, $\bar{x}_{\mathrm{C}}$, $\bar{x}_{\mathrm{T}}$, denote, respectively, the square root of the pooled within treatment group variance, the estimated means for control and treatment groups and n is the number of patients in each of the two treatment groups.

The asymptotic normality of the t-statistic is a consequence of the central limit theorem. If the responses within each treatment group and subset are normally distributed with constant variance σ^2, then the number of patients per treatment n required to achieve power $1-\beta$ for rejecting the null hypothesis at significance level α is asymptotically given by

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{[\gamma(\mu_{0\mathrm{T}} - \mu_0) + (1-\gamma)(\mu_{1\mathrm{T}} - \mu_1)]^2/\{2\sigma^2 + \gamma(1-\gamma)[(\mu_1 - \mu_0)^2 + (\mu_{0\mathrm{T}} - \mu_{1\mathrm{T}})^2]\}} \tag{2}$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ denote percentiles of the standard normal distribution. (See the supplemental material for a derivation of equation (2).) For the targeted design where all patients are typed and only R+ patients are randomized, the number of randomized patients is approximately:

$$n_t = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_{1T} - \mu_1)^2/2\sigma^2} \tag{3}$$

The sample size in (3) is reciprocally proportional to the square of the treatment effect $\mu_{1T} - \mu_1$ expressed in units of the standard deviation of the difference in two responses, in this case $\sqrt{2\sigma^2}$. The sample size in (2) is of the same form. The denominator is the square of the treatment effect for the untargeted study. The term in curly brackets is the square of the corresponding standard deviation for the untargeted design where the responses have mixture distributions.

The total number of required randomized patients will be $2n$ for the untargeted design, and $2n_t$ for the targeted design. The ratio r between the required numbers of patients for the two designs is

$$r = \frac{n}{n_t} = \frac{1 + \dfrac{\gamma(1-\gamma)}{2\sigma^2}[(\mu_1 - \mu_0)^2 + (\mu_{1T} - \mu_{0T})^2]}{\left[1 - \gamma\left(1 - \dfrac{\mu_{0T} - \mu_0}{\mu_{1T} - \mu_1}\right)\right]^2} \tag{4}$$

The treatment effect for the untargeted trial design is a weighted average of the treatment effect for R− patients and the treatment effect for R+ patients, weighted by the proportions γ of R− patients and $1 - \gamma$ of R+ patients. In the case where there is no treatment effect for R− patients and where the average response for control patients is the same for the two subsets, equation (2) simplifies and the ratio of sample sizes (4) becomes

$$r = \frac{n}{n_t} = \frac{1 + \dfrac{\gamma(1-\gamma)}{2\sigma^2}[(\mu_{1T} - \mu_1)^2]}{[1-\gamma]^2} \tag{5}$$

This ratio is the relative efficiency of the targeted design relative to the untargeted design. It can be seen from (5) that this relative efficiency increases approximately with the reciprocal of the square of the proportion of R+ patients because the term $\gamma(1-\gamma)$ in the numerator changes much more slowly than the term $1/(1-\gamma)^2$. For $\gamma = 0$ all of the patients are in the R+ group and consequently the two trial designs are equivalent and the relative efficiency is 1. For γ close to 1, there are very few R+ patients but the treatment effect for the untargeted trial design is close to zero, hence the untargeted trial requires a huge sample size and the relative efficiency approaches infinity. Expression (5) also indicates that the relative efficiency of the targeted design increases with the square of the treatment effect in the R+ subset. When the treatment effect is small, both designs require large numbers of patients. When the treatment effect for R+ patients is large, the targeted design may require very few patients, but the untargeted design may still require a large number of patients depending on the value of γ.

The above results provide some insight into the nature of the relationship between the relative efficiencies of the two designs and the parameters of the model. The results are valid only asymptotically, however. For the untargeted design the finite sample distribution of the t-statistic may be bimodal and the asymptotic approximation may converge slowly unless the treatment effect is small. Consequently, we have derived sample size formulas that should be more accurate for small sample sizes using the untargeted design. Let X denote a response

Copyright © 2004 John Wiley & Sons, Ltd.

from a patient in the control group C and Y from a patient in the treatment group T. If outcomes for groups C and T are compared using a two-sample Wilcoxon test, the power is approximately [23]:

$$1-\Phi\left\{\frac{0.5n^2+z_{1-\alpha/2}\sqrt{n^2\dfrac{2n+1}{12}}-0.5-n^2p_1}{\mathrm{var}(W_{XY})}\right\} \tag{6}$$

where $\mathrm{var}(W_{XY})$ is the variance of the Mann–Whitney statistic given by

$$\mathrm{var}(W_{XY})=n^2\{p_1(1-p_1)+(n-1)(p_2+p_3-2p_1^2)\} \tag{7}$$

$p_1=\Pr(X<Y)$ is the probability that a control outcome X is less than a treatment outcome Y. $p_2=\Pr(X<Y_1,X<Y_2)$ is the probability that a control outcome is simultaneously less than two independent treatment outcomes, and $p_3=\Pr(X_1<Y,X_2<Y)$ is the probability that a treatment outcome is simultaneously greater than two independent control outcomes. In expression (6) Φ denotes the cumulative distribution function for the standard normal distribution.

We have used expressions (6) and (7) to compute the smallest sample size that provides a specified power $1-\beta$ for rejecting the null hypothesis for any specified pair of treatment distributions. Expressions (6) and (7) are quite general and make no restrictions on the nature of the distributions being compared other than that they are continuous. We have applied this result for both the targeted trial design and for the untargeted trial design. We computed the parameters p_1, p_2, and p_3 using Monte-Carlo integration.

If $2n_\mathrm{t}$ denotes the number of patients randomized to the targeted trial, then $2n_\mathrm{t}/(1-\gamma)$ is the expected number of patients screened and typed in order to obtain $2n_\mathrm{t}$ patients to be randomized. For the untargeted design, the number of screened patients equals the number of randomized patients. In the results presented below, we compare the two trial designs with regard to the number of randomized patients and the number of screened patients.

The model described above can be generalized to account for error in the assaying of patients as belonging to subset R$-$ or R$+$. Such errors only affect results for the targeted design. Let λ_sens denote the sensitivity of the assay for diagnosing R$+$ patients and let λ_spec denote the corresponding specificity. The mean response for patients selected for the targeted study becomes

$$\omega_-\mu_0+\omega_+\mu_1$$

for the control group and

$$\omega_-\mu_{0\mathrm{T}}+\omega_+\mu_{1\mathrm{T}}$$

for the treatment group where

$$\omega_+=\lambda_\mathrm{sens}(1-\gamma)/\{(1-\lambda_\mathrm{spec})\gamma+\lambda_\mathrm{sens}(1-\gamma)\} \tag{8}$$

is the positive predictive value (PPV) of the assay and $\omega_-=1-\omega_+$.

The treatment effect for the targeted study is thus diluted by ω_+. If $2n_\mathrm{t}$ denotes the number of patients randomized to the targeted trial, then

$$2n_\mathrm{t}/\{(1-\lambda_\mathrm{spec})\gamma+\lambda_\mathrm{sens}(1-\gamma)\} \tag{9}$$

is the expected number of patients screened in order to obtain $2n_\mathrm{t}$ randomized patients.

Copyright © 2004 John Wiley & Sons, Ltd.

3. COMPARISON OF DESIGN EFFICIENCY

We evaluated the relative efficiency of the untargeted design and the targeted design with regard to: (a) the number of patients required for randomization with each design to achieve 80 per cent statistical power at a 5 per cent two-sided significance level; and (b) the number of patients that each design requires to screen in order to achieve the target number of randomized patients. For the untargeted design, the number screened equals the number randomized, but this is not true for the targeted design.

In comparing the two designs, we evaluated the effect of the proportion of patients who are R+, the size of the treatment effect for the R− patients relative to the R+ patients, the sensitivity of the assay for detecting R+ patients (λ_{sens}) and the specificity of the assay (λ_{spec}). We assumed without loss of generality that $\mu_0 = 0$, and $\sigma = 1$. For the results shown below, $\mu_1 = 0$; that is, there was no prognostic difference between R+ and R− patients in the control group. We examined other cases but found that the relative efficiency results were very insensitive to such a prognostic effect.

The ratio of the number of randomized patients required for the untargeted design relative to the number required for the targeted design is shown in Figure 1. The horizontal axis of each plot is the proportion of R+ patients $(1-\gamma)$. The upper three panels are for scenarios in which the treatment effect is limited to the R+ patients. Those panels are based on values $\mu_{0\text{T}} = 0$, $\mu_{1\text{T}} = 1$; that is, the size of the treatment effect for the R+ patients is equal to one standard deviation and there is no treatment effect for the R− patients. Results were also computed (but not shown) for similar scenarios with the treatment effect for R+ patients being one-half standard deviation. Although the sample sizes changed when the treatment effects were reduced, the relative sample sizes for the two designs remained almost the same as those shown in Figure 1. The lower three panels correspond to scenarios where the treatment effect for the R− patients is one half as large as for R+ patients. The three columns of panels in Figure 1 correspond to assay specificities of 1, 0.8 and 0.6, respectively, and the three curves in each panel correspond to assay sensitivities of 1, 0.8 and 0.6. This range of sensitivities and specificities encompasses those used in most diagnostic medical applications.

Figure 1 shows that the targeted design is more efficient (ratio greater than one) that the untargeted design in term of number of randomized patients, even in the worst situation when the sensitivity and the specificity are 0.6 and the R− patients benefit from the treatment (see the last plot of lower panel of Figure 1). The specificity is more important than the sensitivity; for some of the panels the three curves corresponding to different assay sensitivities cannot even be distinguished. However if the specificity decreases, the efficiency of targeted design decreases because some R− patients selected for inclusion in the targeted trial will dilute the treatment effect.

Figure 1 shows that the targeted design loses much of its efficiency advantage if the treatment effect for the R− patients is half as large as for the R+ patients (lower panels) rather than zero (upper panels). This is because the sample size for the untargeted design decreases when there is a treatment effect for R− patients, whereas the sample size for the targeted design is unaffected.

The relative efficiency of the targeted design with regard to number of randomized patients also tends to increase as the proportion of R+ patients decreases. When the proportion of R+ patients is 1, the two designs are equivalent. When the proportion of R+ patients is small, the treatment effect for the untargeted design is very small if there is no treatment effect for R−

Copyright © 2004 John Wiley & Sons, Ltd.

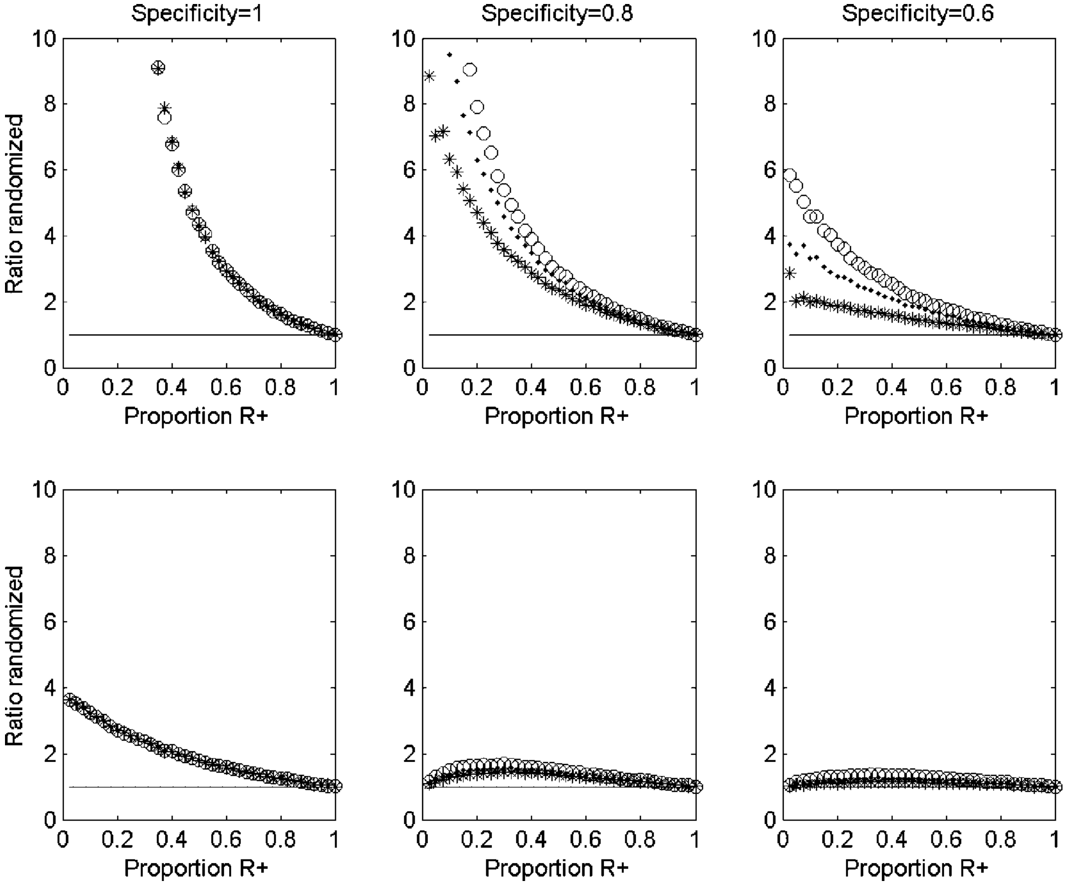

Figure 1. Ratio of number randomized for untargeted versus targeted designs. Upper panel: no treatment effect for R− patients. Lower panel: treatment effect for R− patients half that of R+ patients. ○ Sensitivity = 1; ● Sensitivity = 0.8; $*$ Sensitivity = 0.6.

patients. The effect of the proportion of R+ patients is dependent, however, on the specificity of the assay, particularly when there is a treatment effect for the R− patients. The number of randomized patients required for the targeted design depends on the proportion of assay positive patients who are really R+, that is, the positive predictive value of the assay. When the treatment effect is limited to R+ patients, the treatment effect for the targeted design equals $\omega_+(\mu_{1\mathrm{T}}-\mu_1)$ where the positive predictive value ω_+ is defined by equation (8). Consequently, the treatment effect is diluted by an amount equal to the positive predictive value. When the treatment effect for R− patients is half the size of the treatment effect for the R+ patients, the net treatment effect for the targeted design is $(\mu_{1\mathrm{T}} - \mu_1)((\omega_+ + 1)/2)$. If the specificity of the assay is 1, then the positive predictive value is 1 regardless of the sensitivity or the prevalence of R+ patients. If the specificity is less than 1, however, then the PPV depends on sensitivity and prevalence.

Copyright © 2004 John Wiley & Sons, Ltd.

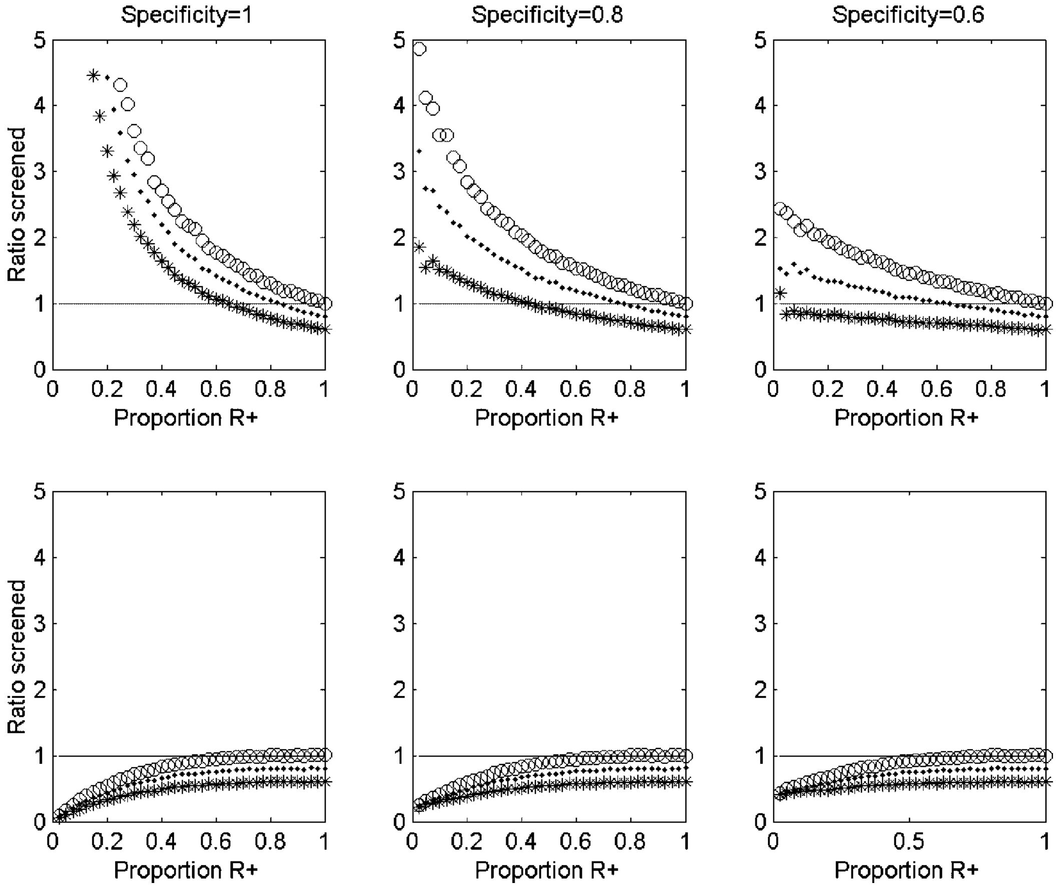

Figure 2. Ratio of number randomized for untargeted design to number screened for targeted design. Upper panel: no treatment effect for R− patients. Lower panel: treatment effect for R− patients half that of R+ patients. ○ Sensitivity = 1; ● Sensitivity = 0.8; ∗ Sensitivity = 0.6.

Figure 2 shows the ratio of the number of patients randomized for the untargeted trial design to the number of patients screened for the targeted design. The format of the figure is the same as for Figure 1. The relative efficacy of the two designs with regard to the number of patients screened depends on the assay parameters in a more complex way. If r denotes the relative efficiency of untargeted to targeted design with regard to number of randomized patients and rs denotes the relative efficiency with regard to number of screened patients, then

$$rs = r[\gamma(1-\lambda_{\text{spec}}) + (1-\gamma)\lambda_{\text{sens}}] \tag{10}$$

As noted above, r depends on the positive predictive value in a non-linear way because it multiplies the treatment effect for the targeted design.

As in Figure 1, the results are quite different when the treatment effect is limited to R+ patients (upper panel) compared to when the treatment effect for the R− patients is half that of the R+ patients (lower panel). When the treatment effect is limited to R+ patients, then

Copyright © 2004 John Wiley & Sons, Ltd.

the number of patients screened for the targeted design is generally less than the number randomized for the untargeted design unless the assay has low sensitivity and either low specificity or the proportion of R+ patients is high. As the sensitivity decreases, more patients must be screened for the targeted design to obtain a specified number to randomize. Changes in the specificity and prevalence of R+ patients influence the ratio of randomized patients and thereby indirectly influence the ratio of screened patients.

If the treatment effect for the R− patients is half that of the R+ patients, then the number of patients to be screened for the targeted design is generally greater than the number of patients required for the untargeted design. The difference depends heavily on the sensitivity of the assay and on the proportion of R+ patients. If the proportion of R+ patients is 50 per cent and the specificity of the assay is 0.8, then 39 per cent more patients are required for screening with the targeted design if the sensitivity is 0.8, and 84 per cent more if the sensitivity is 0.6. The targeted design requires, respectively, 30 and 26 per cent fewer patients to be randomized than the untargeted design under these conditions.

4. DISCUSSION

When the treatment effect is limited to R+ patients and the specificity of the assay is close to 1, the targeted design is generally much more efficient than the untargeted design with regard to both the number of randomized and number of screened patients. As the specificity of the assay degrades however, the advantage of the targeted design decreases. Nevertheless, the targeted design remains advantageous with regard to number of randomized patients even if the assay sensitivity and specificity are 0.6 and the proportion of R+ patients is less than about 70 per cent (Figure 1). If, however, the assay sensitivity is poor (e.g. 0.6), then the savings in randomized patients will be somewhat penalized by an increase in the number of screened patients (Figure 2).

If the treatment effect for the R− patients is half as large as for the R+ patients, the targeted design requires more patients to be screened than the untargeted design even if the assay is perfect. The increased number of screened patients for the targeted design depends critically on the sensitivity of the assay (Figure 2). Under these conditions, there is no substantial advantage of the targeted design with regard to number of randomized patients unless the assay has specificity close to 1. If the specificity is close to 1, then selection of the preferred design will depend on other considerations such as relative costs of screening versus randomizing, and value of a broader labeling indication.

Our results indicate that the relative efficiency of the targeted and untargeted designs do not depend heavily on whether the patients predictive to be responsive to the new treatment are prognostically different than the other patients or on the size of the treatment effect for the responsive patients.

Our results highlight the importance of assay specificity in determining the efficiency of targeted clinical trials with regard to required number of randomized patients and the importance of assay sensitivity on relative efficiency with regard to required number of screened patients.

Our results have indicated that two important features in determining the efficiency of targeted clinical trials are assay specificity and whether a treatment effect is expected for the less responsive group of patients. These two features are related. We have chosen to model them separately, because in some cases it is important to distinguish the adequacy of an assay for a drug target from the effects of multiple drug targets. If preliminary data indicate that substantial responses occur among patients predicted to be R−, then there may not be an advantage of the targeted design. This type of data can be developed during phase II development. In some cases, however, there will be great efficiencies to be earned from devoting the effort to develop accurate predictors of responsive patients during phase II.

In cases where the new treatment is to be compared to a standard treatment rather than a placebo or no-treatment control, the most relevant assay is one which identifies patients likely to be more responsive to the new treatment than the control. This distinction is important when the control treatment is itself highly active. For such settings, it is important to develop such relative-activity assays during phase II development based on data for patients who receive the new treatment and those who receive the regimen that will be used as control in the phase III trial.

Our model can be generalized in many ways. We have addressed continuous response endpoints, and a similar model could be developed for binary response or right-censored survival data. We expect that most of our qualitative findings would also apply for those types of endpoints. The presence of covariates, or unbalanced designs would not be expected to substantially change our findings. Our approach could also be extended to continuous predictive assays with established ROC performance characteristics [24] and one could simultaneously optimize the assay threshold and the clinical trial design.

REFERENCES

1. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 1992; **258**(5038):818–821.
2. Perou CM, Sorlie T, Eisen MB, De Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. *Nature* 2000; **406**(6797):747–752.
3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403**(6769):503–511.
4. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS *et al.* Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature Medicine* 2002; **8**(1):68–74.
5. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *New England Journal of Medicine* 2002; **346**(25):1937–1947.
6. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* 2003; **95**(1):14–18.
7. Balis FM. Evolution of anticancer drug discovery and the role of cell-based screening. *Journal of The National Cancer Institute* 2002; **94**(2):78–79.
8. Fox E, Curt GA, Balis FM. Clinical trial design for target-based therapy. *The Oncologist* 2002; **7**(5):401–409.
9. Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? *Journal of Clinical Oncology* 2001; **19**(1):265–272.
10. Hartwell LH, Szankasi P, Roberts CJ, Murray AW, Friend SH. Integrating genetic approaches into the discovery of anticancer drugs. *Science* 1997; **278**(5340):1064–1068.
11. Grunwald V, Hidalgo M. Developing inhibitors of the epidermal growth factor receptor for cancer treatment. *Journal of The National Cancer Institute* 2003; **95**(12):851–867.
12. Daly AK. Molecular basis of polymorphic drug metabolism. *Journal of Molecular Medicine* 1995; **73**(11): 539–553.
13. Nebert DW. Polymorphisms in drug-metabolizing enzymes: what is their clinical relevance and why do they exist? *American Journal of Human Genetics* 1997; **60**(2):265–271.
14. Rosenwald A, Wright G, Wiestner A, Chan WC, Connors JM, Campo E, Gascoyne RD, Grogan TM, Muller-Hermelink HK, Smeland EB, Chiorazzi M *et al.* The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* 2003; **3**(2):185–197.

Copyright © 2004 John Wiley & Sons, Ltd.

15. Simon R. Heterogeneity and Standardization in Clinical Trials. In *Controversies in Cancer*, *Design of Trials and Treatment*, Tagnon HJ, Staquet MJ (eds). Masson Publishing USA, Inc.: New York, Paris, Barcelona, Milan, Mexico City, Rio De Janeiro, 1978; 37–49.
16. Betensky RA, Louis DN, Cairncross JG. Influence of unrecognized molecular heterogeneity on randomized trials. *Journal of Clinical Oncology* 2002; **20**(10):2495–2499.
17. Brittain E, Wittes J. The run-in period in clinical trials. The effect of misclassification on efficiency. *Controlled Clinical Trials* 1990; **11**:327–338.
18. Lang JL. The use of a run-in to enhance compliance. *Statistics in Medicine* 1990; **9**:87–95.
19. Elston RC, Idury RM, Cardon LR, Lichter JB. The study of candidate genes in drugs trials: sample size considerations. *Statistics in Medicine* 1999; **18**(6):741–751.
20. Fijal BA, Hall JM, Witte JS. Clinical trials in the genomic era: effects of protective genotypes on sample size and duration of trial. *Controlled Clinical Trials* 2000; **21**(1):7–20.
21. Schork NJ, Allison DB, Theil B. Mixture distributions in human genetics. *Statistical Methods in Medical Research* 1996; **5**:155–178.
22. McLachlan G, Peel D. *Finite Mixture Models*, Wiley Series in Probability and Statistics. Wiley: New York, Chichester, Weinheim, Brisbane, Singapore, Toronto, 2000; 9–15.
23. Lehmann EL, D’Abrera HJM. *Nonparametrics*: *Statistical Methods Based on Ranks*. Holden-Day, Inc.: San Francisco, McGraw-Hill International Book Company: Dusseldorf, Johannesburg, Sao Paulo, London, Singapore, Mexico, Sydney, New York, Panama, Toronto, 1975; 1–119.
24. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**(1):29–36.

Copyright © 2004 John Wiley & Sons, Ltd.