

Clinical Trial Designs for Predictive Biomarker Validation

XH Andrew Zhou, a joint work with Kim Young

azhou@u.washington.edu

Professor, Department of Biostatistics, University of Washington

Director of Biostatistics Unit, Seattle VA Medical Center

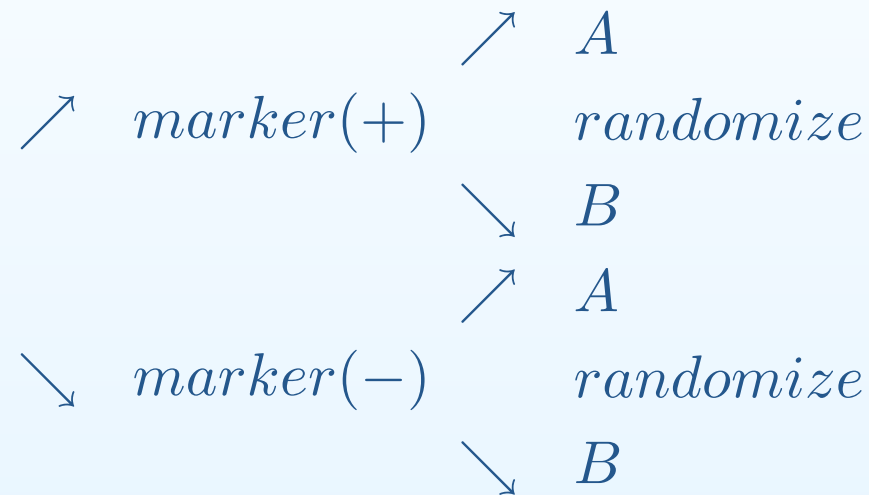
Clinical trial designs for Predictive Marker Validation

- A predictive marker predicts the clinical benefits from a specific therapy based on the marker status. In other words, only a subgroup of the patients may benefit from a particular treatment. This indicates treatment-marker interaction in statistical terms.
- Predictive markers are considered valid if they help improve the treatment efficacy, or help decrease toxicity. Identifying such valid predictive markers will be useful for a targeted group who can benefit from the right treatment and improve their quality of life essentially.

The marker by treatment interaction design

The marker by treatment interaction design (Sargent, 2005)

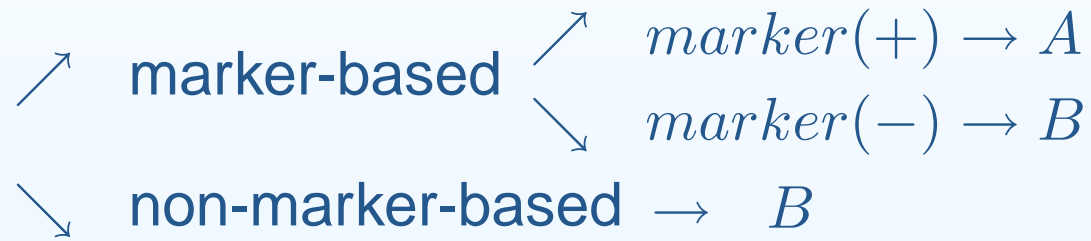
Register \rightarrow Test marker



The marker-based strategy design I, continued

The marker-based strategy design I (Sargent, 2005)

Register \rightarrow Randomize



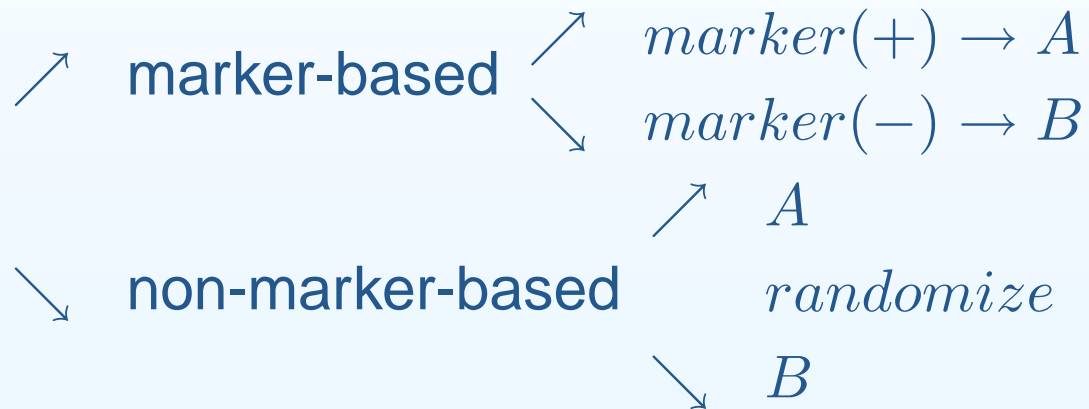
The marker-based strategy design I, continued

- The predictive value of the marker may be evaluated by comparing the outcome of all patients from the marker-based arm to the outcome of all patients from the non-marker-based arm.
- If the marker(+) group is at high risk of toxicity from the alternative treatment, then it is not ethical to design a clinical trial such that they have a chance to receive the alternative treatment.
- One limitation of this design is that we cannot assess if there is any true interaction between treatments and marker status.

The marker-based Strategy Design II

The marker-based strategy design II (Sargent, 2005)

Register \rightarrow Randomize



Sample Size Calculation

- The sample size might be dramatically different in a clinical trial conducted to validate a biomarker than a traditional clinical trial, also called the untargeted design, where participating patients are simply randomized to treatment or placebo.
- Assessment of the efficiency of the non-traditional clinical trial designs by comparing their sample sizes to the traditional design may provide better planning in clinical trials.

Set up

- For simplicity, we assume that the clinical outcome X is continuous. A single marker is used in the clinical trial design and the marker status is binary, measured with error.
- We calculate the sample size needed for a clinical trial to conduct the two sample T-test and the Wilcoxon rank sum test, a non-parametric test.
- The sample size is calculated based on a two-sided test with type-I error $\alpha = 0.05$, and type-II error $\beta = 0.20$ (power = 0.80).
- We presented the methods of sample size calculation for the marker-based strategy designs, which may provide direct assessment of the predictive value of a biomarker.

Notations

- Let T denote the treatment groups: $T = 1$ for treatment A, and $T = 0$ for treatment B.
- Let D denote the true underlying marker status: $D = 1$ for the group with positive status, and $D = 0$ with negative status.
- Let $\gamma = P(D = 0)$.
- Let μ_{0C} and μ_{1C} denote the mean response for the control group with marker status $D = 0$ and $D = 1$, respectively.
- Let μ_{0T} and μ_{1T} denote the mean response for the treatment group with marker status $D = 0$ and $D = 1$, respectively.
- We assume that each of the above four subgroups have the same variance, σ^2 .

Imperfect Assays

- Let R denote the marker status from an imperfect assay, where $R = 1$ for the group with positive marker status and $R = 0$ with negative status.
- $\lambda_{sens} = P(R = 1|D = 1)$: the sensitivity of the imperfect assay for diagnosing patients with positive status.
- $\lambda_{spec} = P(R = 0|D = 0)$: the corresponding specificity.

Imperfect Assays, continued

- Positive Predictive Value of the biomarker:

$$\omega_+ = P(D = 1|R = 1) = \frac{\lambda_{sens}(1 - \gamma)}{\lambda_{sens}(1 - \gamma) + (1 - \lambda_{spec})\gamma},$$

- Negative Predictive Value of the biomarker:

$$\theta_- = P(D = 0|R = 0) = \frac{\lambda_{spec}\gamma}{(1 - \lambda_{sens})(1 - \gamma) + \lambda_{spec}\gamma}.$$

The marker-based Strategy Design I

- We use the two sample t-test.
- In this design, all registered patients are randomized to either marker-based arm or non-marker-based arm.
- In the marker-based arm, since the patients are tested by an imperfect assay to determine their marker status first, the error of the assay will have impact on the sample size.
- The null hypothesis is $H_0 : \nu_m = \nu_n$, where ν_m and ν_n denote the mean response from the marker-based arm ($M = 1$) and the non-marker-based arm ($M = 0$), respectively.

The marker-based Strategy Design I, continued

$$\nu_m = [\mu_{1T}\omega_+ + \mu_{0T}(1 - \omega_+)][\lambda_{sens}(1 - \gamma) + (1 - \lambda_{spec})\gamma] +$$
$$[\mu_{1C}(1 - \theta_-) + \mu_{0C}\theta_-][(1 - \lambda_{sens})(1 - \gamma) + \lambda_{spec}\gamma],$$

$$\nu_n = \mu_{1C}(1 - \gamma) + \mu_{0C}\gamma.$$

The marker-based Strategy Design I, continued

Let τ_m^2 and τ_n^2 be the variance of the response for the marker-based arm and the non-marker-based arm, respectively. We obtain the following formula:

$$\tau_m^2 = Var(X|M = 1) = Var[E(X|M = 1, R) + E[Var(X|M = 1, R)],$$

where

$$\begin{aligned} & Var[E(X|M = 1, R)] \\ &= [\mu_{1T}\omega_+ + \mu_{0T}(1 - \omega_+) - \nu_m]^2[\lambda_{sens}(1 - \gamma) + (1 - \lambda_{spec})\gamma] + \\ & \quad [\mu_{1C}(1 - \theta_-) + \mu_{0C}\theta_- - \nu_m]^2[(1 - \lambda_{sens})(1 - \gamma) + \lambda_{spec}\gamma], \end{aligned}$$

and

$$\begin{aligned} E[Var(X|M = 1, R)] &= [(\mu_{1T} - \nu_{1T})^2\omega_+ + (\mu_{0T} - \nu_{1T})^2(1 - \omega_+) + \sigma^2]X \\ & \quad [\lambda_{sens}(1 - \gamma) + (1 - \lambda_{spec})\gamma] + \\ & \quad [(\mu_{1C} - \nu_{0C})^2(1 - \theta_-) + (\mu_{0C} - \nu_{0C})^2\theta_- + \sigma^2]X \\ & \quad [(1 - \lambda_{sens})(1 - \gamma) + \lambda_{spec}\gamma]. \end{aligned}$$

The marker-based Strategy Design I, continued

Hence, the total sample size required in the clinical trial is asymptotically given by

$$n_s = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2(\tau_m^2 + \tau_n^2)}{(\nu_m - \nu_n)^2},$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ denote the percentiles of the standard normal distribution.

Marker-based strategy design II

- In the non-marker-based arm, patients are further randomized to either treatment A or treatment B in this design. Again, the error of an imperfect assay will also have impact on the sample size.
- The null hypothesis is $H_0 : \nu_m = \nu_{nr}$, where ν_m and ν_{nr} denote the mean response from the marker-based arm ($M = 1$) and the non-marker-based arm ($M = 0$), respectively.
- Let τ_m^2 and τ_{nr}^2 denote the variance of the response for the marker-based arm and the non-marker-based arm, respectively.

Marker-based strategy design II, cont

- The mean ν_m and variance τ_m^2 in the marker-based arm in this design are the same as the ones in Marker-based strategy design I.
- However, the mean ν_{nr} and the variance τ_{nr}^2 in the non-marker-based arm are different here.

$$\nu_{nr} = \frac{(1 - \gamma)(\mu_{1T} + \mu_{1C}) + \gamma(\mu_{0T} + \mu_{0C})}{2},$$

$$\tau_{nr}^2 = \frac{[\mu_{1T}(1 - \gamma) + \mu_{0T}\gamma - \nu_{nr}]^2 + [\mu_{1C}(1 - \gamma) + \mu_{0C}\gamma - \nu_{nr}]^2 + \tau_T^2 + \tau_C^2}{2},$$

$$n_r = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2(\tau_m^2 + \tau_{nr}^2)}{(\nu_m - \nu_{nr})^2},$$

where n_r is the total sample size required to reject the $H_0 : \nu_m = \nu_{nr}$.

Wilcoxon Rank Sum Test

- Since the t -test method for estimating the sample size assumes that the test statistic, the difference in means, is distributed normally, it may not be correct for small sample sizes.
- In cases for which we expect a small sample size or in which we expect the outcome to have a heavy-tailed distribution, it may be better to take a nonparametric approach to estimating the sample size.
- We also derive sample sizes based on the Wilcoxon rank sum test.
- In keeping with the two-sided hypothesis testing approach of the t -test method, we examine here a two-sided test that the sum of the ranks of the outcomes in each arm are the same.

Wilcoxon Rank Sum Test, continued

- Let X_F and X_G denote the outcomes from the two arms of the trial. Call F and G the cumulative distribution functions of X_F and X_G , respectively. The null hypothesis we wish to test is that the outcomes from the two arms arise from a common distribution, that is, $H_0 : F = G$, against the alternative $H_a : F \neq G$.
- Assuming an equal sample size, n , in each of the two arms of the trial, the Mann-Whitney statistic is given by $W = \sum_{i=n+1}^{2n} R_i - \frac{1}{2}n(n+1)$, where the R_i are the ranks of the observations in the second arm of the trial. We reject the null hypothesis if

$$\frac{W - n^2 p_1}{\sqrt{\text{Var}(W)}}$$

is too extreme.

Wilcoxon Rank Sum Test, continued

- Let

$$p_1 = P(X_F < X_G), p_2 = P(X_F < X_{G_1}, X_F < X_{G_2}),$$

$$p_3 = P(X_{F_1} < X_G, X_{F_2} < X_G).$$

It can be shown that

$$\text{Var}(W) = n^2[p_1(1 - p_1) + (n - 1)(p_2 + p_3 - 2p_1^2)].$$

Given the sample size of n in each arm, the power of the Wilcoxon rank sum test, with a continuity correction, is given approximately by [9]:

$$1 - \Phi \left(\frac{\frac{1}{2}n^2 + z_{1-\alpha/2} \sqrt{\frac{n^2(2n+1)}{12}} - \frac{1}{2} - n^2 p_1}{\sqrt{\text{Var}(W)}} \right), \quad (1)$$

where Φ is the cumulative standard normal distribution function. We can estimate p_1 , p_2 and p_3 stochastically using Monte-Carlo integration.

Comparison of Efficiency

- Using the t -test and the Wilcoxon rank sum test, we compare the efficiency of each of the two marker-based strategy designs relative to the traditional design with regard to the number of patients needed to be randomized for each design to have specified values of type I error and power to detect a given difference in means.
- We evaluate this quantity, the ratio of the number of patients involved in a trial with the traditional design to the number in the alternative design, as a function of the sensitivity and specificity of the marker assay, the size of the treatment effect, and the prevalence of the marker in the population of interest.
- Since the results from the two-sample t -test and from the Wilcoxon rank sum test were very similar, we present only the results from the former.

Settings for numerical studies

- We consider two scenarios: (1) there is a treatment effect in patients with positive marker status ($D+$ group), but no treatment effect in those with negative marker status ($D-$ group); and (2) there is a treatment effect in the group with negative status, and the treatment effect is half of the treatment effect in the group with positive status. In the both cases, we assume that the within-patient variance of response, σ^2 , is the same for each of the four subgroups.

- The scenarios we consider are:

$$\text{Scenario (1) : } \mu_{1T} = 1, \mu_{1C} = 0, \mu_{0T} = 0, \mu_{0C} = 0, \sigma^2 = 1;$$

$$\text{Scenario (2) : } \mu_{1T} = 1, \mu_{1C} = 0, \mu_{0T} = \frac{1}{2}, \mu_{0C} = 0, \sigma^2 = 1.$$

- Since the marker assay is not perfect in general, we consider various levels of quality for the marker assay: sensitivities and specificities of 1.0, 0.8, and 0.6.
- We consider values of the marker prevalence $(1 - \gamma)$ between 0% and 100%, at 5% intervals.

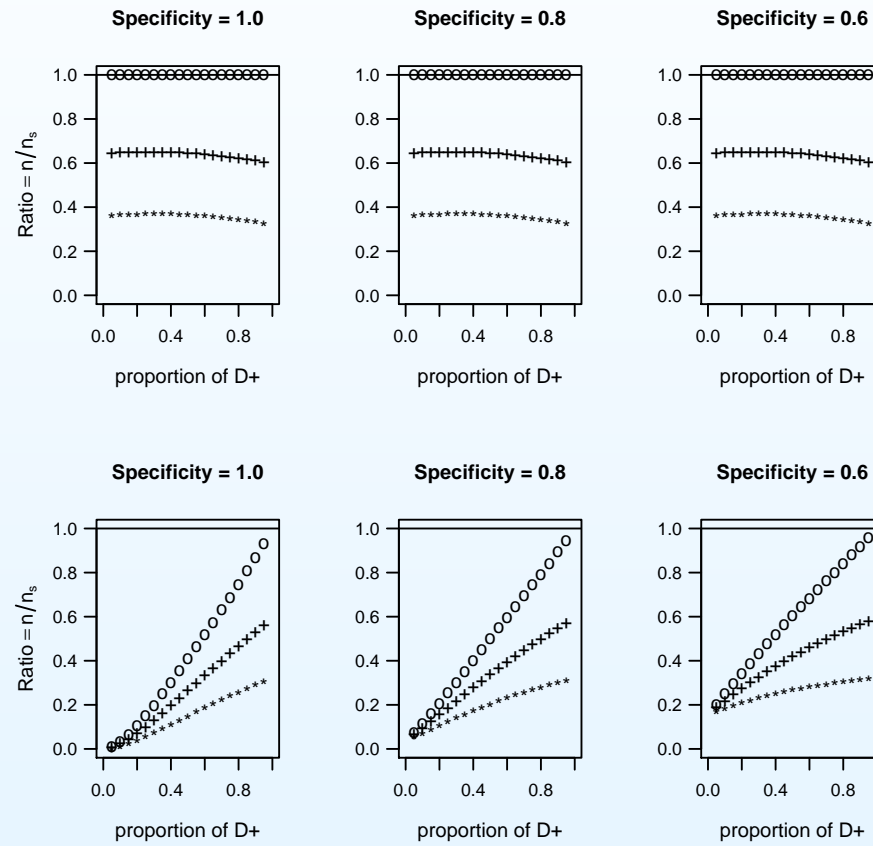
Some cautionary notes

- Since for each of the marker-based strategy designs the efficiency comparison is based on two hypothesis tests that are different in nature, the plots and observations presented are meant as guidelines only.

Result I - Captions

- Ratio of number randomized for untargeted versus marker-based strategy design WITHOUT randomization in non-marker-based arm (n/n_s).
- Upper panel: Scenario 1: No treatment effect in D- group.
- Lower panel: Scenario 2: Treatment effect in D- group is 1/2 of the treatment effect in D+ group. o Sensitivity=1.0; + Sensitivity=0.8; * Sensitivity=0.6

Result I - Marker-based Strategy Design I



Result I - continued

- We may see that if there is no treatment effect in the D- group, the marker-based strategy design without randomization in non-marker-based arm is as efficient as the untargeted design when the sensitivity is 1.0 regardless of the specificity and the proportion of D+ patients.
- When the sensitivity is less than 1.0, the marker-based strategy design without randomization in non-marker-based arm becomes less efficient. The lower the sensitivity, the less efficient this design becomes.
- Given a fixed sensitivity, the efficiency decreases slightly as the proportion of D+ patients becomes larger. However, changes in the specificity do not seem to have impact on the efficiency at all. Therefore, as long as the sensitivity of the assay is relatively high, using this design will not lose much efficiency compared to the untargeted design.

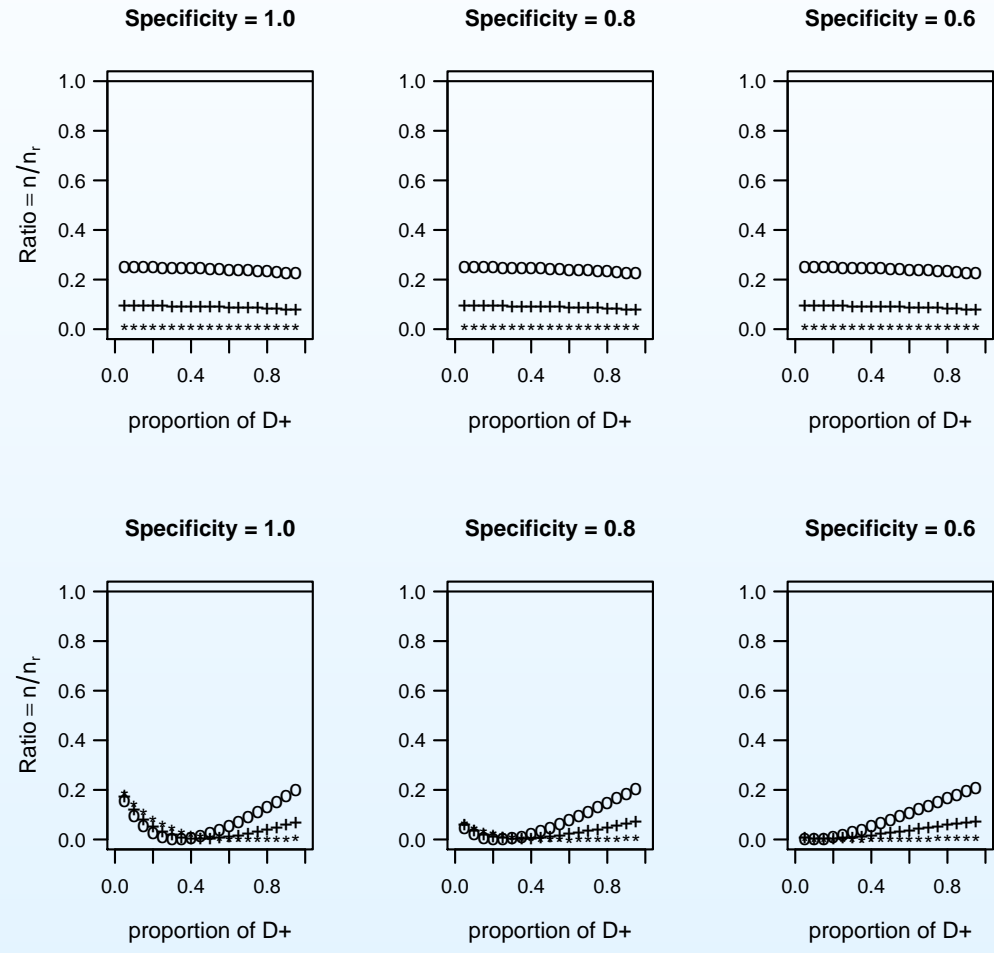
Result I - continued

- When the treatment effect in the D- group is half of the treatment effect in the D+ group, the assay sensitivity continues to play an important role, with a significant impact when the proportion of the D+ patients is large.
- On the the hand, changes in the specificity do have impact on the efficiency; as the specificity decreases, the efficiency increases. This is because some D- patients with positive test status help increase the mean response of the marker-based arm, thus the treatment effect between these two arms increases, and the corresponding sample size decreases.
- In addition, as the proportion of the D+ patients increases, so does the efficiency. As a result, when considering sample size, a clinical trial using this design is recommended if the proportion of the D+ patients and the sensitivity are high.

Result II - caption

- Ratio of number randomized for untargeted versus marker-based strategy design WITH randomization in non-marker-based arm (n/n_s).
- Upper panel: Scenario 1: No treatment effect in D- group.
- Lower panel: Scenario 2: Treatment effect in D- group is 1/2 of the treatment effect in D+ group. o Sensitivity=1.0; + Sensitivity=0.8; * Sensitivity=0.6

Result II - graphs



Result II - continued

- In general, the marker-based strategy design with randomization in non-marker-based arm is much less efficient than the untargeted design regardless of the proportion of the D+ group, the sensitivity and specificity of the assay.
- When there is no treatment effect in the D- group, the patterns of the figure look similar to the previous design except that the efficiency is much lower. This is expected since additional randomization is also performed in the non-marker-based arm, and much more patients are needed for the trial.
- In addition, the efficiency decreases slightly as the proportion of the D+ patients increases even though the sensitivity is 1.0. If the goal of the study requires us to use this design, then the sensitivity of the assay needs to be very high in order to be feasible to conduct such a huge clinical trial.

Result II - continued

- When the treatment effect in the D- group is half of the treatment effect in the D+ group, the efficiency might become non-monotone. This is because we cannot guarantee that the mean response in the marker-based arm is always bigger than the mean response in the non-marker-based arm. Which arm has a bigger treatment effect and the magnitude of the treatment effect depends on the proportion of the D+ patients, the assay sensitivity and specificity.
- When the proportion of the D+ patients is high, the efficiency of this design increases as the proportion increases; the efficiency also increases as the sensitivity increases regardless of its specificity. When the proportion of the D+ patients is low and the specificity is high, an increase in sensitivity actually decreases the efficiency although the impact of sensitivity is not big; an increase in the proportion of the D+ patients also decreases efficiency.
- In order to be feasible to conduct a trial using this design, the sensitivity and proportion of the D+ patients need to be very high. An additional feasible case is when the specificity is very high and the proportion of the D+ patients is very low. However, when the proportion of the D+ patients is very low or very high, it might not be worthwhile to pursue and validate the predictive value of the biomarker.

Summary and recommendation

- We have assessed the efficiency of the marker-based strategy designs compared to the untargeted design. The sample size of the marker-based strategy designs involved with biomarker status test by an imperfect assay are influenced by three factors with addition to the usual parameters in a regular sample size calculation. The three additional factors are: (1) the proportion of the D+ patients, (2) the sensitivity of the assay and (3) the specificity of the assay.
- The marker-based strategy designs are less efficient than the untargeted design in general. If there is no treatment effect in the D- group, it is still feasible to use the design without randomization in the non-marker-based arm if the assay sensitivity is high. If the treatment effect in the D- group is half of the effect in the D+ group, the proportion of the D+ patients needs to be relatively high and the sensitivity of the assay needs to be very high. The design with randomization in the non-marker-based arm is not recommended in general since its efficiency is very low compared to the untargeted design.

Future research

- There are lots of other potential areas for further research.
- First, we considered continuous clinical endpoints only; however, binary data and survival data are also used in clinical trials. Thus, the efficiency of the trials for such data may be a possible research area in the future.
- Second, only binary marker status was used in this paper, but categorical or continuous marker values may be considered.
- Third, we only considered a single marker in the clinical trial designs. However, in practice, multiple markers may be used to predict clinical outcomes. Fourth, the simulation studies assumed equal variance among the four subgroups, and possible different variances may be used in further simulation studies.
- Fifth, we did not include additional covariates in our model though we do not expect our findings will be substantially different with the presence of covariates.

Acknowledgements

- Kim Young
- Amy Laird