Published in final edited form as:

Stat Med. 2015 November 30; 34(27): 3503–3515. doi:10.1002/sim.6564.

Designing a Study to Evaluate the Benefit of a Biomarker for Selecting Patient Treatment

Holly Janes^{1,2,*}, Marshall D Brown¹, and Margaret S Pepe^{1,2}

¹Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

²University of Washington, Seattle, Washington, USA

Abstract

Biomarkers that predict the efficacy of treatment can potentially improve clinical outcomes and decrease medical costs by allowing treatment to be provided only to those most likely to benefit. We consider the design of a randomized clinical trial in which one objective is to evaluate a treatment selection marker. The marker may be measured prospectively or retrospectively using samples collected at baseline. We describe and contrast criteria around which the trial can be designed. An existing approach focuses on determining if there is a statistical interaction between the marker and treatment. We propose three alternative approaches based on estimating clinically relevant measures of improvement in outcomes with use of the marker. Importantly, our approaches accommodate the common scenario in which the marker-based rule for recommending treatment is developed with data from the trial. Sample sizes are calculated for powering a trial to assess these criteria in the context of adjuvant chemotherapy for the treatment of estrogen-receptor-positive, node-positive breast cancer. In this example, we find that larger sample sizes are generally required for assessing clinical impact than for simply evaluating if there is a statistical interaction between marker and treatment. We also find that retrospectively selecting a case-control subset of subjects for marker evaluation can lead to large efficiency gains, especially if cases and controls are matched on treatment assignment.

Keywords

treatment selection; biomarker; randomized controlled trial; study design; statistical interaction; predictive marker

1 Introduction

Markers that predict treatment efficacy hold great potential for improving clinical outcomes and decreasing medical costs. If a marker can identify which patients are likely to benefit from a treatment, provision of treatment can be restricted to this subgroup. Such an approach has the benefit of sparing the remaining subjects unnecessary, and potentially toxic or costly therapy. Treatment selection markers, sometimes called “predictive” [1–3] or “prescriptive” [4] markers, are of particular interest in contexts such as cancer treatment where treatment

* Address correspondence to: Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2 C200, Seattle, Washington 98109 USA, phone: 206 667 6353, fax: 206 667 4378, hjan@fhcc.org.

effects are thought to be heterogeneous, and where the toxicity or cost of the treatment is high so that only subgroups with sufficiently large treatment effects should be recommended it. However, these markers need rigorous evaluation before they can be applied in clinical practice to ensure that selective treatment improves outcomes and that the right patients are being recommended treatment.

There is a large literature on study designs related to treatment selection markers. Potential designs include “enrichment” designs that only enroll biomarker-positive subjects who are thought likely to respond to treatment, and “biomarker strategy” designs that randomize participants to having treatment assignment determined by the marker or not. Randomized and controlled trials (RCTs) that simply compare treatment options can also be used (see [1–3,5–19] and [20] for a recent review). RCTs do not make use of the marker in the conduct of the trial but can retrospectively provide information about the value of candidate treatment selection markers. As discussed in [20], the various designs are suitable for addressing very different scientific questions. In this paper, we focus on the policy question: What is the benefit of using a marker to select patient treatment? This is in contrast to other questions such as “Is treatment effective in biomarker-positive subjects?” and “Does the biomarker predict treatment efficacy?” For answering the policy question, the RCT design is most appropriate [20].

The existing trial design methodology is limited. Most of the literature focuses on powering studies to test for a statistical interaction between the marker and treatment. While an interaction is necessary, it is insufficient for measuring marker performance [21,22]. We propose instead to focus on assessing the clinical impact of using the marker to assign treatment, using previously proposed measures of clinical impact [4,21,23–27]. We power studies to assess clinical impact with precision. Another shortcoming of existing methodology is that it focuses exclusively on the setting where a biomarker-based rule for treatment assignment— a biomarker “positivity rule”— exists in advance of the study. More often, there is considerable uncertainty about what positivity criterion should be used, and accordingly our design calculations accommodate a continuous marker where the data are used to identify a positivity rule and to evaluate its performance for treatment selection.

The structure of this manuscript is as follows. We set the context and describe relevant background in Section 2. Section 3 describes existing and new approaches to powering an RCT for evaluating a treatment selection marker. The approaches are illustrated and contrasted by designing a future trial to evaluate markers used for identifying women with estrogen-receptor-positive (ER+), node-positive breast cancer who are not likely to benefit from adjuvant chemotherapy, the current standard of care. Extensions of the approach are described in Section 4.

2 Context and Background

2.1 Setting

We consider the setting where there are two treatment options, for simplicity called “treatment” ($T = 1$) and “no treatment” ($T = 0$). The binary clinical outcome of interest, denoted by D , indicates that a bad event (e.g. death, disease recurrence, or a composite

outcome incorporating treatment-associated toxicity) has occurred following the provision of treatment/no treatment. For didactic reasons we assume that treatment is the standard of care and that it is beneficial marginally in the sense that $P(D = 1|T = 1) \equiv \rho_1 < P(D = 1|T = 0) \equiv \rho_0$. The hope is that a marker, Y , will be useful for identifying a subgroup of subjects for whom treatment can be avoided. The opposite scenario where $T = 0$ is the default approach or standard of care and Y is used to identify a subgroup to treat can be handled analogously and details are provided in Section 4.

Let the marker, Y , be either a single measurement or a pre-specified score that combines multiple marker measurements. Methods for developing the score variable within the study being planned are discussed in Section 5. We assume that Y is continuous and that the study will be used to determine a rule based on Y to recommend for or against treatment. We call the treatment assignment rule the “marker-positivity” rule. The simpler setting where Y is binary (or where a pre-specified treatment rule exists) is accommodated easily and is discussed in Section 4.

2.2 Motivating clinical context

For illustration we consider the breast cancer treatment context. Women diagnosed with ER + and node-positive breast cancer are typically treated with both hormone therapy (e.g. tamoxifen) and adjuvant chemotherapy following surgery. This is despite the fact that it is generally well-accepted that only a subset of these women benefit from the adjuvant chemotherapy, and the remaining women only suffer its toxic side effects, not to mention the burden and cost of unnecessary treatment [28]. A high public health priority is to identify markers that can be used to predict which women are and are not likely to benefit from the adjuvant chemotherapy [29]. The Oncotype DX recurrence score is currently being used in clinical practice for this purpose. The marker is a proprietary combination of 21 genes whose expression levels are measured in the tumor tissue obtained at surgery [30–32]. It has been shown to have some value for identifying a subset of women who are unlikely to benefit from chemotherapy [31,32]. The question we consider is how to design a future study to further evaluate the benefit of using the Oncotype DX marker, or other potentially better-performing markers, to identify women who can forego adjuvant chemotherapy.

2.3 Measuring performance of a treatment selection marker

The first step in measuring marker performance is to specify a rule for recommending no treatment on the basis of marker value, i.e. to propose the marker-positivity criterion. Let

$(Y) = P(D = 1|T = 0, Y) - P(D = 1|T = 1, Y)$ be the absolute treatment effect given a subject’s marker value Y . The optimal rule for assigning treatment is to recommend treatment to those for whom $(Y) \geq 0$ and to recommend against treatment if $(Y) < 0$ [24,25,33]. We write this rule as $A(Y) = I((Y) \geq 0)$. That is, $A(Y) = 0$ subjects with $(Y) < 0$ are called “marker-negative” and recommended $T = 0$. Analogously, $A(Y) = 1$ indicates marker-positivity and corresponds to a recommendation for treatment. In the breast cancer context, where D is an indicator of breast cancer recurrence or death, $A(Y) = 0$ subjects have a higher recurrence/death rate with chemotherapy than without. For these subjects chemotherapy may increase the non-cancer mortality rate to an extent that it is not offset by a reduction in the recurrence rate— they should be recommended to forego chemotherapy.

A variety of statistical measures are useful for evaluating the performance of this treatment rule (see [27] for a summary):

- Proportion of subjects that can forego treatment,

$$P_{neg} = P(A(Y)=0)$$

- Average benefit of foregoing treatment among marker-negatives,

$$B_{neg} = P(D=1|T=1, A(Y)=0) - P(D=1|T=0, A(Y)=0)$$

- Average benefit of treatment among marker-positives,

$$B_{pos} = P(D=1|T=0, A(Y)=1) - P(D=1|T=1, A(Y)=1)$$

- Improved outcomes, as measured by the decrease in the population event rate, under marker-based treatment assignment,

$$\begin{aligned} \Theta &= [P(D=1|T=1)] - [P(D=1|T=1, A(Y)=1)P(A(Y)=1) + P(D=1|T=0, A(Y)=0)P(A(Y)=0)] \\ &= [P(D=1|T=1, A(Y)=0) - P(D=1|T=0, A(Y)=0)] P(A(Y)=0) \\ &= B_{neg} \cdot P_{neg}, \end{aligned}$$

where we define $P(D=1|T, A(Y)=0) = 0$ if $P(A(Y)=0) = 0$.

The measure Θ is widely advocated as a global measure of treatment selection marker performance [4,21,23–27]. Θ varies between 0 for an entirely useless marker, for which there is no difference in the event rate with and without use of the marker, and $\Theta = \rho_1$ for a perfect marker that yields an event rate of zero under marker-based treatment. The constituents of Θ , namely B_{neg} and P_{neg} , are informative and helpful for dissecting the impact of the marker. The measures B_{neg} and B_{pos} quantify the average benefit of the treatment policies recommended to marker-negative and marker-positive subjects, respectively.

Janes et al. [27] describe methods for estimating and making inference about these measures, given data from an RCT comparing $T=0$ to $T=1$ (or a design-based subsample thereof). (Y) is estimated by using a parametric risk model with parameters γ , $P(D=1|T, Y; \gamma) = f(T, Y; \gamma)$, and calculating $\hat{(Y)} = P(D=1|T=0, Y; \gamma) - P(D=1|T=1, Y; \gamma)$. “Empirical” and “model-based” estimates of the marker performance measures have been proposed. Empirical estimates, indicated by e , use the risk model to identify marker-positive and marker-negative subjects, but estimate the performance of the treatment assignment rule empirically. Model-based estimates, indicated by m , estimate performance using modelled estimates of (Y) . If $\hat{P}$ denotes an empirical probability estimate and $\hat{E}$ denotes an empirical mean,

$$\begin{aligned}
\hat{P}_{neg} &= \hat{P}(\hat{\Delta}(Y) < 0) \\
\hat{B}_{neg}^e &= \hat{P}(D=1|T=1, \hat{\Delta}(Y) < 0) - \hat{P}(D=1|T=0, \hat{\Delta}(Y) < 0) \\
\hat{B}_{neg}^m &= \hat{E}(-\hat{\Delta}(Y) | \hat{\Delta}(Y) < 0) \\
\hat{B}_{pos}^e &= \hat{P}(D=1|T=0, \hat{\Delta}(Y) > 0) - \hat{P}(D=1|T=1, \hat{\Delta}(Y) > 0) \\
\hat{B}_{pos}^m &= \hat{E}(\hat{\Delta}(Y) | \hat{\Delta}(Y) > 0) \\
\hat{\Theta}^e &= \hat{B}_{neg}^e \cdot \hat{P}_{neg} \\
\hat{\Theta}^m &= \hat{B}_{neg}^m \cdot \hat{P}_{neg}
\end{aligned}$$

Empirical estimators are less efficient but also less reliant on risk model assumptions than model-based estimators [23,25,27]. Software for estimation and inference can be found at <http://labs.fhcrc.org/janes/index.html>.

3 Designing a marker evaluation study

3.1 RCT design

Consider the task of determining N , the total sample size of an RCT that is needed to evaluate a candidate marker's performance. We assume a fixed treatment allocation probability, $\lambda(\lambda N = N_1$ assigned $T = 1$, $(1 - \lambda)N = N_0$ assigned $T = 0$); commonly $\lambda = 0.5$ corresponding to equal treatment allocation. Primary objectives typically drive the original choice of N , but when assessing marker performance is a secondary or even an exploratory study objective, the following criteria, explained below, may motivate a larger N in order to ensure that treatment selection markers can also be evaluated given data from the trial.

Criterion 1 (Power for Interaction) α_1 -level test of H_0 : no interaction between marker and treatment has power $1 - \beta_1$ given a specified marker-by-treatment interaction;

Criterion 2 (Detecting Improved Outcomes) Lower bound (LB) of the $1 - \alpha_2\%$ CI for Θ lies above 0 with high probability: $P(\text{LB} > 0 \mid \Theta = \Theta_a, H_0 \text{ rejected}) = 1 - \beta_2$ under design alternative Θ_a ;

Criterion 3 (Precise Estimation of Improved Outcomes) Θ is estimated with sufficient precision: $P(|\hat{\Theta} - \Theta_a| \leq \varepsilon_1 \mid \Theta = \Theta_a, H_0 \text{ rejected}) = 1 - \beta_3$ for a specified $\varepsilon_1 > 0$;

Criterion 4 (Errors in Treatment Rule) Largest treatment effect among “estimated marker-negatives” is sufficiently small: for $Y_t = \arg \max_{Y: \hat{\Delta}(Y) < 0} \hat{\Delta}(Y)$, $P(\hat{\Delta}(Y_t) < \varepsilon_2 \mid \Theta = \Theta_a, H_0 \text{ rejected}) = 1 - \beta_4$ for specified $\varepsilon_2 > 0$.

Criterion (1) is used to ensure adequate power to determine whether the marker has any value at all for treatment selection. A statistical interaction is necessary but not sufficient for good marker performance, so concluding that a statistical interaction exists does not imply that clinical outcomes are improved with use of the marker. Criterion (2) directly addresses improvements in outcomes with use of the marker. It ensures that $\Theta = 0$ can be ruled out with high probability. Criterion (3) may be more stringent (depending on ε_1), requiring precise estimation of the extent of improved outcomes. Criterion (4) addresses the uncertainty in the estimated marker-positivity threshold. However, rather than focusing on the marker threshold itself, (4) requires that the consequences of being designated marker-

negative are appropriate. Specifically, recall that the default choice is $T = 1$ and the goal is to identify subjects who do not benefit from treatment. Criterion (4) considers the subgroup of “estimated marker-negative” subjects whose treatment recommendations change from $T = 1$ to $T = 0$ following marker measurement. It requires that the largest true benefit of treatment ($\Delta(Y)$) in this subgroup is sufficiently small, in order to ensure that subjects are not recommended $T = 0$ if in fact there is a meaningful benefit of $T = 1$ for them.

Note that, for criteria (2)–(4), we condition the probability statement on rejecting H_0 : no marker-by-treatment interaction. This is because inference about our measures of marker performance is problematic under H_0 [27]. Therefore in practice we recommend only evaluating marker performance if H_0 is rejected, and accordingly we design studies to allow assessment of marker performance given that H_0 is rejected (criteria (2)–(4)). Our acknowledgement of the conditioning is somewhat unique; commonly analysts ignore the hypothesis testing that determines whether they proceed to estimate effect sizes. The problematic performance under H_0 also implies that criterion (2) should not be characterized as a hypothesis test; rather it simply assesses whether the $(1 - \alpha_2)\%$ CI rules out $\Theta = 0$. (Janes et al. [27] verified proper coverage of empirical and model-based confidence intervals for Θ , both marginally and conditional on H_0 rejection.) We also ensure that studies have adequate power both to reject H_0 and to assess marker performance, i.e. to satisfy criterion (1) and criterion (k) for $k = 2, 3, 4$.

Importantly, we advocate verifying the power of the design for testing and estimating treatment effects in marker-positive and marker-negative subgroups. Such analyses justify the recommendation for treatment among marker-positives and against treatment among marker-negatives. Typically, designs meeting criteria (1)–(4) have ample power for these subgroup analyses, as demonstrated in our example below.

For testing H_0 , we use likelihood ratio tests of marker-by-treatment interaction in the context of a parametric risk model—logistic regression in our examples. For evaluating Θ , we explore both empirical and model-based estimators and use the percentile bootstrap to generate CIs. The marker-positivity threshold is estimated as $Y_t = \max(Y: \hat{\Delta}(Y) < 0)$ since the risk functions we assume yield $\Delta(Y)$ that cross 0 only once. We use simulations to determine the sample sizes required to satisfy criteria (1)–(4).

3.1.1 Illustration—We illustrate the approach by designing a future study to evaluate the Oncotype DX marker, denoted by Y_1 , and two hypothetical markers, one weak (although stronger than Oncotype DX; Y_2), and one very strong (Y_3). The phase 3 trial SWOG SS8814, which compared adjuvant chemotherapy plus tamoxifen versus tamoxifen alone for treating post-menopausal women with ER+, node-positive breast cancer [32,34], is used as the basis for the simulations. The primary outcome, D , is an indicator of 5-year breast cancer recurrence or death, and the marginal treatment effect is $\rho_0 - \rho_1 = 0.24 - 0.21 = 0.03$. In our simulations, N patients are enrolled and randomized with equal probability to tamoxifen alone ($T = 0$) or tamoxifen plus chemotherapy ($T = 1$). Y_1 mimics the Oncotype DX distribution [32]: $\sqrt{Y_1}$ is normally distributed with mean 4.8 and standard deviation 1.8. Y_2 and Y_3 follow standard normal distributions. Each marker is related to D via a linear logistic regression model,

$$\text{logit } P(D=1|T, Y) = \gamma_0 + \gamma_1 T + \gamma_2 Y + \gamma_3 YT. \quad (1)$$

The parameters of the risk models are constrained to produce the marginal event rates $\rho_0 = 0.24$ and $\rho_1 = 0.21$ seen in SS8814 [32]. Figure 1 shows the risk of 5-year recurrence or death given tamoxifen alone or tamoxifen plus chemotherapy as a function of marker value, and the distribution of marker-specific treatment effect (γ_3), for each of the three markers.

In analyzing data from each simulated trial we fit the linear logistic regression model (1). The design alternatives for Y_1 are a marker-by-treatment interaction of $\gamma_{3a} = -0.0192$, and, much more interpretable, an associated 1% reduction in the event rate under marker-based treatment ($\Theta_a = 0.01$). The design alternatives for Y_2 are $\gamma_{3a} = -0.5$ and $\Theta_a = 0.02$, and for Y_3 they are $\gamma_{3a} = -1.5$ and $\Theta_a = 0.08$. We consider the following design criteria:

Criterion 1 80% or 90% power for 0.05-level test of $H_0: \gamma_3 = 0$ under $\gamma_3 = \gamma_{3a}$;

Criterion 2 LB of the 95% CI for Θ lies above 0 with 80% or 90% probability: $P(\text{LB} > 0 \mid \Theta = \Theta_a, H_0 \text{ rejected}) = 0.80 \text{ or } 0.90$;

Criterion 3 Θ is estimated to within 50% of its true value with 80% or 90% probability: $P(|\hat{\Theta} - \Theta_a| \leq 0.5 \cdot \Theta_a \mid \Theta = \Theta_a, H_0 \text{ rejected}) = 0.80 \text{ or } 0.90$;

Criterion 4 Largest treatment effect among estimated marker-negatives is sufficiently small with 80% or 90% probability: $P(\gamma_l < 0.015 \mid \Theta = \Theta_a, H_0 \text{ rejected}) = 0.80 \text{ or } 0.90$.

Estimating Θ to within, say, 20% of its true value could be entertained as a more stringent criterion (3). The 0.015 cutoff for γ_l in criterion (4) is selected as half the marginal treatment effect, and a magnitude of treatment effect below which avoiding chemotherapy may be a reasonable treatment recommendation. We emphasize again that criteria (2)–(4) are all assessed conditional on rejection of H_0 , i.e. for a given sample size we calculate the fraction of simulated trials where the criterion is satisfied among those where H_0 is rejected.

We simulate 2,500 trials with sample sizes ranging from 50 to 10,000 in increments of 25, and for each sample size calculate the proportion of trials for which each design criterion is satisfied. Thus the Monte Carlo error for each estimated proportion is at most 0.01

$(\max_p \sqrt{p(1-p)/2500} = 0.01)$. To obtain sample sizes ensuring criteria are satisfied with 80 or 90% probability we interpolate between sample sizes using loess smoothing (span = 0.4) and round the resultant sample size up to the nearest 25.

Table 1 shows the sample sizes required to satisfy each design criterion for the weak simulated marker, Y_2 . Using model-based estimation, we require $N = 1050, 1050, 4675$, and 5225 to satisfy criteria (1)–(4) with 90% probability, respectively. While the sample size is smallest for satisfying criteria (1) and (2), it is important to note that if the trial is sized only to satisfy these criteria estimates of Θ will be imprecise. For example, the $N = 1050$ participant study yields an expected 95% CI for Θ of 0.004 to 0.0524 when $\Theta = 0.02$ (the true value for Y_2). Moreover, errors in treatment recommendation are likely: with $N = 1050$ there is only an 0.69 probability that the largest treatment effect among estimated marker-

negatives is below 0.015 (criterion (4)). In other words, there is a 31% chance that the treatment rule derived from the trial will inappropriately recommend no treatment to a group of subjects that could benefit meaningfully from it.

Also shown in Table 1 are the sample sizes required to satisfy criterion (1) and criterion (k) for $k = 2, 3, 4$. We see, for example using model-based estimation and targeting 90% power, that a somewhat larger sample size ($N = 1375$ vs $N = 1050$) is required both to test for an interaction and to rule out $\Theta = 0$. However the largest required sample size does not increase when considering the criteria jointly ($N = 5225$).

Importantly, a trial sized to satisfy criteria (1)–(4) also has ample power to evaluate average treatment effects in each marker subgroup. For example, using the $N = 5225$ design and model-based estimation the expected 95% CI for B_{neg} is 0.03 to 0.06, a fairly precise estimate of the true value of B_{neg} which is 0.05 for Y_2 . Similarly, for B_{pos} the expected 95% interval is 0.07 to 0.11 around the true value of 0.09. In addition, 100% of trials have a lower limit of the 95% CI for B_{pos} above 0, and 100% have a lower limit for B_{neg} above 0— the efficacy of chemotherapy among marker-positives is reliably established, as is the superiority of no chemotherapy for marker-negatives.

Table 1 reveals the loss of efficiency associated with empirical estimation of Θ [23,25,27]. Sample sizes satisfying criterion (3) are 18–34% larger, and those satisfying criterion (2) are substantially larger if the empirical estimator is used.

Table 1 also shows results for the strong simulated marker, Y_3 . As expected, sample sizes are generally smaller for this marker, with one exception: the N required to satisfy criterion (4) is slightly larger. This is a consequence of the fact that for this marker the treatment effect function, (Y) , has greater variation in Y (see Figure 1). Therefore, error in the marker-positivity threshold has greater consequence. The largest sample size continues to be that satisfying criterion (4), while rejecting H_0 and ruling out $\Theta = 0$ requires the smallest sample size. Here again, an $N = 150$ participant study sized to test H_0 with 90% power would yield a very wide expected 95% CI for Θ using model-based estimation, from 0.024 to 0.176 when the true value of Θ is 0.08. Errors in treatment recommendation would be more likely than not, with only a 0.39 probability that the largest treatment effect among estimated marker-negatives is below 0.015 (criterion (4)).

For the Oncotype DX-like marker, Table 1 shows that sample sizes are much larger in general, a consequence of the relatively poor performance of this marker. More than 10,000 participants are required to satisfy criteria (1)–(4) with 80 or 90% probability using model-based estimation.

In the appendix, we describe the results of a sensitivity analysis for marker Y_2 that assesses, for the sample sizes required to satisfy each criterion with 90% probability under the original risk model parameters, the impact of varying the parameters over values consistent with the SS8814 trial data. We find that the probabilities of satisfying the design criteria vary from 70% to 100% when the risk model parameters span the interquartile range of their estimated distributions. Probabilities vary more for criteria (1)–(3) than for criterion (4).

In practice, we recommend requiring at least 80% power for criterion (1). Note that, in some settings (e.g. markers Y_2 , Y_3) criterion (2) may be satisfied with a smaller N , but this sample size is insufficient; there must be adequate power both to reject H_0 and to evaluate marker performance conditional on H_0 rejection (criterion (1) *and* criterion (2)). Criterion (3) may be used in addition to or instead of (2) to ensure that Θ is estimated with sufficient precision. Criterion (4) is compelling, ensuring that the estimated treatment rule does not yield meaningfully incorrect treatment recommendations, but satisfying (4) may require a substantially larger N as we observed in Table 1 for marker Y_3 . In many settings, while empirical estimation of Θ is appealing, the substantial increase in N may render the estimator impractical in applications. Finally, sensitivity analyses are warranted for studying the impact of varying the risk model parameters, and perhaps also the parametric risk model itself.

3.2 Case-control design nested within an RCT

Next we consider designing a case-control study nested within an RCT for marker evaluation. Specifically, we entertain retrospectively sampling a subset of subjects from the RCT and measuring Y only on them. Retrospective sampling could be based on case ($D = 1$) or control ($D = 0$) status, and potentially also on treatment assignment, T . Methods for estimating marker performance given nested case-control data are described in [27].

We determine the total N required for the RCT if the marker is measured on n subjects in the case-control sample. Let $N = N_{D0} + N_{D0^-} + N_{D1} + N_{D1^-}$ where N_{Dt} and N_{Dt^-} denote the numbers of cases and controls in treatment group t , $t = 0, 1$, respectively. The case-control designs we consider sample $n_{D0} = f_0 \cdot N_{D0}$ and $n_{D1} = f_1 \cdot N_{D1}$ cases in the $T = 0$ and $T = 1$ arms of the trial, for fixed proportions f_0 and f_1 , and $n_{D0^-} = k_0 \cdot n_{D0}$ and $n_{D1^-} = k_1 \cdot n_{D1}$ controls for fixed control:case ratios k_0 and k_1 . Candidate design criteria are as described above in Section 3.1.

3.2.1 Illustration—The breast cancer context is used for illustration, first to demonstrate that the case-control design can be very efficient if the outcome event $D = 1$ is rare. In particular it achieves nearly the same efficiency as a study that measures Y on the entire RCT cohort but uses a substantially smaller sample size. Table 2 shows the RCT sample sizes required to satisfy the design criteria of Section 3.1.1 with 90% probability for markers Y_2 and Y_3 . The table shows sample sizes under 5 possible designs: measuring Y for all subjects in the trial (“RCT cohort”); case-control designs that sample all cases and 1 or 2 controls per case (“1:1 case-control” and “1:2 case-control”), and designs that additionally match controls and cases on treatment assignment (“1:1 matched case-control” and “1:2 matched case-control”). For marker Y_2 , criteria (1)–(4) are satisfied with 90% probability with $N = 5225$ participants under the RCT cohort design, compared to just $n = 3625$ under 1:1 case-control sampling. However, to obtain $n = 3625$ participants in the case-control sample requires an RCT cohort of $N = 7975$. When controls are matched on treatment assignment to cases under 1:1 case-control sampling, only a slightly larger cohort is needed ($N = 5450$), and Y need only be measured on $n = 2469$ trial participants. Interestingly, matching on treatment has very little impact on the n required to satisfy criterion (1), a modest impact for criteria (2) and (3), and a large impact for criterion (4). Increasing from 1

to 2 matched controls per case requires measuring the marker on more subjects ($n = 3551$) but decreases the required size of the cohort back to the original $N = 5225$. For Y_2 , the 1:2 matched case-control design may be the best choice.

Similar results are seen for marker Y_3 , where we note that $N = 225$ is the smallest RCT cohort size that supports case-control sampling; with any smaller N it is not uncommon for there to be simulated datasets with no cases in one of the two treatment arms. Again we see that 1:2 matched case-control sampling may be a good design choice, requiring a slightly larger cohort than a design that measures the marker on all study participants ($N = 5450$ vs. $N = 5400$) but requires measuring the marker on about 2/3 of participants ($n = 3687$). In general, the tradeoff between the cost of measuring Y on a participant and the cost of enrolling a participant in the RCT must be weighed to choose the optimal design.

4 Extensions

4.1 When not treating is the default strategy

The design approach described above also applies to settings where the default treatment strategy is not to treat and the purpose of the marker is to identify a subset who benefit from treatment. For example, suppose that a new therapy is being efficacy-tested in the population of interest for the first time, and there is a hypothesis that only a subset of patients may benefit from it. The value of using Y to identify those who benefit is measured by:

- Proportion of subjects to whom treatment can be recommended,

$$P_{pos} = P(A(Y)=1) = P(\Delta(Y) \geq 0)$$

- Average benefit of not treating marker-negatives,

$$B_{neg} = P(D=1|T=1, A(Y)=0) - P(D=1|T=0, A(Y)=0)$$

- Average benefit of treating marker-positives,

$$B_{pos} = P(D=1|T=0, A(Y)=1) - P(D=1|T=1, A(Y)=1)$$

- Improvement in outcomes using marker-based treatment vs. the standard of not treating,

$$\begin{aligned} \Theta &= [P(D=1|T=0)] - [P(D=1|T=1, A(Y)=1)P(A(Y)=1) + P(D=1|T=0, A(Y)=0)P(A(Y)=0)] \\ &= [P(D=1|T=0, A(Y)=1) - P(D=1|T=1, A(Y)=1)] P(A(Y)=1) \\ &= B_{pos} \cdot P_{pos}, \end{aligned}$$

and the design criteria are identical with the exception of (4) which becomes:

Criterion 4 Smallest treatment effect among “estimated marker-positives” is sufficiently large: for $Y_t = \arg \min_Y: \hat{\gamma}_Y(0) - \hat{\gamma}_Y(1), P(\hat{\gamma}_Y(1) > \varepsilon_2 | \Theta = \Theta_a, H_0 \text{ rejected}) = 1 - \beta_4$ for specified $\varepsilon_2 < 0$.

4.2 Binary markers

If the marker of interest is binary, $Y = 0$ or 1 , the rule for withholding treatment is pre-defined by $A(Y) = I(Y = 0)$ (or perhaps $I(Y = 1)$) and the marker performance measures directly characterize the properties of this rule. The design criteria remain unchanged except that criterion (4), which relates to the choice of threshold, does not apply.

4.3 General treatment rules

The rule “do not treat if $(Y) < 0$ ” is only optimal if all effects of treatment are captured in the outcome, D . In some settings there may be additional downsides of treatment, for example toxicities, cost, or patient burden, that is not captured in D . In this case, the optimal rule is “do not treat unless the benefit is at least δ ”. In other words, the marker-positivity rule is $A(Y) = I((Y) \geq \delta)$ where $\delta > 0$ represents the burden of treatment relative to that of the event, $D = 1$ [33,35]. The measures of marker performance apply equally well to this rule and the design criteria also apply where criterion (4) becomes:

Criterion 4 Largest treatment effect among “estimated marker-negatives” is sufficiently small: for $Y_t = \arg \max_Y: \hat{Y} < \delta(Y), P((Y_t) < \delta + \varepsilon_2 | \Theta = \Theta_a, H_0 \text{ rejected}) = 1 - \beta_4$ for specified $\varepsilon_2 > 0$.

4.4 Optimal treatment allocation

An additional issue in the design of an RCT is the choice of treatment allocation probability, $\lambda = P(T = 1)$. Having identified the N required to satisfy the design criteria using equal treatment allocation ($\lambda = 0.5$), one can perform an additional set of simulations, fixing N at the required value and examining the probabilities of satisfying the design criteria as a function of λ . There may be settings in which unequal treatment allocation is advantageous. However, in our breast cancer example $\lambda = 0.5$ was found to be approximately optimal for all 4 design criteria as shown in Table 3. (Note that with $N = 450$ there is a slight advantage of $\lambda > 0.5$ for criterion (2); however this advantage disappears for larger values of N .)

5 Discussion

For randomized trials in which evaluating the utility of a candidate treatment selection marker is a study objective, this paper proposes several possible design criteria which can be used to guide the choice of trial sample size. Specifically, we consider the existing approach of ensuring adequate power to determine if the marker has a statistical interaction with treatment. This criterion is contrasted with other criteria based on estimating a clinically relevant measure of marker performance with sufficient precision, and with estimating the marker-positivity rule with enough accuracy to avoid meaningful errors in treatment recommendations. In our examples, we find that larger sample sizes are generally required for assessing clinical impact in addition to simply evaluating if there is an interaction between marker and treatment. Our design methods apply more broadly than existing methods because they focus on clinically relevant measures of marker performance, do not assume that a marker-positivity criterion exists in advance of the trial (and thus capture the uncertainty in the associated criterion), and allow for case-control sampling from within a randomized trial to identify participants for marker measurement.

In settings where a pre-specified marker-positivity rule exists in advance of the trial, it can be efficient to stratify the randomization in the RCT on marker-positivity (see [7,12,19,36,37]). This ensures balance between treatment arms in the distribution of marker-positivity, which is only guaranteed under unstratified randomization in large samples. Note that implementing stratified randomization requires measuring the marker at baseline on all trial participants.

We have assumed throughout that, in settings with multiple markers, a score combining the marker measurements exists in advance of the trial. If not, the trial could be used both to develop and to evaluate the score. However, to avoid over-optimism, part of the data would need to be set aside for score development, and the remainder saved for score evaluation. Alternatively, a cross-validation or bootstrapping procedure could be used to correct for over-optimism. Friedlin and Simon [5] have also developed methods for designing a trial in which a score is developed and evaluated, but importantly their design criteria differ from ours. Specifically, Friedlin and Simon propose to use data for a subset of subjects (the “stage 1” data) to develop a score. Next they test for an overall treatment effect using data from all subjects, and then only if the overall effect is not statistically significant do they test for a treatment effect in the marker-defined subset of subjects in the validation (“stage 2”) data. They define a “positive result” to be a significant treatment effect either overall or within the marker-specific subgroup. The study is then powered to detect this type of positive result. This analysis strategy does not answer the question as to the value of the marker for treatment selection, noting for example that a positive treatment effect overall does not rule out substantial heterogeneity in treatment effects as a function of marker value. Rather, it seems aimed at achieving a different objective of finding a subgroup of patients who can be recommended treatment.

A fundamental challenge in designing studies to evaluate treatment selection markers is that, unless markers are expected to be very strong predictors of treatment effect, very large sample sizes are needed to allow for marker evaluation. This unfortunate reality is evidenced by several trials currently underway in breast cancer: the MINDACT trial which enrolled 6,7000 women [38], the TAILORx trial which enrolled over 10,000 women [39], and the RxPONDER trial enrolling 4,000 women [40]. (Note that these trials were not designed to assess the criteria we propose here.) Meta-analyses of multiple randomized, controlled trials may be one solution to the sample size dilemma [41]. Group sequential “futility” monitoring for the marker-by-treatment interaction coefficient should also be entertained, in order to cap enrollment at the sample size required to satisfy primary study objectives in the event that interim data do not support the hypothesis that the marker predicts treatment efficacy.

Acknowledgments

Funded by: R01 CA152089 and R01 GM054438 from the National Institutes of Health

References

1. Sargent D, Allegra C. Issues in clinical trial design for tumor marker studies. *Seminars in Oncology*. 2002; 29(3):222–230. [PubMed: 12063675]

2. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*. 2004; 10:6759–6763. [PubMed: 15501951]
3. Simon R. Lost in translation: Problems and pitfalls in translating laboratory observations to clinical utility. *European Journal of Cancer*. 2008; 44(18):2707–2713. [PubMed: 18977655]
4. Gunter, L.; Zhu, J.; Murphy, S. Proceedings of the 11th conference on Artificial Intelligence in Medicine. Springer Verlag; 2007. chap. Variable selection for optimal decision making
5. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*. 2005; 1(11):7872–7878. [PubMed: 16278411]
6. Mandrekar S, Grothey A, Goetz M, Sargent D. Clinical trial designs for prospective validation of biomarkers. *American Journal of Pharmacogenomics*. 2005; 5(5):317–325. [PubMed: 16196501]
7. Sargent D, Conley B, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*. 2005; 23(9):2020–2027. [PubMed: 15774793]
8. Simon R, Wang S. The use of genomic signatures in therapeutics development in oncology and other diseases. *Pharmacogenomics Journal*. 2006; 6(3):166–173. [PubMed: 16415922]
9. Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute*. 2007; 99:1036–1043. [PubMed: 17596577]
10. Simon R. Development and validation of biomarker classifiers for treatment selection. *Journal of Statistical Planning and Inference*. 2008; 138(2):308–320. [PubMed: 19190712]
11. Lijmer J, Bossuyt P. Various randomized designs can be used to evaluate medical tests. *Journal of Clinical Epidemiology*. 2009; 62(4):364–373. [PubMed: 18945590]
12. Mandrekar S, Sargent D. Clinical trial designs for predictive biomarker validation: One size does not fit all. *Journal of Biopharmaceutical Statistics*. 2009; 19:530–542. [PubMed: 19384694]
13. Simon R, Paik S, Hayes D. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *Journal of the National Cancer Institute*. 2009; 101:1446–1452. [PubMed: 19815849]
14. Taube S, Clark G, Dancey J, McShane L, Sigman C, Gutman S. A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. *Journal of the National Cancer Institute*. 2009; 101:1453–1463. [PubMed: 19855077]
15. Mandrekar S, Sargent D. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of Clinical Oncology*. 2009; 24(24):4027–4034. [PubMed: 19597023]
16. Clark G, McShane L. Biostatistical considerations in development of biomarker-based tests to guide treatment decisions. *Statistics in Biopharmaceutical Research*. 2011; 3:549–560.
17. Mandrekar S, Sargent D. Predictive biomarker validation in practice: lessons from real trials. *Clinical Trials*. 2010; 7:567–573. [PubMed: 20392785]
18. McShane L. Statistical challenges in the development and evaluation of marker-based clinical tests. *BMC Medicine*. 2012; 10:52. [PubMed: 22642713]
19. Freidlin B, McShane L, Korn E. Randomized clinical trials with biomarkers: Design issues. *Journal of the National Cancer Institute*. 2010; 102:151–160.
20. Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: a systematic review and classification. *Clinical Cancer Research*. 2013; 19:4578–88. [PubMed: 23788580]
21. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine*. 2011; 154(4):253–259. [PubMed: 21320940]
22. Huang Y, Gilbert PB, Janes H. Assessing Treatment-Selection Markers using a Potential Outcomes Framework. *Biometrics*. 2012; 68(3):687–696. [PubMed: 22299708]
23. Song X, Pepe MS. Evaluating markers for selecting a patient's treatment. *Biometrics*. 2004; 60(4): 874–883. [PubMed: 15606407]
24. Brinkley J, Tsiatis AA, Anstrom KJ. A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics*. 2010; 66:512–522. [PubMed: 19508237]
25. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. *Biometrics*. 2012

26. Zhao Y, Zeng D, Rush AJ, Kosorock MR. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*. 2012; 107:1106–1118. [PubMed: 23630406]
27. Janes H, Brown MD, Pepe MS, Huang Y. An approach to evaluating and comparing biomarkers for Patient treatment selection. *International Journal of Biostatistics*. 2014; 10:99–121. [PubMed: 24695044]
28. Early Breast Cancer Trialists Collaborative Group. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomized trials. *Lancet*. 2005; 365:1687–1717. [PubMed: 15894097]
29. Dowsett M, Goldhirsch A, Hayes DF, Senn HJ, Wood W, Viale G. International web-based consultation on priorities for translational breast cancer research. *Breast Cancer Research*. 2007; 9:R81. [PubMed: 18034879]
30. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*. 2004; 351(27):2817–2826. [PubMed: 15591335]
31. Paik S, Tang G, Shak S, Chungyeul K, Baker J, Kim W, Cronin M, Baehner FL, Watson D, Bryant J, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen-receptor-positive breast cancer. *Journal of Clinical Oncology*. 2006; 24:3726–3734. [PubMed: 16720680]
32. Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh IT. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: A retrospective analysis of a randomized trial. *Lancet Oncology*. 2010; 11:55–65. [PubMed: 20005174]
33. Janes H, Pepe MS, Huang Y. A framework for evaluating markers used to select patient treatment. *Medical Decision Making*. 2013; 34:159–67. [PubMed: 23811760]
34. Albain KS, Barlow WE, Davdin PM, Farrar WB, Burton GV, Ketchel SJ, et al. Adjuvant chemotherapy and timing of tamoxifen in postmenopausal patients with endocrine-responsive, node-positive breast cancer: A phase 3, open-label, randomized controlled trial. *Lancet*. 2009; 274:2055–2063. [PubMed: 20004966]
35. Vickers AJ, Kattan MW, Sargent D. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*. 2007; 8:14. [PubMed: 17550609]
36. Pusztai L, Hess KR. Clinical trial design for microarray predictive marker discovery and assessment. *Annals of Oncology*. 2004; 15:1731–1737. [PubMed: 15550577]
37. Simon R. The use of genomics in clinical trial design. *Clinical Cancer Research*. 2008; 14:5954–5958. [PubMed: 18829473]
38. Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ. Clinical application of the 70-gene profile: the MINDACT trial. *Journal of Clinical Oncology*. 2008; 26:729–35. [PubMed: 18258980]
39. Sparano JA. TAILORx: trial assigning individualized options for treatment (Rx). *Clinical Breast Cancer*. 2006; 7:347–50. [PubMed: 17092406]
40. Barlow, WE. *Handbook of Statistics in Clinical Oncology*. Chapman and Hall/CRC; 2012. chap. Design of a Clinical Trial for Testing the Ability of a Continuous Marker to Predict Therapy Benefit
41. Tajik P, Bossuyt PM. Genomic markers to tailor treatments: waiting or initiating? *Human Genetics*. 2011; 130:15–8. [PubMed: 21487770]

Appendix

This section describes the results of a sensitivity analysis for marker Y_2 in the breast cancer example. We assess the impact of varying the true risk model parameters over a range of values consistent with the SS8814 trial data, identified as follows. We bootstrap 500 times from a simulated dataset mimicking the SS8814 data. In each bootstrap sample, we fit the linear logistic risk model (1) and calculate true marker performance measures under that

fitted model. This gives us a distribution of risk model parameters and marker performance measures consistent with the original data. Then we choose a set of risk model parameters $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ that yield $\hat{\gamma}_3$ and Θ values spanning the 10th to 90th percentiles of the marginal distributions of these quantities. In all, 19 sets of risk model parameters are chosen. Next, we simulate data under the linear logistic model (1) with these parameters and calculate the probability of satisfying the design criteria of Section 3.1.1. The sample size is fixed at N such that each criterion is satisfied with 90% probability using model-based estimation and the original risk model parameters. Figure 2 shows the results for criteria (1)–(3). We see that if γ_3 is varied over its interquartile range, power for criterion (1) varies from about 75% to 100%. When Θ is at its 25th percentile, criteria (2) and (3) are satisfied with about 70% probability whereas the target is 90%. The probability of satisfying criterion (4) varies relatively little, from 83% to 92% across all 19 scenarios. These calculations could be used to support a modest increase in sample size for Y_2 to ensure that criteria (2) and (3) are satisfied even if the marker has lower performance than evident based on pilot data.

Figure 1.

Risk of D given $T = 0$ or $T = 1$ as a function of the marker, for the Oncotype DX-like marker (Y_1), weak simulated marker (Y_2), and strong simulated marker (Y_3) in the breast cancer example. The distribution of the marker-specific treatment effect, $(Y) = P(D = 1|T = 0, Y) - P(D = 1|T = 1, Y)$, is also shown for each marker.

Figure 2.

Probability of satisfying criteria (1) (top) and (2) and (3) (bottom) where 19 sets of risk model parameters, consistent with the SS8814 data, are used as the basis for the simulation-based power calculations. For each criterion, the sample size is fixed at N ensuring 90% probability of satisfying the criterion under the original risk model parameters: $N = 1050$ for criteria (1) and (2) and $N = 4675$ for criterion (3). For criterion (1), the probability is plotted against $|\gamma_3|$, the absolute value of the marker-by-treatment interaction coefficient, and is monotone in this parameter (up to Monte-Carlo error). For criteria (2) and (3), the probability is plotted versus θ , which explains most but not all the variation in the probability. Smoothing splines are overlaid.

Total sample size (N) needed to satisfy each of the design criteria of Section 3.1.1, or combinations of these criteria, with 80% or 90% probability, for markers Y_1 , Y_2 , and Y_3 in the breast cancer example. Results for criteria (2) and (3) depend on whether the model-based or empirical estimator of Θ is used. Sample sizes are rounded up to the nearest 25. Y_1 has the same performance as Oncotype DX; Y_2 and Y_3 are hypothetical markers with weak and strong performance, respectively.

Table 1

Criterion	Y_1			Y_2			Y_3		
	80% probability			80% probability			80% probability		
	$\hat{\Theta}^m$	$\hat{\Theta}^f$	$\hat{\Theta}^m$	$\hat{\Theta}^m$	$\hat{\Theta}^f$	$\hat{\Theta}^m$	$\hat{\Theta}^m$	$\hat{\Theta}^f$	$\hat{\Theta}^m$
1	1475	1475	1975	800	800	1050	100	100	150
2	4275	> 10000	5750	450	3600	1050	75	75	150
3	> 10000	> 10000	> 10000	2825	3800	4675	275	325	500
4	2400	2400	5150	2225	2225	5225	2275	2275	5400
1+2	4300	> 10000	5775	1075	3600	1375	150	300	375
1+3	> 10000	> 10000	> 10000	2825	3800	4675	250	450	650
1+4	2800	2800	5175	2250	2250	5225	2275	2275	5400

Total sample size for the RCT cohort (N) and for various case-control designs needed to satisfy each of the design criteria of Section 3.1.1 with 90% probability, for markers Y_2 and Y_3 in the breast cancer example. The RCT cohort design is compared to case-control designs that sample n out of N subjects consisting of: all cases and 1 control per case (1:1 case-control), all cases and 2 controls per case (1:2 case-control), all cases and 1 control per case matched on treatment assignment (1:1 matched case-control), and all cases and 2 controls per case matched on treatment assignment (1:2 matched case-control). Results are shown for the model-based estimate of Θ only. Sample sizes for the RCT cohort are rounded up to the nearest 25.

Stat Med. Author manuscript; available in PMC 2016 November 30.

Probability (%) of satisfying each design criterion of Section 3.1.1 as a function of λ , the treatment allocation probability, for marker Y_2 in the breast cancer example. Under one approach (“Satisfy each criterion”), the total sample size of the trial is set to the N which ensures that each design criterion is satisfied with at least 80% or 90% probability under $\lambda = 0.5$ using model-based estimation. Under another approach (“Satisfy all criteria”), the total sample size is set to that which ensures that all criteria are satisfied with 80% (90%) probability. Sample sizes are rounded up to the nearest 25.

Table 3

		λ							
Prob.	Criterion	N	0.2	0.3	0.4	0.5	0.6	0.7	0.8
80%	Satisfy each criterion	1	800	62	72	79	80	80	74
		2	450	74	77	78	81	81	82
		3	2825	69	75	79	80	81	77
		4	2225	76	78	80	81	79	77
90%		1	1050	73	84	88	90	89	84
		2	1050	81	85	89	91	90	89
		3	4675	78	86	89	90	90	87
		4	5225	86	88	91	91	89	88
80%	Satisfy all criteria	1	2825	99	100	100	100	100	100
		2	2825	96	98	100	100	100	99
		3	2825	69	75	78	79	80	76
		4	2825	80	81	82	84	83	78
90%		1	5225	100	100	100	100	100	100
		2	5225	100	100	100	100	100	100
		3	5225	82	87	92	91	90	90
		4	5225	86	88	91	91	89	88