Evaluating Putative Predictive Biomarkers in Randomized Clinical Trials

Guidance Document
August 2012

Authors:

Parvin Tajik\textsuperscript{1,2}, Aleiko H. Zwinderman\textsuperscript{1}, Ben W. Mol\textsuperscript{2}, Patrick M. Bossuyt\textsuperscript{1}

\textsuperscript{1}Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam

\textsuperscript{2}Department of Obstetrics and Gynaecology, Academic Medical Center, University of Amsterdam
Table of Contents

1 General introduction and outline .................................................................................................................. 3
  1.1 Definition and classification of biomarkers ......................................................................................... 4
  1.2 Outline of the document ....................................................................................................................... 5

2 Study questions in biomarker studies ....................................................................................................... 5

3 Study designs for biomarker evaluation ................................................................................................... 6
  3.1 Single-arm ........................................................................................................................................ 7
  3.2 Enrichment ...................................................................................................................................... 7
  3.3 Randomize all .................................................................................................................................. 8
  3.4 Biomarker-strategy ............................................................................................................................ 9
    3.4.1 Biomarker-strategy, with biomarker assessment in the control arm .................................... 9
    3.4.2 Biomarker-strategy, without biomarker measurement in the control arm ........................ 10
    3.4.3 Biomarker-strategy, with treatment randomization in the control arm ............................ 10
  3.5 Combined designs ............................................................................................................................... 11

4 Issues and techniques in analysis of biomarker trials ............................................................................. 12
  4.1 Statistical issues for biomarker trials ............................................................................................... 12
    4.1.1 Multiplicity ........................................................................................................................... 12
  4.2 Analysis Plans .................................................................................................................................. 13
    4.2.1 Single-arm design ................................................................................................................ 13
    4.2.2 Enrichment design ............................................................................................................... 13
    4.2.3 Randomize all design ........................................................................................................... 13
    4.2.4 Biomarker-strategy .............................................................................................................. 16

5 Sample size and efficiency of the designs ................................................................................................ 17
  5.1 Sample size calculations ...................................................................................................................... 17
    5.1.1 Single-arm design ................................................................................................................ 17
    5.1.2 Enrichment design ............................................................................................................... 17
    5.1.3 Randomize all design ........................................................................................................... 17
    5.1.4 Biomarker-based strategy design ........................................................................................ 18
  5.2 Comparative efficiency ....................................................................................................................... 19

6 A proposed strategy for design selection ................................................................................................. 21
7 Case studies from ZonMW trials (HYPITAT and EORTC)......................................................................................................................... 24

7.1 HYPITAT - Cervical favorability and the decision for labor induction in gestational hypertension or mild preeclampsia at term ........................................................................................................................................... 24

7.2 EORTC 55971 - Primary surgery or neoadjuvant chemotherapy in stage IIIC or IV ovarian cancer 26

8 Concluding remarks .................................................................................................................................................................................. 28

9 Tables .............................................................................................................................................................................................................. 29

10 References ........................................................................................................................................................................................................ 32
1 General introduction and outline

One of the challenges faced by health care professionals is dealing with the variability of the clinical signs and symptoms that are associated with diseases. The fact that two patients with the same disease are never identical and in many cases respond differently to managements, makes treatment decisions always complicated and the outcome uncertain. This issue has gained importance with the increasing availability of drugs and interventions of high costs and considerable side-effects. With current global efforts to reduce costs in health care systems, decision-makers would like to restrict the use of interventions to those for whom the health gains are sufficiently large to justify the burden of treatment to patients and to society. Developments in biotechnology, genomics and imaging technologies have provided biomarkers for indicating the heterogeneity and grasping the biological basis of disease.

Advances in our understanding of the heterogeneous nature of diseases leads to new challenges for clinical trial design. The standard paradigm for the design of Phase III clinical trials involves defining broad eligibility criteria and basing conclusions on comparing aggregated outcomes; demonstrations of treatment effectiveness are based on tests of the overall null hypothesis that the treatment effect is zero. This paradigm leads to designing large clinical trials to identify small treatment effects in heterogeneous groups of patients. This approach can be inefficient and may result in practice standards in which many patients are treated with toxic and expensive drugs from which they little or no benefit.

Methodologists are developing new and alternative approaches to clinical trial design and analysis, entering a new era of predictive medicine, in which appropriate treatments can be matched to qualified patients in a reliable manner. The original emphasis on broad eligibility criteria for RCT has been based on concern that drugs found effective in clinical trials might subsequently be used in a broad patient population. The focus on the overall null hypothesis was based on concerns about the multiple testing involved in the commonly practiced exploratory post hoc subset analysis and the assumption that qualitative interactions are unlikely. The common advice was to perform subset analyses without believing in them. The famous subset analysis of a trial based on patient astrological sign is still prominent in the minds of many researchers and clinicians. This paradigm was based on implicit assumptions that qualitative interactions are unlikely and that drugs are inexpensive and without serious side effects. Today these assumptions are not always appropriate for many drugs. Treating the majority for the benefit of a minority may no longer be an effective public health strategy.

Today, we are challenged to develop a new paradigm of clinical trial design and analysis that enables the development of a predictive medicine that is science based and reliable. Currently we have powerful tools for characterizing the diseases biologically. We now can use this characterization to provide a stronger basis for the design and analysis of clinical trials. The primary analysis of the new generation of clinical trials must likely consist of more than just testing the null hypothesis of no average effect. It is also clear that the tradition of post hoc data dredging subgroup analysis is not an adequate basis for predictive medicine. We need prospective analysis plans that provide for both preservation of Type I experiment-wise error rate and for focused predictive analyses that can be used to reliably select patients in clinical practice for use of the new treatments. These two primary objectives are not inconsistent and clinical trials can be sized for both purposes. By moving from ‘subset analysis’ to
‘classifier development’ the problem is moved, from one of multiple testing to one for evaluation of classification and prediction

1.1 Definition and classification of biomarkers

The National Institutes of Health Biomarker Definitions Working Group has defined a biological marker or biomarker as ‘a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.’ According to this definition, biomarkers could cover a rather wide range of data types, from, for example, biochemistry laboratory tests to genetic tests, imaging, and clinical characteristics or a combination of them.

Biomarkers that are informative for clinical outcome are broadly classified as prognostic or predictive biomarkers. The Biomarker Definition Working Group defines predictive biomarkers as pretreatment markers that predict who will benefit from a particular therapy, separating them from those who will not. Hence, there is a differential treatment benefit, according to the value of a predictive biomarker. Treatment benefit has to be defined in terms of the comparison between experimental and control treatment. In statistical terms, this difference in the benefit from treatment effect can be investigated by testing for a biomarker by treatment interaction. As an example, HER2 is a well-known predictive biomarker. It distinguishes breast cancer patients who would benefit from trastuzumab more than other second-line treatment options.

Prognostic biomarkers classify patients treated with standard therapies - including no treatment, if that is standard practice - into subgroups with distinct expected clinical outcomes. A prognostic marker measured at diagnosis may predict a shorter survival compared to that observed in patients without the marker, for example, both in the absence of as well as in the presence of treatment.

Janes and colleagues have recently coined the term ‘treatment selection markers’. Considering the logic of treatment decisions, one can support the decision to start treatment if the outcome under treatment is better than the outcome under the alternative (no treatment) and if the benefit exceeds a prespecified treatment threshold. The threshold for treatment is often defined by treatment harms, burden and costs. Prognostic and predictive biomarkers whose information has implications for treatment decisions can both be considered here as treatment selection markers.

As an example, the MammaPrint® gene signature (Netherlands Cancer Institute and Agendia BV, Amsterdam, The Netherlands) is a validated prognostic biomarker which predicts the risk of metastasis within 5-year of resection. This prognostic information may have implications for treatment selection; if the signature can correctly distinguish women who have a low risk of metastasis, these low risk women may be spared from cytotoxic treatments, because their baseline risk is low and the potential risk reduction due to treatment is less than the harms and costs of the chemotherapy. In this case a prognostic tool could serve as a treatment selection marker.

Another example is KRAS gene expression in colorectal cancer tumors. Patients without KRAS mutations were shown to have significantly better outcomes with anti–epidermal growth factor receptor (EGFR) treatment than those with KRAS mutations, who derive essentially no benefit from it. This makes...
KRAS expression, as a predictive biomarker, useful for selecting treatment. The U.S. Food and Drug Administration labeling for 2 EGFR inhibitors, cetuximab and panitumumab, has been changed, stating that these drugs are not recommended for the treatment of colorectal cancer in patients with KRAS mutations in codon 12 or 13.

1.2 Outline of the document
The focus of this report is on methods for evaluating the potential of biomarkers as treatment selection biomarkers. We first review the typical study questions in biomarker studies in Section 2. In Section 3, we present a brief review and classification of the biomarker trial designs. In section 4 the issue of multiplicity is discussed and different analysis plans of each trial design are presented. In section 5, methods of sample size and power calculation of different designs are summarized and the efficiency of designs are compared. Section 6 includes a proposed strategy for selecting a design or analysis plans for biomarker trials. In section 7 we present two examples of biomarker analyses we have performed on two clinical trials. In section 8 we discuss these approaches and present our discussion and conclusion.

2 Study questions in biomarker studies
When evaluating treatment selection biomarkers three types of effects can be studied. To ease the presentation, we consider a scenario of evaluating an experimental treatment (E) versus a control treatment (C) in the presence of a binary biomarker, of which the values can be classified as positive or negative. However, the discussions are generalizable to conditions with more than two treatment options and where the biomarker(s) under study is (are) not binary.

When studying our binary biomarker and the experimental treatment, one class of effects to be studied is that of treatment effects, which include the effect of treatment in biomarker-positives, the treatment effect in biomarker-negatives, and the treatment effect in the overall patient population.

In many situations, such as in targeted therapies, there is strong prior evidence that the experimental treatment is only effective in biomarker-positive patients, therefore a biomarker-based treatment strategy can be proposed, in which biomarker-positive patients receive treatment and biomarker-negative patients are managed with the control treatment. We then are interested in the effect of the biomarker-based strategy versus a non-biomarker-based strategy, such as managing all patients with the control treatment. In the same way we can compare the biomarker-based strategy with treating all patients with the experimental treatment.

Biomarker effects are the other group of effects which are usually investigated in biomarker studies. The strength of the association between the biomarker and the outcome in the control arm shows if the biomarker is prognostic. One can also investigate if biomarker status is associated with the outcome in the experimental arm.

Finally, one may also compare the treatment benefit in biomarker-positives with that in biomarker-negatives, to see whether the biomarker status can identify who will benefit from treatment. Box 1 summarizes the type of questions that biomarker studies usually aim to answer. 11
Box 1. Study questions in the biomarker studies

1. **Treatment effects**
   a. How does the experimental treatment compare to the control treatment in biomarker-negatives?
   b. How does the experimental treatment compare to the control treatment in biomarker-positives?
   c. How does the experimental treatment compare to the control treatment in overall study population?
   d. How does the biomarker-based treatment strategy compare to the control treatment in the overall study population?
   e. How does the biomarker-based treatment strategy compare to the experimental treatment in the overall study population?

2. **Biomarker effect**
   a. Is the biomarker status associated with the outcome in the standard of care group? (Is the biomarker prognostic?)
   b. Is the biomarker status associated with the outcome in the experimental treatment group?

3. **Biomarker by treatment effect**
   a. Is the biomarker status associated with a benefit of experimental treatment? (Is the biomarker predictive?)

### 3 Study designs for biomarker evaluation

The main objectives of studies for the evaluation of treatment selection biomarkers are treatment effect evaluations. While observational studies of various biomarkers can provide useful preliminary insights for defining a promising subgroup or hypotheses for future definitive testing, a randomized clinical trial is needed to more fully exploit the potential of treatment selection biomarkers. Biomarker trials differ from conventional clinical trials of treatment effects because they should be designed in a way to be able to answer the biomarker-related questions (Box 1).

A rapid review of the biomarker trial designs reported in literature suggests a substantial variability in the designs, as well as in the terms or labels proposed by authors for defining the designs. We performed a systematic review of literature on trial designs for evaluating treatment selection markers. The review was based on a comprehensive and systematic search in multiple databases (MEDLINE, EMBASE, Cochrane Methodology Register and MathSciNet) and hand searching of references and citing papers of all included papers using the Web of Science database. Eligible for inclusion were methodological articles that described one or more trial designs for identification and/or validation of prognostic or predictive biomarkers for tailoring treatment.

The initial search yielded 1,758 abstracts, of which 98 were deemed eligible based on title and abstract. After assessing the full texts of these articles, 55 were included. Seventeen other papers could be added by hand searching of references and citing articles, resulting in a total of 72 articles in the final analysis. From all included articles and for each proposed study design we extracted detailed data on the proposed design label, design objectives, patient flow elements, and analysis plan. Our definition of patient flow is composed of the biomarker status of patients deemed eligible for the study (biomarker-positives only, both biomarker-positives and -negatives, or biomarker status unknown), the intervention participants are assigned to (whether or not biomarker status is used for assigning study participants to...
the experimental treatment) and the comparison assignment (control treatment or both control and experimental treatments). The combination of patient, intervention and comparator of the design determines what type of comparisons can be made in evaluating the value of the putative predictive biomarker in that design.

We started our analysis by developing the list of study labels from all included studies. For each label in this list we then explored the participant-intervention-comparator components of patient flow. We then clustered all study labels with identical components into disjoint categories. The most frequently used label for describing designs of that patient flow category was selected for labeling the corresponding class of study designs. Table 1 summarizes the design labels and their corresponding patient flow elements that we extracted in the review. The heterogeneity in the labeling of designs is noticeable. Labels clustered under the same bullet describe completely identical trial design. Other labels in the same patient flow category, have distinguished objectives and analysis plans.

By comparing the patterns of all the included designs we could distinguish four basic and distinct patient flow categories, as well as a fifth category, consisting of combinations (serial or parallel) of two or more of the five basic patient flow elements. (Table 1) In the section below, we discuss the patient flow categories extracted from the study designs in more detail.

3.1 Single-arm
In single-arm trials all patients, irrespective of their biomarker status, are included in the trial and all undergo the experimental treatment (Figure 1; Exp.: experimental). With this design one can evaluate whether the biomarker status is associated with the outcome of therapy. However, it would not be clear if the biomarker is related to the overall patient’s prognosis or the patient’s benefit from the specific therapy. Therefore, this design does not help in evaluating the potential of biomarker as a treatment selection marker; an association with a difference in outcome after treatment does not provide evidence of an association with a difference in benefit from treatment, since benefit has to be defined in terms of the comparison between experimental and control treatment. Hence, an evaluation of treatment selection biomarkers requires statistical data from RCTs that include both experimental and control treatment. (Table 2)

![Figure 1. Single-arm design](image)

3.2 Enrichment
In an enrichment trial, all potentially eligible patients are first tested for the biomarker under study and only patients with a certain biomarker value are randomly assigned to experimental or control treatment. Patients who have biomarker values other than the desired value are excluded from further investigations (Figure 2; Ctrl: control treatment). Enrichment trials allow us to compare the effect of experimental treatment versus control treatment in biomarker-positive patients; it does not provide evidence whether the biomarker is prognostic or predictive. (Table 2) Identification of a treatment
selection biomarker requires statistical data from RCTs that include both biomarker-positive and biomarker-negative patients.

![Figure 2. Enrichment design](image)

The pivotal trial for trastuzumab and human epidermal growth factor receptor2 (HER2)\textsuperscript{17} is a well-known example of an enrichment design. Patients with HER2-positive breast adenocarcinoma were enrolled and randomized to chemotherapy with or without trastuzumab. This study provided strong evidence that trastuzumab combined with chemotherapy improves outcomes among women with HER2-positive breast cancer.

### 3.3 Randomize all

In this trial category the biomarker-status is not part of the eligibility criteria. All patients, irrespective of their biomarker status, enter the study and are randomly allocated to experimental or control treatment. Afterwards associations between biomarker status and treatment outcome are evaluated (Figure 3). There exist several variations in the design and analysis of trials of this category.

![Figure 3. Randomize all design](image)

In some trials biomarker measurement is obligatory before randomization; however, in some others the measurement can happen retrospectively, using data and stored biospecimens collected in previously completed RCT’s. A typical example of retrospective introduction of the biomarker question to the trial was the evaluation of KRAS mutations in colorectal cancer. There a prospectively specified analysis of data from a previously conducted randomized Phase III trial of panitumumab versus best supportive care showed a significant treatment by KRAS interaction (p<0.0001).\textsuperscript{18} This analysis, supported by similar findings in other independent trials, demonstrated that the benefit from panitumumab is restricted to patients with wild type KRAS status, with no clinical benefit for patients with KRAS mutations in codon 12 or 13.
The type of randomization is another source of variability of trials in this category. In some trials, researchers apply a simple 1:1 randomization procedure for all patients. In some other trials, where the biomarker under evaluation is binary or categorical with few categories, randomization is stratified according to the biomarker status. In other words, biomarker-positive and biomarker-negative patients are separately randomized to experimental or control treatment. This randomization scheme causes the biomarker status to be more balanced in both arms, and more importantly it makes sure that the biomarker status is available for analysis for all trial participants. In cases where the outcome of interest can be assessed in a relative short period of time and the trial has a good infrastructure for timely and frequent follow up and interim analysis, one may apply an ‘outcome-based adaptive randomization’ scheme, in which, the randomization probabilities of the next patients are adapted by the response rates of the previous patients in the trial. Then patients with a specified biomarker value will be randomized with a higher probability to the treatment to which the former patients of that biomarker category have responded most.

In trials within the randomize all category, the eligibility criteria are independent of the biomarker status and all biomarker subgroups are allocated randomly to experimental or control treatment. Therefore, if the sample size is large enough, the trial potentially provides direct evidence for all of the biomarker-related hypotheses except two. The questions whether treatment selection based on biomarker status leads to better outcomes than assigning all patients to control treatment, or compared to assigning all patients to experimental treatment, can only be answered indirectly. Depending on the prior evidence, randomization scheme and the objectives of the biomarker analysis, various statistical plans are proposed for a trial with randomize all patient flow. In many cases authors have coined a design label for referring to a randomize all design when analyzed based on a specific analysis plan. Examples of such labels are as ‘stratified analysis’, ‘marker by treatment interaction’, ‘biomarker analysis’, ‘sequential testing’, ‘prospective subset’, ‘adaptive threshold’, ‘adaptive biomarker’, ‘adaptive signature’, ‘cross-validated adaptive signature’ and ‘generalized adaptive signature’. All these designs have a randomize all patient flow, but they differ in their analysis plans. In the next section (see Section 5.1.4) we will discuss these analysis plans in detail.

### 3.4 Biomarker-strategy

The distinguishing feature of designs in the biomarker-strategy patient flow category is the inclusion of a new management strategy. This is not experimental or control treatment, but a prespecified maker-based treatment strategy. For example, biomarker-positive patients receive experimental therapy while all biomarker-negative patients get the control treatment. Eligible patients are randomized to this biomarker-based treatment strategy, or to control treatment. Designs with biomarker-strategy patient flow allow a direct comparison of the outcomes with a biomarker-based treatment strategy versus those of control treatment. This comparison addresses the clinical utility of a predictive biomarker. We can distinguish three subtypes in this design category.

#### 3.4.1 Biomarker-strategy, with biomarker assessment in the control arm

Biomarker status is assessed in all patients enrolled in the trial, who are then randomized to the biomarker-strategy arm or to control treatment (Figure 4). By analyzing the association between
biomarker status and outcome in the control arm one can also evaluate whether the biomarker is prognostic.\textsuperscript{11,24} Yet, this design cannot be used to evaluate if the biomarker is predictive. Moreover, any findings regarding the efficacy of the biomarker-directed approach to therapy cannot distinguish between a true effect of biomarker-directed strategy, and an improved experimental regimen, regardless of biomarker status (Table 2).

**Figure 4.** Biomarker-strategy design with biomarker assessment in the control arm

### 3.4.2 Biomarker-strategy, without biomarker measurement in the control arm

In settings where it is not feasible or ethical to test the biomarker in all patients, biomarker status is only acquired in patients allocated to the biomarker-strategy arm (Figure 5). As a result, the biomarker status of patients in the control strategy is unknown. This limitation confines direct assessment of prognostic value of the biomarker.

**Figure 5.** Biomarker-strategy, without biomarker assessment in the control arm

### 3.4.3 Biomarker-strategy, with treatment randomization in the control arm

Sargent and Allegra have proposed a modification to the first subtype, wherein a second randomization between experimental versus control therapy is added to the comparator arm\textsuperscript{22} (Figure 6). This modification allows evaluating what is the best treatment in biomarker-positive and biomarker-negative patients. This variation also allows indirect evaluation whether the biomarker is predictive.\textsuperscript{11,16}
3.5 Combined designs

A final category of study designs consists of trials in which two or more of the aforementioned basic designs are combined to form a new design. Combinations of designs are usually required when trial aims at evaluating multiple hypotheses, multiple biomarkers, multiple treatments, or when the trial has several stages.

The simplest combination is when in enrichment trial biomarker-negative patients are not excluded from the study but are assigned to control treatment and undergo outcome assessment. Here, an enrichment flow is combined in parallel with a single arm trial of control treatment in biomarker-negative patients. This variation with collecting specimens and follow-up information from all patients allows for future testing for other potential prognostic biomarkers. An example of a hybrid design is the Trial Assigning Individualized Options for Treatment (TAILORx). This study aims at the evaluation of Oncotype Dx, a 21-gene recurrence score, in tamoxifen-treated patients with breast cancer. In this ongoing trial, patients are divided into 3 subgroups of low, intermediate and high risk based on their Oncotype Dx® recurrence score. Low risk patients receive hormonal therapy, high risk patients receive both hormonal and chemotherapy, while patients with an intermediate risk are randomized to hormonal therapy or chemotherapy plus hormonal therapy.

Staged trial designs have also, in essence, a combination of basic patient flows. For instance, the proposed ‘two-stage sample-enrichment’ design by Liu and colleagues starts with accruing only biomarker-positive patients during the initial stage of the trial. At the end of the first stage, an interim analysis is performed comparing the outcome of experimental treatment versus control for the biomarker-positive patients. If the results are not promising for the experimental treatment, accrual stops and no treatment benefit is claimed. If the results are promising for the biomarker-positive patients at the end of the first stage, accrual continues with recruiting unselected population. This design can be either an enrichment trial only or a combination of enrichment and a traditional flow, conditional on the result of the interim analysis.
4 Issues and techniques in analysis of biomarker trials

4.1 Statistical issues for biomarker trials

4.1.1 Multiplicity

Multiplicity is an important issue for multiple testing in the planning, data analysis, and interpretation of biomarker studies. Simultaneous considerations of a set of statistical inferences are common in biomarker studies. Biomarker studies frequently incorporate one or more of the following design features: subgroup analysis to address particular concerns on efficacy and safety of the drug in specific patient subgroups (e.g., biomarker-positive/negative), repeated tests of significance as the study progresses (interim analyses) to ensure early detection of effective treatments, exploration of multiple prognostic biomarkers, multiple outcome measures, comparison between more than two treatment groups, and various combinations of these features.

The concept of a Type I error rate originates in the problem of testing a single hypothesis. When more than one test statistic is used, the chance of rejecting the null hypothesis and incorrectly declaring a benefit increases. When ten statistically independent statistical tests are performed, each with a significance level of 0.05, and the null hypothesis is actually true (no true difference), the chance of at least one test being significant is no longer 0.05, but approximately $0.40 = 1 − (1 − 0.05)^{10}$.

The hypotheses of interest are considered together, as a family, and when doing multiple comparisons the family wise error rate should be controlled to stay at 0.05 level. Two important types of multiple testing procedures are single-step and sequential procedures. Single-step procedures are multiple testing procedures for which the decision to reject any hypothesis does not depend on the decision to reject any other hypothesis. In other words, the order in which the hypotheses are tested is not important. The Bonferroni procedure is an example of a single-step procedure.

Unlike single-step procedures, sequential procedures are carried out in a progressive manner. Some hypotheses are not tested explicitly and may be retained or rejected by implication. Main forms of sequential procedures include fixed-sequence, closed-testing and stepwise procedures. These techniques provide an attractive alternative to single-step procedures because they can reject more hypotheses without inflating the overall error rate. Some of these techniques are used in the design and analysis of biomarker trials and are presented below. Interested reader for methods of multiplicity adjustment are referred to the papers published by Goeman, Bauer, Westfall and Young and Dmitrienko and colleagues.

For pharmacogenomic studies and genome-wide association studies that focus on finding sets of predictive genes, an alternative approach to multiple testing considers the false discovery rate (FDR), which is the probability that a given gene identified as differentially expressed is a false positive. The FDR is typically computed after a list of differentially expressed genes has been generated. Unlike a significance level, which is determined before looking at the data, FDR is a post-data measure of confidence. It uses information available in the data to estimate the proportion of false-positive results that have occurred.
4.2 Analysis Plans

4.2.1 Single-arm design
Analysis of a single arm trial includes evaluating the association between the baseline biomarker values of patients and treatment response or treatment outcome. Depending on the type of biomarker (binary, nominal or continuous) and type of the outcome (binary, continuous or time to event) a specific type of analysis would be suitable. For example, when both biomarker and outcome are binary a simple two-by-two table and chi-square test is suitable, while for a continuous biomarker and time to event outcome a Cox proportional hazard model is appropriate.

Specific attention should be paid in this design to the issue of multiple testing due to exploration of multiple prognostic biomarkers. Although hypothesis testing and strict type I error control are traditionally part of the confirmatory data analysis, there is now more need for multiple comparison procedures in exploratory setting. The main reason for this is that researchers want to protect themselves from following up on too many false leads and doing too many unsuccessful validation experiments. Recently, Goeman and Solari\textsuperscript{36} have proposed a new procedure for this purpose and developed an R package (cherry) for implementing this technique.

4.2.2 Enrichment design
In enrichment design only marker-positive patients are randomized to experimental (E) and control (C) treatment and the comparison is within that subset. Therefore, the analysis simply consists of comparing the outcome of experimental treatment versus control treatment in biomarker-positive patients. The null hypothesis to be tested is $h_0: \mu_{+,e} = \mu_{+,c}$, where $\mu_{+,e}$ is the outcome in biomarker-positives by experimental treatment and $\mu_{+,c}$ is the outcome in biomarker-positives by control treatment.

4.2.3 Randomize all design
Depending on the type of the biomarker, the prior evidence and objectives of the analysis, a trial with the randomize all design can have different analysis plans. Below we summarize a list of more common analysis plans.

4.2.3.1 Treatment effects
This trial design aims at assessing the treatment effect in the overall population and in subpopulations defined by biomarker values. There are three possible analysis plans for this objective. One may choose between any of these analysis plans, based on prior knowledge of the treatment effects in the overall populations and in marker-defined subpopulations and on the sample size required for each analysis plan.

i) \textit{Separate testing}: Comparing experimental and control treatments separately in biomarker-positive and biomarker-negative subsets each.\textsuperscript{22}

ii) \textit{Fall-back testing}: First comparing experimental and control treatments for all randomized patients at a reduced two-sided alpha level (e.g. 0.03). If the overall test is not significant, then experimental treatment is compared with control in biomarker-positive subset using the remaining significance threshold (e.g. 0.02).\textsuperscript{39} One may use this two-step plan for situations where there is no strong prior confidence in the biomarker.
iii) **Closed testing:** Comparing E to C in the biomarker-positive subset at two-sided 0.05 level. If that is significant, then the biomarker-negative subset at two-sided 0.05 level will be evaluated. This plan is suitable in cases where one does not *a priori* expect the treatment to be effective in biomarker-negative patients unless it is effective in the biomarker-positive patients.  

iv) **Test of interaction:** One approach is first testing for an interaction between the magnitude of the treatment effect and biomarker status. If this interaction is not significant, then proceeding by just comparing treatments overall at two-sided significance level 0.05. Otherwise, comparing treatments within subsets at two-sided 0.05 level. 

In practice interaction tests are often performed by using statistical models. For example, consider a trial to evaluate a predictive biomarker for progression-free survival. A Cox proportional hazard model is often applied to the survival time and formulation of the model is as follows:

\[
\text{Hazard (Treat, Bio)} = h_0 \exp(h_{1}\text{Treat} + h_{2}\text{Bio} + h_{3}\text{Treat} \times \text{Bio})
\]

Treat = 0,1 indicates whether the patients were allocated experimental or control treatments, respectively; Bio = 0, 1 denotes negative or positive biomarker status, respectively. Treat × Bio indicates the interaction term. Here \(h_0\) is a baseline hazard, and \(h_1\) and \(h_2\) are the effects of treatment and biomarker status, respectively. An interaction effect by treatment and biomarker status is \(h_3\).

To identify whether the biomarker is predictive, the interaction effect \(h_3\) is tested for statistical significance. In addition, the hazard ratio (HR) is a useful statistic for quantitative interpretation of the treatment effect and interaction effect. In particular, the HR of the test treatment to the standard treatment in biomarker-negative patients is estimated as \(\exp(h_1)\), and the HR for biomarker-positive patients is estimated as \(\exp(h_1 + h_3)\), as derived from Equation above. Additionally, a logistic model for a binary outcome and analysis of (co)variance (ANOVA or ANCOVA) model for a continuous outcome are often applied to evaluate the interaction effect by treatment and biomarker status.

4.2.3.2 **Adaptive threshold**

Jiang and colleagues has proposed an “adaptive threshold” analysis plan in which the treatment effect is first evaluated in the overall population and, if not positive, a cutpoint for a single promising continuous biomarker is identified. This applies if the cutpoint is not well defined a priori. In this fallback analysis, first E to C is compared for all randomized patients at a reduced two-sided alpha level (e.g. 0.04). If it is significant, the trial is considered successful in showing effectiveness in the overall population and the procedure is done. However, if the overall test is not significant, the second stage of the analysis starts aiming at finding an optimal threshold for the biomarker, to define a subgroup of patients who would benefit from the experimental treatment. This analysis uses the remaining part of overall alpha (e.g. 0.01). Different methods can be applied for choosing the cutpoint.
(1) Jiang and colleagues originally propose choosing the cutpoint which yields the strongest statistical evidence for a treatment benefit based on a log-likelihood measure of treatment effect.

(2) The tail-oriented subpopulation treatment effect pattern plot (STEPP)\textsuperscript{41,42} graphs the estimated benefit of experimental versus control treatment among patients with a marker level greater than a cutpoint (thereby specifying the tail of the distribution) as a function of different cutpoints.

(3) The sliding-window subpopulation treatment effect pattern plot (STEPP)\textsuperscript{41,42} graphs the estimated benefit of experimental versus control treatment among persons with a marker level within an interval (thereby defining a sliding window) as a function of marker level.

(4) The selection impact curve\textsuperscript{43} graphs the benefit of marker-based treatment selection as a function of marker cutpoints.

(5) The marker-by-treatment predictiveness curves\textsuperscript{7} graph the risks of outcome separately under experimental and control treatments for persons with a marker in the interval.

4.2.3.3 Adaptive biomarker – multiple candidates

Simon proposes a generalized version of “adaptive threshold” analysis plan for situations where a small list of candidate binary biomarkers is available in advance but none of them have been selected as the potential treatment selection marker. The analysis starts with comparison of experimental to control treatment for all randomized patients at a reduced two-sided alpha level (e.g. 0.04). If it is significant, the trial is considered successful in showing effectiveness in the overall population and the procedure is done. However, if the overall test is not significant, the second stage of the analysis starts aiming at selecting one of the binary biomarkers to define a subgroup of patients who would benefit from the experimental treatment. The biomarker which has the maximum log-likelihood measure of treatment effect for patients who are positive for that biomarker is then selected. The statistical significance of the treatment effect in this biomarker-positive subset is determined by permuting the treatment group labels of the patients and the re-evaluating the treatment effect within the positive subset of the biomarker. Using bootstrap resampling, one can evaluate the proportion of the times that each patient is included in the positive subset of the selected biomarker and obtain a confidence interval for the treatment effect in the selected subset.\textsuperscript{26}

4.2.3.4 Adaptive signature

Freidlin and Simon have proposed “adaptive signature” analysis plan in which the treatment effect in first evaluated in the overall population and, if not positive, a classifier is developed using multiple biomarkers.\textsuperscript{27} This is also a fallback analysis; first E to C is compared for all randomized patients at a reduced two-sided alpha level (e.g. 0.04). If the treatment effect is significant, the trial is considered successful in showing effectiveness in the overall population and the analysis is complete. However, if the overall test is not significant, the second stage of the analysis starts aiming at identifying and evaluating a potentially promising subgroup using the following algorithm. First, using the training sample, a model is constructed to predict the benefit of experimental versus control treatment as function of baseline variables. This model
produces what is called a benefit function. The benefit function is then computed for each participant in the test sample to obtain a benefit score. A pre-specified cutpoint for the benefit score is used to specify a subgroup. The estimated benefit of treatment in this subgroup is tested at a statistical significance of 0.01 to determine if it is greater than the benefit threshold for T (T=0 in the original formulation). According to Freidlin and Simon the adaptive signature design is especially attractive for allowing pharmaceutical companies to invest in the development of pharmacogenomics signatures without the risk of losing indications where supported by the results of phase III trials.

Different benefit functions can be used:

1. Minimum odds ratio for treatment effect in at least \( k \) biomarkers with treatment-biomarker interaction that are statistically significant at some level \( c \), where \( k \) and \( c \) are specified in advance.
2. The risk difference: the probability of outcome in the experimental group conditional on baseline variables minus the probability of that same outcome in the control group conditional on the same variables.

### 4.2.4 Biomarker-strategy

The analysis plans of different subtypes of the biomarker-strategy design are different from the previous ones. We distinguish the following.

- **Biomarker-strategy, without biomarker measurement in the control arm**

  The differential effect of the two strategies can be compared by simply testing the null hypothesis of \( h_0: \mu_{bs} = \mu_c \) where \( \mu_{bs} \) is the outcome in the biomarker-strategy arm and \( \mu_c \) is the outcome in the control arm. Using this design, one can indirectly test two other null hypotheses:

  - \( h_0: \mu_{+,e} = \mu_{+,c} \) (treatment effect in the biomarker-positive group)
  - \( h_0: \mu_{+,c} = \mu_{-,c} \) (prognostic effect of the biomarker)

- **Biomarker-strategy, with biomarker assessment in the control arm**

  Again, the differential effect of the two strategies is compared by simply testing the null hypothesis of \( h_0: \mu_{bs} = \mu_c \). Measurement of the biomarker values in the control arm, allows us to test the following hypotheses directly:

  - \( h_0: \mu_{+,e} = \mu_{+,c} \) (treatment effect in the biomarker-positive group)
  - \( h_0: \mu_{+,c} = \mu_{-,c} \) (prognostic effect of the biomarker)
  - \( h_0: \mu_{+,e} = \mu_{-,e} \)

- **Biomarker-strategy, with treatment randomization in the control arm**

  In this design, the aggregated outcome in the biomarker-based arm (\( \mu_{bs} \)) is compared with the aggregated outcome in the randomized arm (\( \mu_r \)); \( h_0: \mu_{bs} = \mu_r \). Measurements of the biomarker values in all participants and randomization of patients in the non-biomarker-based arm between
experimental and control arm allow direct evaluation of all the biomarker-treatment hypotheses (Table 2).

5 Sample size and efficiency of the designs

5.1 Sample size calculations

5.1.1 Single-arm design
An issue which needs to be considered for the analysis and sample size estimation of these trials is the inherent multiple testing problem. The sample size can be calculated using the conventional formulae but the type I error rate for comparisons needs to be adjusted for the number of comparisons using one of the existing adjustment methods, like Bonferroni or exact. (Zaslavsky & Scott, 2012)

5.1.2 Enrichment design
The sample size calculation can be done by the standard formulae for traditional randomized trials. (Maitournam & Simon, 2005) For example for a binary outcome, the sample size of each trial arm will be estimated by:

\[ n = \frac{2(\bar{p})(1-\bar{p})(Z_\alpha + Z_\beta)^2}{(p_e - p_c)^2} \]

Where \( p_e \) is the outcome rate in the experimental arm, \( p_c \) is the outcome rate in the control arm. \( \alpha \) and \( \beta \) are the accepted Type I and II error rates, respectively. \( Z \) denotes percentile of the standard normal distribution. \( \bar{p} \) is the average of \( p_e \) and \( p_c \).

Zhao and Simon have made the methods of sample size planning for the design of enrichment trials available online at http://brb.nci.nih.gov. One consideration about the enrichment designs is the number of patients needed to be screened, which is very much dependent on the prevalence of biomarker positivity and the accuracy of biomarker assay. It is better to report the number needed to be screened along with the number needed to be randomized.

5.1.3 Randomize all design
This design can be analyzed with different analysis protocols. Depending on the protocol, the size of the sample required to provide enough statistical power for the comparison of interest differs.

Depending on the prior knowledge about the performance of the biomarker, researchers may choose one of approaches mentioned below:

- In case of applying sequential analysis techniques, sample size calculation needs also to be staged; first applying the conventional formulae using the reduced alpha and estimates of treatment effect in the first comparison, then estimating the detectable effect size in the 2\textsuperscript{nd} step analysis with an acceptable power (e.g. 80% or 70%) and possibly adjusting the overall sample size to have reasonable power for this analysis. (BRB website)
- In case of testing for interactions there are 3 options for sample size calculation:
Calculating the sample size needed for the overall test to have 90% power at 0.05 significance level, and running two simulations: (BRB website)

- With uniform treatment effect to calculate the probability of non-significant interaction and significant overall test (2-sided 0.05)
- With zero treatment effect in biomarker-negative subset and treatment effect delta in biomarker-positive subset to calculate probability of significant interaction and significant treatment effect for delta in biomarker-positive subset

Calculating the sample size needed to detect a prespecified odds ratio of interaction. Calculations can be done using the following formulae (Tajik et al., 2012):

\[
 n = \frac{(Z_\beta + Z_{\alpha/2})^2}{\sigma^2 \delta^2} \left( \frac{1}{p_e (1 - p_e)} + \frac{1}{p_c (1 - p_c)} \right)
\]

Here, \( p_e \) and \( p_c \) are response rates in overall E and C treated patients. \( \sigma^2 \) is the variance of the biomarker (which is: biomarker prevalence(1-prevalence)). \( \Delta \) is the logarithm of the interaction odds ratio (Ln OR) which we would like to detect. By this calculation a multiplicative interaction will be detected.

Another approach for sample size calculation would be based on the hypothesis:

\[
 h_0: p_{m+} - p_{m-} = 0,
\]

where \( p_{m+} \) is the treatment outcome in the biomarker-positives and \( p_{m-} \) is the expected treatment outcome in the biomarker-negatives. Then:

\[
 n = \frac{2(\bar{p})(1 - \bar{p})(Z_\beta + Z_{\alpha/2})^2}{(p_{m+} - p_{m-})^2}
\]

For adaptive threshold and adaptive signature designs interested readers may refer to papers by Freidlin and Simon.25,27

### 5.1.4 Biomarker-based strategy design

#### 5.1.4.1 With or without biomarker assessment in the control arm

The main hypothesis of the design is related to a difference in outcome between the biomarker-based strategy (\( p_{bs} \)) and the control strategy (\( p_c \)). Therefore, the sample size calculation is based on the null hypothesis of \( H_0: p_{bs} = p_c \), which is:

\[
 n = \frac{2(\bar{p})(1 - \bar{p})(Z_\beta + Z_{\alpha/2})^2}{(p_{bs} - p_c)^2}
\]

In this design, fraction of the control arm which is marker negative is actually receiving its optimal treatment which is the control treatment. This results in the dilution of the effect size between the control arm and the marker-based arm and therefore, a larger sample size is required. In other
words, the lower the marker-positivity prevalence, the higher the dilution of the effect size and therefore, the lower the efficiency of the design.

5.1.4.2 With treatment randomization in the control arm

The main hypothesis of the design is also the same as in the design without treatment randomization in the control arm. Therefore the comparison is again between the outcome in the biomarker-strategy arm ($p_{bs}$) and the outcome in the control arm ($p_c$). The sample size required for this design can be calculated with the conventional formula; however, $p_c$ will be the outcome in the randomized arm, which is the average of the outcome with experimental and with control in the overall population. In this design, irrespective of marker prevalence, half of the patients in the randomized arm receive the optimal treatment just by chance. This way we always have a 50% fixed dilution of the effect size, which results in the low efficiency of the design. (Sargent, Conley, Allegra, & Collette, 2005)

It is deducible that where the biomarker-positivity prevalence is less than 50%, biomarker-strategy design with randomization in the control arm is more efficient than the other two subtypes of the biomarker-strategy design. Both of these designs are less efficient than biomarker-stratified design and traditional design. 1,16,21,46

5.2 Comparative efficiency

There are several papers in the literature comparing the efficiency of different biomarker designs in terms of the sample size required for the designs. However, we emphasize that the hypothesis tests for each are different in nature and clinical value. Therefore, the comparative efficiencies are meant as rough guidelines only and should not be interpreted as if one design is superior or inferior to the other designs.

Simon and Maitournam47 evaluated the efficiency of an enrichment design versus a randomize all design. They found that the efficiency of the enrichment design depends on two factors; the prevalence of biomarker-positivity and the magnitude of the experimental treatment effect in biomarker-negative patients. When fewer than half of the patients are biomarker-positive and the new treatment has relatively no effect in biomarker-negative patients, the number of randomized patients required for an enrichment design is dramatically smaller than the number of randomized patients required for a randomize all design.

For example, if the experimental treatment has no effect in biomarker-negative patients, then the ratio of number of patients required for randomization in the randomize all design relative to the number required for the enrichment design is approximately $1/p^2$ where $p$ denotes the proportion of patients who are biomarker-positive. This equals a factor of 4 when half the patients are biomarker-positive.1

The treatment may have some effect on biomarker-negative patients either because the biomarker testing assay is imperfect for measuring the putative biomarker or because the drug has off-target effects. Even if the new treatment is half as effective in biomarker-negative patients as in biomarker-positive patients, the randomization ratio is approximately $4/(p+1)^2$. This equals about 2.56
when marker-positivity prevalence \((p)\) is 25%, indicating that the enrichment design reduces the number of randomized patients by a factor of 2.56. \(^{26}\)

In another paper, Lee and colleagues have performed simulations to compare efficiency of the designs in different scenarios. They assumed a case of two treatments with a binary outcome, and one binary biomarker with 50% prevalence. They compared efficiency of designs in five different scenarios of treatment and marker effects combinations including no treatment effect, treatment effect but no biomarker effect, no treatment effect but prognostic biomarker effect, predictive biomarker effect, and both prognostic and predictive biomarker effects. The simulations showed that in almost all of the scenarios and for all of the hypotheses tested, randomize all and enrichment design (in terms of number needed to be screened) have comparable efficiency and the biomarker-strategy design with randomized control arm had much lower efficiency. \(^{16}\)

Hoering and colleagues investigated the required sample sizes for enrichment, randomize all and biomarker-strategy designs as well. They explored settings in which the biomarker under study has a continuous underlying distribution. The scenarios they explored in their simulation study included (1) non-informative biomarker, (2) prognostic biomarker, (3) predictive biomarker where experimental treatment does not help biomarker-negatives more than control treatment, but has additional benefit for biomarker-positives increasing with the biomarker value, (4) predictive biomarker where effect of experimental treatment increases with biomarker value, but effect of control treatment is constant over the range of biomarker values, and (5) predictive biomarker where the experimental therapy benefits biomarker-positives but is worse than control treatment in biomarker-negatives. \(^{46}\)

Their simulations showed underlying biomarker effect distribution has a large influence on the design power. This highlights the need for a thorough investigation of biomarker properties before committing to a specific design. They showed that the enrichment design performs the best in all scenarios with an underlying true predictive biomarker. The randomize all design performs at least as efficient as biomarker-strategy design and in most cases performs better. Therefore, they recommend using the randomized all design over the strategy design except for cases in which the actual strategy hypothesis is of greater interest than the efficacy hypothesis.

Based on their results they recommend using the enrichment design if it is known with little uncertainty that the experimental treatment does not help all patients to some degree, if the biomarker prevalence is small, and if the cutpoint of biomarker positivity is relatively well established. If the cutpoint is not yet well established, the power of the study can be severely compromised. If the treatment works better for biomarker-positives but also may benefit biomarker-negatives, or if the cutpoint determining the biomarker positivity has not yet been established, and if the biomarker prevalence is large enough, they recommend using the randomized all design with the power adjusted for multiple comparisons such that both the overall and the subgroup hypothesis can be tested.

Therefore, all the above-mentioned studies concluded that in terms of efficiency randomize all and enrichment designs perform comparably. While biomarker-strategy designs perform worse or at most similar to a randomize all design.
6 A proposed strategy for design selection

In this section we propose a strategy to help researchers in choosing between the presented trial designs when planning for a study to evaluate the potential of a biomarker as a treatment selection biomarker (Figure 7). The process starts with the biomarker under study:

1. If the assay for biomarker measurement is analytically validated and the cutpoint for classifying patients as biomarker-positive and biomarker-negative is well defined, designs should be chosen that allow for confirmatory validation of a specified biomarker(s). Designs that allow confirmatory validation of biomarkers are enrichment, randomize all, biomarker-strategy or combination designs.

2. If the biomarker is not prespecified or it is specified but the cutpoint for classifying patients that benefit from treatment is not determined yet, one of the randomize all trials should be chosen, which allows for setting the cutpoint, or for exploratory identification of new biomarkers or combinations of biomarkers.

3. In both of the above mentioned situations one may use available datasets of previously completed randomize all trials to identify or validate a biomarker or biomarker combination. In this case, the analysis should be done according to a predefined analysis protocol. In case the biomarker-based treatment strategy is complex (comprises of many treatments or biomarker levels) the most efficient option is using one of the biomarker-based strategy designs. For example, Coumagen-II is an ongoing trial to evaluate the clinical impact of applying pharmacogenetic algorithms to individualize warfarin dosing. The dosing of warfarin for each individual patient is determined using a mathematical formulae consisting of genotype and clinical information. (Circulation 2012 Anderson). In this scenario, the number of biomarker levels and treatment levels to be tested are high and it would be neither feasible nor efficient to use a randomize all approach. ¹¹

4. Even if the biomarker-based treatment strategy is not complex, we still recommend researchers to explore the feasibility of conducting a biomarker-based strategy design with treatment randomization in the control arm. Although all simulation studies have shown that this design relative to a randomize all design requires much larger sample sizes, it has a brilliant capacity to provide direct evidence for all biomarker related hypotheses. The final goal of evaluating treatment selection biomarkers is designing a biomarker-based treatment strategy. In a randomize all trial both biomarker-positive and biomarker-negative patients are allocated to receive experimental or control therapy. In this situation, one may estimate the outcome in the biomarker-based strategy using the estimates of biomarker prevalence, outcome data for experimental therapy in biomarker-positive patients and outcomes with control therapy in biomarker-negative patients, and compare it to the strategy of treating all with control therapy (estimated by biomarker prevalence, outcome with control therapy in biomarker-positive and outcomes with control therapy in biomarker-negative patients). This indirectly estimated efficacy comparison may seem logical in mathematical terms, but there is extensive literature on the various ways tests may affect the response rates through
Figure 7. Flowchart of the proposed strategy for choosing between trial designs when planning for a study to evaluate the potential of a biomarker as a treatment selection biomarker.
psychosocial pathways. (Ref) For example it is observed that when pregnant women are reassured that their pelvic dimensions (measured in MRI images) are suitable for normal delivery, they were more tolerant to pain and continued longer to have a normal delivery. (Ref) Therefore, the indirectly estimated effect of biomarker-based treatment strategy does not necessary equate its direct estimate. Although biomarker-based strategy designs may need a larger sample size than randomize all designs, they provide more realistic estimates of the impact of biomarker-based treatment selection in the population.

5. If it is not feasible to recruit the sample size required for a biomarker-based strategy design, the next best choice is one of the designs of the randomize all category. However, in contexts where there is a strong biological basis for believing that biomarker-negative patients will not benefit from the new treatment and including them would raise ethical concerns, the only option is testing the new therapy in only biomarker-positive patients using an enrichment design. In many situations, the biological basis is strong but not compelling. The enrichment design does not provide data on the effectiveness of the new treatment compared with control for biomarker-negative patients. Consequently, unless there is compelling biological or Phase II data that the new treatment is not effective in biomarker-negative patients, the enrichment design may not provide adequate data for regulatory approval of the biomarker.⁴

6. Trials in the randomize all category are efficient and provide indirect evidence for all biomarker-related hypotheses. In some specific contexts there are special types of randomize all design which have an improved efficiency. For example in case the outcome can be assessed in a relatively short period of time and the trial has a good infrastructure for data monitoring and interim analysis, application of outcome-based adaptive randomization scheme may result in a more efficient trial. By outcome-based adaptive randomization, more patients are allocated to more effective treatments as the trial progresses and information accumulates. Therefore the trial gains efficiency through modeling and appropriate use of prior information.

Another example is in settings where the biological basis suggests that experimental treatment is not beneficial for biomarker-negative patients. In this scenario, there are three possible options;

i. A randomize all trial

ii. A two-stage trial, starting with accruing just biomarker-positive patients and doing an interim analysis at the end of stage I. If the results are not promising for the new treatment, accrual stops and no treatment effect is claimed. Otherwise, accrual continues with recruiting unselected population. Liu and colleagues have proposed this design and called it “two-stage sample-enrichment” design.³¹

iii. An “adaptive parallel two-stage” design in which case the trial starts with a randomize all design using a biomarker-stratified randomization. If the interim analysis at the end of stage I shows promising results for both biomarker-positive and biomarker-negative patients, trial continues with recruiting unselected population. While if it shows treatment effect only in biomarker-
positive patients, recruitment continues only for biomarker-positive patients. And if trial shows no evidence of efficacy in stage I, the study stops.⁴⁸

7 Case studies from ZonMW trials (HYPITAT and EORTC)
We applied the above mentioned statistical techniques for exploring possible treatment selection biomarkers in two already conducted and reported trials; HYPITAT (Hypertension and Pre-eclampsia Intervention Trial At Term)⁴⁹ and EORTC- 55971 (Neoadjuvant Chemotherapy or Primary Surgery in Stage IIIC or IV Ovarian Cancer)⁵⁰.

On these medium-sized already existing datasets of randomize all trials, we explored the potential of recorded participants’ baseline characteristics for treatment selection. In HYPITAT trial we explored the association of two cervical favorability measures (cervical length and Bishop score) with the benefit obtained from intervention using test of interaction and marker-by-treatment predictiveness curves. The paper of this successful exploratory analysis is now published.⁵¹ In EORTC trial we tested the potential of a short list of five baseline characteristics for treatment selection using STEPP methodology. We then explored possibility of developing a multivariable treatment selection algorithm. Application of such analyses if done correctly and with a prespecified analysis plan will result in new insights and hypotheses for treatment selection that can be replicated or validated in other independent datasets or new trials.

7.1 HYPITAT - Cervical favorability and the decision for labor induction in gestational hypertension or mild preeclampsia at term

Objective: To examine whether cervical favorability (measured by cervical length and the Bishop score) should inform obstetricians’ decision regarding labor induction for women with gestational hypertension or mild preeclampsia at term.

Design: A posthoc analysis of the Hypertension and Pre-eclampsia Intervention Trial At Term (HYPITAT).⁴⁹

Setting: Obstetric departments of 6 university and 32 teaching and district hospitals in the Netherlands.

Population: 756 women diagnosed with gestational hypertension or preeclampsia between 36+0 and 41+0 weeks of gestation randomly allocated to induction of labor or expectant management.

Methods Data were analyzed using logistic regression modeling.

Main outcome measures: The occurrence of a high risk maternal situation defined as either maternal complications or progression to severe disease. Secondary outcomes were caesarean delivery and adverse neonatal outcomes.

Results: The superiority of labor induction in preventing high risk situations in women with gestational hypertension or mild preeclampsia at term varied significantly according to cervical
favorability. In women who managed expectantly, the longer the cervix the higher the risk of developing maternal high risk situations, while in women in whom labor was induced, cervical length was not associated with a higher probability of maternal high risk situations (test of interaction p=0.03; Figure 8). Similarly, the beneficial effect of labor induction on reducing the caesarean section rate was stronger in women with an unfavorable cervix.

**Conclusion:** Against widely held opinion, our exploratory analysis showed that women with gestational hypertension or mild preeclampsia at term who have unfavorable cervix benefited more from labor induction than other women.

**Figure 8.** Plots of the primary and secondary outcomes of HYPITAT trial as a function of the cervical length; A, risk of maternal high risk situations, B, risk of caesarean delivery, C, risk of developing adverse neonatal outcomes. Lines illustrate model-based estimated risk with 95% confidence intervals (dashed lines). Small circle and triangle markers show the observed risks in the studied patients.
7.2 EORTC 55971 - Primary surgery or neoadjuvant chemotherapy in stage IIIC or IV ovarian cancer

*Background:* In patients with advanced ovarian cancer primary debulking surgery followed by chemotherapy (PDS-CT) performs on average the same as neoadjuvant chemotherapy and interval debulking surgery followed by postsurgical chemotherapy (NACT). Since there may be contraindications to their use, based on side-effects or cost, we investigated subgroups in which NACT may be more or less beneficial.

*Patients and methods:* The international EORTC 55971 trial randomized 718 women to either PDS-CT or NACT. We assessed whether age, CA125 at study entry, WHO performance status, largest metastatic tumor size and high-grade serous tumor histology predicted the magnitude of the effect of PDS-CT compared with NACT. Subpopulation treatment effect pattern plot (STEPP) analysis was performed for the 1 and 5 year survival rate.

*Results:* Individually, none of the biomarkers significantly predicted differential treatment effects for one-year survival. However, women whose tumors were stage IV or had metastatic tumor larger than 100 mm had lower 5-year survival with PDS-CT compared to NACT (p-value of interaction 0.024 and 0.044, respectively- Figure 9). STEPP analysis of a composite measure of prognostic risk revealed no pattern of association between risk score and comparative efficacy of the two treatment options (Figure 10).

*Conclusion:* Patients with a stage IV tumor and/or at least one metastatic tumor larger than 100 mm benefit more from treatment by NACT.

**Figure 9.** Subpopulation treatment effect pattern plots
Figure 10. Subpopulation treatment effect pattern plots of the composite risk score
8 Concluding remarks

The 2012 policy statement of the Dutch Ministry of Health, Welfare and Sport (VWS) starts with the following statement “In the past few years, healthcare costs have risen faster than the economy growth, increasing pressures on public spending”. The concerns about rising health care costs is not limited to the Netherlands. Many other western countries are facing similar trends and hence similar challenges. One strategy for cost containment would be to limiting interventions to individuals who benefit from them. If we could be able to identify individuals who would benefit from an intervention in advance, then we would be able to confine the intervention provision to only persons who are predicted to benefit from it. This will result in cutting down the costs of health care provision without compromising the obtained health gains.

The advent of new technologies such as all “omics” platforms (genomics, proteomics, metabolomics and so on) and novel imaging tools has allowed a much more detailed characterization of individual patients at the time a decision is to be made for an intervention. More detailed biological characterization helps in disentangling the heterogeneity in the benefit patients’ gain from interventions. In this new era, the traditional paradigm for the design and analysis of randomized clinical trials which compares the aggregated treatment effects in the overall study population is not efficient anymore.

We believe that exploring the heterogeneity in benefits gained from interventions should be systematically incorporated into the design phase of clinical trials. The availability of more robust statistical techniques allow for such analyses. Therefore, when a trial ends, the conclusion is not limited to a statement on the aggregated intervention effect. It also includes a statement whether the intervention is beneficial in all individual participants. If the trial has shown that the intervention has not been beneficial for all, then researchers report on whether they could successfully develop a validated classification algorithm for identifying individuals who would benefit from the intervention using baseline biomarker information of the trial participants.

Successful development of valid treatment selection algorithms often requires two foundations. First, large enough sample sizes to provide adequate statistical power, particularly in view of the multiple-testing inherent in biomarker analyses. Moving toward forming large-scale consortia of clinical trials would be a big step toward reaching this foundation. Second, baseline information databases which are rich in terms of factors which explain the heterogeneity in treatment benefits. Considering the rapidly growing field of explanatory biomarkers, this obstacle might be moderated by generating biobanks to accompany clinical trials. Biobanks allow precise future biomarker investigations, which may be beyond imagination or technical possibilities at the time of collection. The presence of large RCT consortia associated with well-designed biobanks would have the capacity to perform as a laboratory, allowing for continued discoveries, validation and updating of treatment selection algorithms. This approach can potentially save a lot of time and expenses required for conducting exploratory or replication trials for each newly proposed biomarker.

In the years to come, investigators designing and conducting clinical trials have an ethical and financial responsibility to maximize the scientific knowledge gained from the participation of the subjects.
9 Tables
Table 1. Patient flow of the extracted design labels for trials evaluating prognostic and predictive biomarkers. Labels clustered in the same bullet describe identical trial designs

<table>
<thead>
<tr>
<th>Design labels</th>
<th>Patients</th>
<th>Intervention</th>
<th>Comparison</th>
<th>Patient Flow Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-arm Uncontrolled cohort pharmacogenetic study</td>
<td>M', M</td>
<td>Exp</td>
<td>-</td>
<td>1. Single arm</td>
</tr>
<tr>
<td>Enrichment, Marker enrichment, Biomarker-Enrichment, Targeted, Efficient targeted, Selection, RCT of Test positives, Gene-enrichment</td>
<td>M</td>
<td>Exp</td>
<td>Std</td>
<td>2. Enrichment</td>
</tr>
<tr>
<td>Randomize all, Traditional, Conventional, Completely randomized, simple randomization, untargeted, Test as baseline in RCT, Prospective/retrospective, Biomarker analysis within an existing RCT, Controlled cohort pharmacogenetic study</td>
<td>M', M *</td>
<td>Exp</td>
<td>Std</td>
<td>3. Randomize all</td>
</tr>
<tr>
<td>Marker strategy, Biomarker-stratified, Stratified randomized, Separate randomization, Stratification, Non-targeted RCT (stratified by marker), Indirect predictive biomarker-based, Planned analysis, Marker by treatment interaction, Treatment by marker interaction, Marker×treatment interaction, Interaction</td>
<td>M', M</td>
<td>Exp</td>
<td>Std</td>
<td>4. Biomarker-strategy</td>
</tr>
<tr>
<td>Marker strategy, Biomarker-strategy, Direct predictive biomarker-based, Random disclosure, Classical, Customized strategy, Parallel controlled pharmacogenetic study, Marker-based strategy design I, Strategy, Classifier randomization design, Biomarker-guided</td>
<td>M', M</td>
<td>M-based strategy</td>
<td>Std</td>
<td>a. with biomarker assessment in the control arm</td>
</tr>
<tr>
<td>Biomarker-strategy without biomarker assessment in the control arm, Biomarker-strategy design with standard control, RCT of testing, Test-treat, Parallel controlled pharmacogenetic diagnostic study, Customized</td>
<td>Unknown</td>
<td>M-based strategy</td>
<td>Std</td>
<td>b. without biomarker assessment in the control arm</td>
</tr>
<tr>
<td>Modified marker-based strategy, Marker-based strategy design II, Marker-strategy, Augmented strategy</td>
<td>M', M</td>
<td>M-based strategy</td>
<td>Exp, Std</td>
<td>c. with treatment randomization in the control arm</td>
</tr>
</tbody>
</table>

Table 2. List of questions that can be answered by the trials of each patient flow category

<table>
<thead>
<tr>
<th>Questions trial can answer</th>
<th>Single-arm</th>
<th>Enrichment</th>
<th>Randomize all</th>
<th>Biomarker-strategy with biomarker measurement in the control arm</th>
<th>Biomarker-strategy without biomarker measurement in the control arm</th>
<th>With treatment randomization in the control arm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How does the experimental treatment compare to the control treatment in biomarker-negatives?</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓ indirect</td>
<td>✓</td>
</tr>
<tr>
<td>How does the experimental treatment compare to the control treatment in biomarker-positives?</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>How does the experimental treatment compare to the control treatment in overall study population?</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>How does the biomarker-based treatment strategy compare to the control treatment in the overall study population?</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>How does the biomarker-based treatment strategy compare to the experimental treatment in the overall study population?</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>indirect</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Biomarker effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is the biomarker status associated with the outcome in the standard of care group? (Is the biomarker prognostic?)</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>indirect</td>
</tr>
<tr>
<td>Is the biomarker status associated with the outcome in the experimental treatment group?</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Biomarker by treatment effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is the biomarker status associated with a benefit of experimental treatment? (Is the biomarker is predictive?)</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
</tbody>
</table>
10 References


