# Randomized Discontinuation Trials with Binary Outcomes

Valerii V. Fedorov[a,*], Tao Liu[b]

[a] *Quintiles, 5927 South Miami Boulevard, Morrisville, NC 27560, USA*
[b] *Department of Biostatistics, Center for Statistical Sciences, Brown University, Providence, RI 02912, USA*

**Abstract**

Randomized discontinuation trial (RDT) has gained popularity across a number of therapeutic areas. Oncology is one of the most known. In the simplest case, at the initial open-label stage all patients are treated with the experimental treatment to identify a population of responders. This stage is followed by a randomized two-arm trial to compare two conditions, e.g. treatment versus placebo. Potentially RDT increases the efficiency of trials relatively to traditional designs gaining information from a sensitized population if the open-label stage provides a reliable separation of responders from non-responders. Often a sensitized population is called an enriched population and respectively the RDT is called "enrichment experiment". We compare RDT with the traditional two-arm randomize clinical trials (RCT) for binary outcomes assuming that the population of interest consist of three groups: placebo responders, treatment-only responders and non-responders. Our results are derived in the "parameter estimation" setting and they are based on the comparison of estimator variances. We identify conditions under which RDT is either superior or inferior to RCT in terms of response rates, misclassification rates, and clinical ethics. Extension of our results to design optimization, hypothesis testing, and sample size calculation are rather straightforward.

*Keywords:* Enrichment experiments; Experimental design; Randomized discontinuation trial; Open-label first stage.

## 1. Introduction

The first description of *randomized discontinuation trial* (RDT) can be traced back to 1970's, when Amery and Dony [1] proposed this design over the classical placebo-controlled *randomized clinical trial* (RCT) to reduce the duration and degree of subjects' exposure to placebo. This design can be viewed as a special case of data enrichment strategies [11, 16]. In the simplest case, RDT allocates all qualified subjects to an active treatment at the first stage (also called the open-label stage/phase). Non-responders to the treatment and those showing serious adverse effect are excluded from further study. At the second phase, the open-label stage responders are randomized to the same active treatment or placebo. Potentially RDT can be superior to traditional RCT because of the open-label stage serves as a filter for removing patients who would be unethically exposed to placebo and do not contribute information about the part of population for whom the treatment can be useful, and after excluding these patients, the second stage distinguishes better whether the treatment adds anything over the placebo effect. RDT relies on several assumptions to be a valid design. One assumption is that "the treatment will not cure the condition during the open-label stage" [1]. For this reason, RDT is generally applied under conditions that require sustained use of a treatment such as in treating chronic diseases and stabilizing tumor growth [4]. Another assumption is that the treatment effect(s) of the open-label stage will not be carried over to the second stage. This assumption can be satisfied

*Corresponding author. Tel: (919) 998-1855.
*Email addresses:* Valerii.Fedorov@quintiles.com (Valerii V. Fedorov), tliu@stat.brown.edu (Tao Liu)

through the set-up of adding a wash-out period between the two stages. There exist a number of modifications of RDTs [8, 10, 13, 15, 17, 18, 22, 23] and other assumptions are often needed.

Capra [3] compared the power RDT with that of RCT, when the primary endpoints are subjects' survival times. Kopec et al. [12] evaluated the utility and efficiency of RDT when the endpoints are binary. They compared the relative sample size required for a fixed power of RDT versus RCT under different scenarios and parameter settings. Their comparisons however are based on the outcomes solely from the second stages, treating the open-label stage as a screening process. This simplifies the statistical analysis, but the information contained in the open-label stage is mostly wasted.

We propose the approach that allows to directly use data from both stages. The approach is based on relatively simple model and the efficiency of various designs is assessed by comparing the asymptotic variances of the respective *maximum likelihood estimators* (MLEs).

This paper is the updated but shorter version of the technical report [7], the latter is available upon request.

## 2. Notations and model

Let the population of interest consist of the following three exclusive sub-populations: placebo responders, treatment-only responders and non-responders. We assume that the placebo responders also respond to the active treatment, i.e. {treatment responders} = {placebo responders} ∪ {treatment-only responders}. For the sake of simplicity we consider binary responses and all our derivations are done in "parameter estimation" setting. We avoid the discussion of how the open-label stage responders are identified. However we introduce probabilities of false positive and false negative classification ($p_1$ and $p_2$ respectively) that makes possible to apply our results to various response models including dichotomized continuous or multi-category responses (cf.[9]) .

Let the fractions of treatment responders and placebo responders be $\pi_+$ and $\pi_p$, respectively. The treatment effect thus can be expressed as the difference between these two fractions $\pi_t = \pi_+ - \pi_p$, or their ratio $R = \pi_p/\pi_+$. Note that $\pi_t$ is the fraction of treatment-only responders.

An RCT is set up as the following. Suppose for a sample of $N$ patients, we randomize $\kappa N$ to placebo and the rest $(1-\kappa)N$ to the treatment arm. Let $n_+$ and $n_p$ be the numbers of responders in the treatment arm and in the placebo arm, respectively. We use the RCT with equal arm allocations ($\kappa = 0.5$) as the reference for evaluating the trial efficiency.

For a fair comparison, we assume that RDT recruits the same number of subjects, $N$. RDT assigns all of them to the treatment at the open-label stage. Let $n_0$ be the number of open-label-stage responders. At the second stage, we randomize $n_{0p} = \gamma n_0$, where $\gamma$ is pre-specified, of the open-label stage responders to the placebo arm, and the rest $n_{0t} = n_0 - n_{0p}$ to the treatment arm. At the end of the second stage, let $n_{1p}$ be the number of responders in the placebo arm and $n_{1t}$ in the treatment arm. Figure 1 visualizes the setup of RDT. To add to the fairness of comparison of RDT and RCT, we assume that the open-label stage takes a shorter time and therefore only some surrogate endpoint [2] can be measured. This endpoint does not provide the result identical to the measurements at the end of RDT. For instance, in oncology the previous one can be an indicator of the tumor size change while the actual endpoint (the same as in RCT) can be cure/non-cure. In most cases "cure" sounds too optimistic and should replace it with either "tumor-free" or "complete responder". The introduction of false positive and false negative classification helps to address this problem.

## 3. Randomized clinical trial

For RCT, the likelihood function is

$$L_{RCT} = (\pi_+)^{n_+}(1-\pi_+)^{(1-\kappa)N-n_+}\pi_p^{n_p}(1-\pi_p)^{\kappa N-n_p}. \tag{1}$$

One can readily derive the MLE of $\pi_+$ and $\pi_p$ and their variances [14, cf. Ch.6]:

$$\hat{\pi}_+ = \frac{n_+}{(1-\kappa)N} \quad \text{and} \quad \mathrm{Var}[\hat{\pi}_+] = \frac{\pi_+(1-\pi_+)}{(1-\kappa)N}, \tag{2}$$

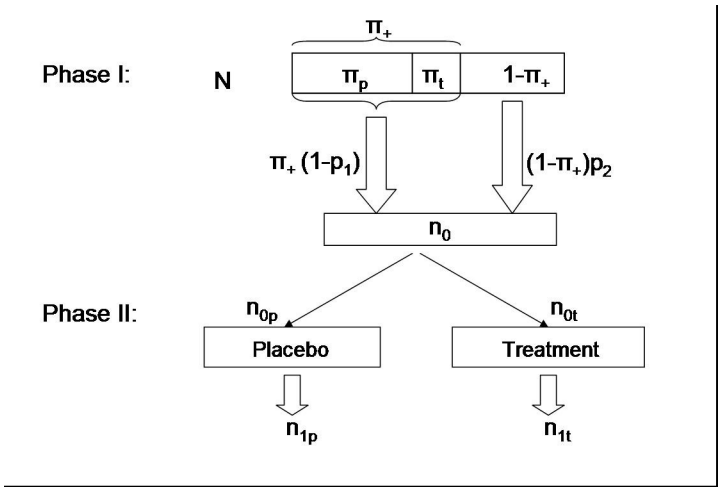



Figure 1: Diagram of Randomized Discontinuation Trial with misclassification

$$\hat{\pi}_p = \frac{n_p}{\kappa N} \quad \text{and} \quad \mathrm{Var}[\hat{\pi}_p] = \frac{\pi_p(1-\pi_p)}{\kappa N}. \tag{3}$$

We used the fact that asymptotically ($N \to \infty$), the variance of the MLE is $M^{-1}$, where $M$=Var$[\nabla \ln L]$ is the Fisher information matrix. Actually, for the estimators defined by (1) formulae (2) and (3) are exact and their information matrix $M$ is diagonal. However in what follows we face more complicated likelihood functions and we work with the asymptotic variance-covariance matrices, if it is not stated otherwise. From (2) and (3), it immediately follows that

$$\hat{\pi}_t = \hat{\pi}_+ - \hat{\pi}_p \quad \text{and} \quad \mathrm{Var}[\hat{\pi}_t] = \frac{1}{N}\left(\frac{\pi_+(1-\pi_+)}{1-\kappa} + \frac{\pi_p(1-\pi_p)}{\kappa}\right), \tag{4}$$

$$\hat{R} = \frac{\hat{\pi}_p}{\hat{\pi}_+} = \frac{n_p(1-\kappa)}{n_+\kappa} \quad \text{and} \quad \mathrm{Var}[\hat{R}] \cong \frac{R^2}{N}\left(\frac{1-\pi_+}{\pi_+(1-\kappa)} + \frac{1-\pi_p}{\pi_p\kappa}\right) . \tag{5}$$

To find Var$[\hat{R}]$ we used the "delta method" (the first order approximation, see [14]). Note that

$$\mathrm{Prob}[\hat{\pi}_+ - \hat{\pi}_p] = P_N > 0$$

and therefore with a non-zero probability the MLE of $\pi_t$ may be negative. However $\lim_{N\to\infty} P_N = 0$ and for a reasonably large $N$ this probability can be neglected, see also the simulation results from Section 7. In what follow we continue to assume that the probability of getting negative estimates for the positive parameters ($\pi_t$, $R$, etc.) can be neglected.

## 4. Randomized discontinuation trial

### *4.1. Model and likelihood estimator*

As it was mentioned before, by allowing outcome misclassifications of responders in the open-label stage of RDT, we try to take into account the fact that at the end of the open-label stage only surrogate endpoints are available to detect responders, and usually surrogate endpoints allow the less accurate classification than the primary endpoints. So, at the open-label stage the probability that a subject will be labeled as a responder equals

$$\zeta_+ = \pi_+(1-p_1) + (1-\pi_+)p_2 \ ,$$

where $p_1$ is the probability of false labeling a true responder as a non-responder and $p_2$ is the probability of false labeling a true non-responder as a responder. We assume that at the second stage the actual end point is observed with no measure error. At this stage, the probability that a subject sampled from the first stage responders is a true responder (and therefore will respond to treatment at the second stage) is $\zeta_{1t} = \pi_+(1-p_1)/\zeta_+$. A subject sampled from the same set of subjects, i.e. from the subjects that are labeled as the first stage responders, will respond to placebo with probability $\zeta_{1p} = \pi_p(1-p_1)/\zeta_+$.

The likelihood function, which includes the observation from both stages, can be written as

$$L_{RDT} = \zeta_+^{n_0}(1-\zeta_+)^{N-n_0} \cdot \zeta_{1p}^{n_{1p}}(1-\zeta_{1p})^{n_{0p}-n_{1p}} \cdot \zeta_{1t}^{n_{1t}}(1-\zeta_{1t})^{n_{0t}-n_{1t}}, \tag{6}$$

where $N, n_0,\ n_{0p},\ n_{1p},\ n_{0t},\ n_{1t}$ are defined in Figure 1.

The maximum likelihood estimators for $\zeta = (\zeta_+, \zeta_{1t}, \zeta_{1p})^T$ are

$$\hat{\zeta}_+ = \frac{n_0}{N}\ , \quad \hat{\zeta}_{1p} = \frac{n_{1p}}{n_{0p}}\ , \quad \hat{\zeta}_{1t} = \frac{n_{1t}}{n_{0t}}\ , \tag{7}$$

their Fisher information is

$$M = N \begin{bmatrix} \frac{1}{\zeta_+(1-\zeta_+)} & 0 & 0 \\ 0 & \frac{\gamma\zeta_+}{\zeta_{1p}(1-\zeta_{1p})} & 0 \\ 0 & 0 & \frac{(1-\gamma)\zeta_+}{\zeta_{1t}(1-\zeta_{1t})} \end{bmatrix} \tag{8}$$

and their variance-covariance matrix is $V = M^{-1}$.

*4.2. Estimation of R*

From the definition of $\zeta_{1t}, \zeta_{1p}$ and (7) it follows that the MLE of $R$ is

$$\hat{R} = \frac{\hat{\zeta}_{1p}}{\hat{\zeta}_{1t}} = \frac{n_{1p}}{n_{1t}} \cdot \frac{n_{0t}}{n_{0p}} = \frac{n_{1p}}{n_{1t}} \cdot \frac{1-\gamma}{\gamma}\ , \tag{9}$$

and using (8) one can verify that

$$\mathrm{Var}[\hat{R}] \cong \frac{R^2}{N\zeta_+}\left(\frac{1-\zeta_{1p}}{\gamma\zeta_{1p}} + \frac{1-\zeta_{1t}}{(1-\gamma)\zeta_{1t}}\right)\ . \tag{10}$$

The efficiency comparison between RCT and RDT for estimating $R$ when $p_1 = p_2 = 0.10$ is shown in Figure 2. Along each curve presented at in this figure variances of $\hat{R}$ for RCT and RDT are equal, i.e. given $\gamma$ the curve is defined by the equation

$$\frac{1-\pi_+}{\pi_+(1-\kappa)} + \frac{1-\pi_p}{\pi_p\kappa} = \frac{1-\zeta_{1p}}{\gamma\zeta_{1p}} + \frac{1-\zeta_{1t}}{(1-\gamma)\zeta_{1t}}\ ,$$

see (5) and (10). For each $\gamma$ the region below the curve towards the $\pi_p$ -axes is where RDT is more efficient.

At $\gamma = 0.5$, RDT is more efficient in a relatively large region. Obviously from the statistical point of view, one should recommend to use the larger $\gamma$. However, the ethical considerations call for smaller $\gamma$. For small values of $\gamma$, RDT is more efficient only in detecting small treatment effect, $\pi_t$.

When $p_1 = p_2 = 0$, formula (10) reduces to

$$\mathrm{Var}[\hat{R}] \cong \frac{R(1-R)}{N\gamma\pi_+}. \tag{11}$$

Figure 3 shows the comparison between RDT and RCT in estimating $R$ for various values of $\gamma$. Except scenarios with the smaller fraction of patients on placebo than on treatment, RDT performs better (statistically, not ethically!) than RCT in a rather broad region; and this superiority is guaranteed for all $\gamma \geq 0.5$. Note that the selection of the large $\gamma$ means that at the second stage we deprive the large number of subjects responding to treatment of that potentially efficacious treatment.

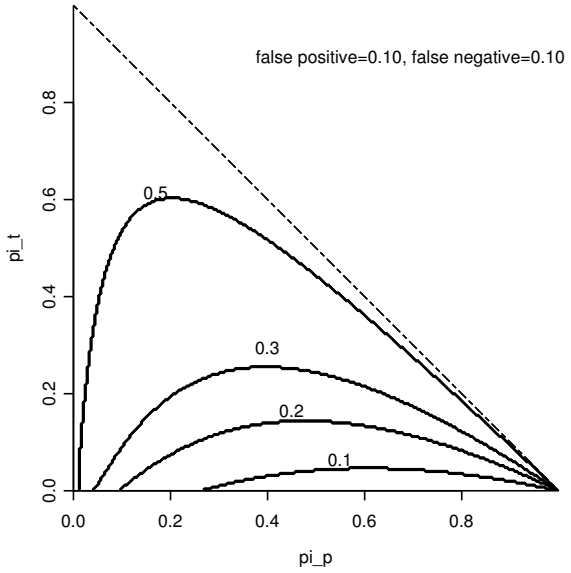



Figure 2: Comparison of RCT and RDT for estimating $R$ when $p_1 = p_2 = 0.10$. Numbers next to the separating lines denote the fraction of patient randomized to placebo ($\gamma$). The region under each curve is the condition when a RDT is more efficient.

*4.3. Estimation of $\pi_t$*

To find the MLE of $\pi_t$, we need to express $\pi_t$ through $\zeta_+$, $\zeta_{1p}$ and $\zeta_{1t}$ and then to replace the latter by their estimators defined in (7). Obviously we cannot estimate more than three parameters. Usually the misclassification rates $p_1$ and $p_2$ can be viewed as nuisance parameters. One way to avoid the non-identifiability issue is to impose at least one constraint on these parameters. Note that for the RCT case we do not need to estimate neither $p_1$ or $p_2$. It is assumed at the end of trials the actual end point is observed and there are no misclassification errors, see also comments at the beginning of Section 4.1.

When $p_1$ and $p_2$ are known, we have

$$\hat{\pi}_t = \frac{\hat{\zeta}_+(\hat{\zeta}_{1t} - \hat{\zeta}_{1p})}{1 - p_1} \ , \tag{12}$$

and its asymptotic variance is

$$\mathrm{Var}[\hat{\pi}_t] \cong \frac{\zeta_+}{N(1-p_1)} \left[ (\zeta_{1t} - \zeta_{1p})^2 (1 - \zeta_+) + \frac{\zeta_{1p}(1-\zeta_{1p})}{\gamma} + \frac{\zeta_{1t}(1-\zeta_{1t})}{1-\gamma} \right].$$

In Figure 4, we compare RDT with RCT for estimating $\pi_t$ when $p_1 = p_2 = 0.10$. For these misclassification rates, the superiority of RDT is guaranteed for $\gamma = 0.5$ for almost entire range of $\pi_t$ and $\pi_p$. For smaller values of $\gamma$, we see that RDT is more efficient only for "small" $\pi_t$ and small $\pi_p$.

For the simplest (albeit not very realistic) scenario, when $p_1 = p_2 = 0$, the variance for $\hat{\pi}_t$ is

$$\mathrm{Var}[\hat{\pi}_t] \cong \frac{1}{\pi_+ N} \left[ (\pi_+ - \pi_p)^2 (1 - \pi_+) + \frac{\pi_p (\pi_+ - \pi_p)}{\gamma} \right] . \tag{13}$$

One can notice that, $\mathrm{Var}[\hat{\pi}_t]$ achieves its minimum, $\pi_t(1-\pi_t)/N$ at $\gamma = 1$, i.e. assigning all the open-label-stage responders to the placebo arm. This choice of $\gamma$ provides the most informative but

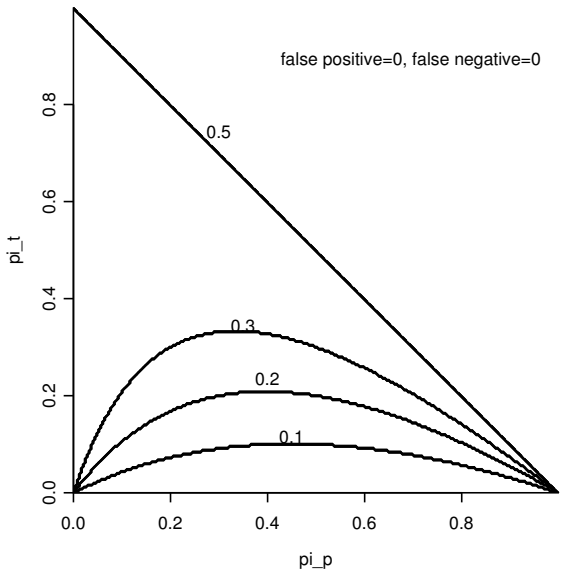



Figure 3: Comparison of RCT and RDT for estimating $R$ for various $\gamma$ when $p_1 = p_2 = 0$. The number assigned to each curve is the value of $\gamma$ and the region under the curve is where RDT is more efficient.

the least ethical trial design. Actually for any $\gamma \geq 0.5$, RDT is always more efficient (allows the better estimation of $\pi_t = \pi_+ - \pi_p$ ) than RCT. Note that the later statement is true only in the case of no misclassification. Figure 5 compares the efficiency between RDT and RCT for the this case.

Curiously enough when $p_2$ is unknown but $p_1$ is known, the estimate of $\pi_t$ and its asymptotic variance is the same as for (12).

When the rate of false positive misclassification $p_2$ is known but $p_1$ is not, then the MLE for $\pi_t$ is

$$\hat{\pi}_t = \hat{\zeta}_+(\hat{\zeta}_{1p} - \hat{\zeta}_{1t}) \left( \frac{1 - \hat{\zeta}_{1t}}{p_2 \hat{\zeta}_{1t}} - \frac{1}{\hat{\zeta}_{1t}\hat{\zeta}_+} \right), \tag{14}$$

with the asymptotic variance for $\hat{\pi}_t$ is

$$\mathrm{Var}[\hat{\pi}_t] \cong \frac{1}{N} \left( A^2 \zeta_+(1-\zeta_+) + \frac{B^2 \zeta_{1p}(1-\zeta_{1p})}{\gamma \zeta_+} + \frac{C^2 \zeta_{1t}(1-\zeta_{1t})}{(1-\gamma)\zeta_+} \right)$$

where

$$A = \frac{1-\zeta_{1t}}{p_2} \left( 1 - \frac{\zeta_{1p}}{\zeta_{1t}} \right),\ B = \frac{p_2 - \zeta_+(1-\zeta_{1t})}{p_2 \zeta_{1t}}\ ,\ C = \frac{\zeta_+}{1-\zeta_{1t}} A + \frac{\zeta_{1p}}{\zeta_{1t}} B\ .$$

So far to avoid the non-identifiability problem we assume that either $p_1$ or $p_2$ is known, or both of them are known. One may also assume that the value of some simple function of $p_1$ and $p_2$, for instance, $q = p_1/(1-p_2)$ is known. In this case

$$\hat{\pi}_t = \frac{(\hat{\zeta}_{1t} - \hat{\zeta}_{1p})q}{1 - \hat{\zeta}_{1t}(1-q)}$$

and

$$\mathrm{Var}[\hat{\pi}_t] = \frac{q^2 \left( \frac{\gamma(1-\zeta_{1t})\zeta_{1t}(1-\zeta_{1p}(1-q))^2}{1-\gamma} + \frac{(1-\gamma)(1-\zeta_{1p})\zeta_{1p}(1-\zeta_{1t}(1-q))^2)}{\gamma} \right)}{\zeta_+ N (1-\zeta_{1t}(1-q))^4}\ .$$

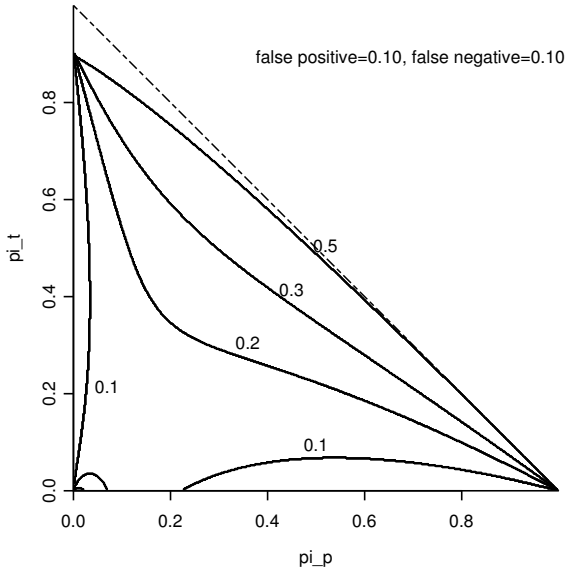



Figure 4: Comparison of RCT and RDT for estimating $\pi_t$ when $p_1 = p_2 = 0.10$ for various values of $\gamma$. The number assigned to each curve is the value of $\gamma$. In the region(s) under and left to the curve(s) towards the two axes RDT is more efficient.

Another way to avoid non-identifiability is to be less demanding and to estimate $\pi_t^- = \pi_t(1 - p_1)$ instead of $\pi_t$. Its MLE is $\hat{\pi}_t^- = (\hat{\zeta}_{1t} - \hat{\zeta}_{1p})/\hat{\zeta}_+$ and it includes neither $p_1$ nor $p_2$, see (7). If one manages to prove that $\pi_t^-$ is statistically and clinically significant, the same will be true for $\pi_t$.

## 5. Comparison between optimal RDT and optimal RCT

The optimal design can trim down the budget for its smaller sample size requirement, speed up the drug development process for a shorter recruitment period, and reduce the chance of exposing human subjects to possible inferior treatments and inactive placebo. In general, an optimal design depends on unknown parameters ($\pi_t$, $\pi_p$, $p_1$, $p_2$). Their wrong guestimates may lead to a design that is not optimal for the true values of unknown parameters. Following the optimal design theory [6, p.117] it is more accurate to call the designs, which will be reported below, by "locally optimal designs". Similar to Fedorov and Atkinson [5], one can explore robustness of locally optimal design with respect to potential uncertainties in initial guesses.

In Table 1, we give the optimal design conditions for RDT and RCT in term of $\kappa$ and $\gamma$ such that the variance of the parameter ($\pi_t$ or $R$) estimate is minimal. Table 2 summarizes the estimator variances under these optimal conditions.

Note that the optimal randomization rates depend on unknown parameters ($\pi_t$, $\pi_p$, $p_1$, $p_2$) or at least some of them. If one is interested in the estimation of two or more parameters then optimality criteria that depend on their variance-covariance matrices should be used, see [6, Ch.2].

If there is no misclassification, RDT is optimized by assigning all patients to placebo at the second stage (i.e. $\gamma = 1$). *The optimal RDT in this case is always superior to optimal RCT.* Indeed, let $\mathrm{Var}^*[\hat{\theta}]$ be the variance of $\hat{\theta}$ for optimal design. Then:

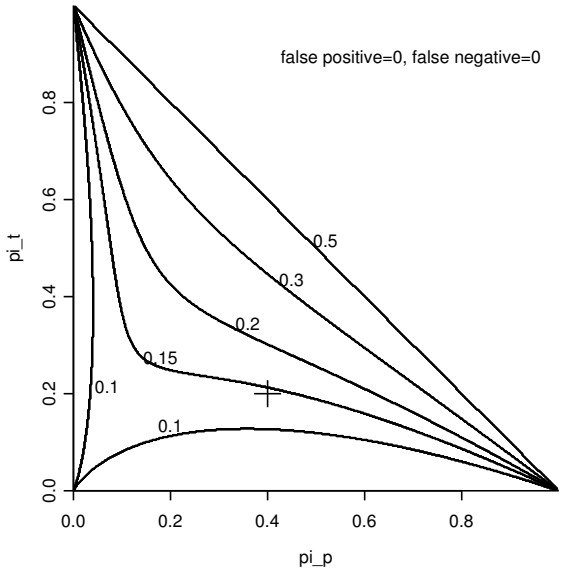



Figure 5: Comparison of RCT and RDT for estimating $\pi_t$ when $p_1 = p_2 = 0$ for various $\gamma$. The region(s) under and left to the curve(s) is where RDT is more efficient. The number assigned to each curve is the value of $\gamma$.

- the variance of $\hat{\pi}_t$

$$\begin{aligned}
\mathrm{Var}^*_{RCT}(\hat{\pi}_t) &= \frac{1}{N}\left(\sqrt{\pi_p(1-\pi_p)} + \sqrt{\pi_+(1-\pi_+)}\right)^2 \\
&\geq \frac{1}{N}[\pi_p(1-\pi_p) + (\pi_p+\pi_t)(1-(\pi_p+\pi_t))] \\
&= \frac{1}{N}[2\pi_p(1-\pi_+) + \pi_t(1-\pi_t)] \geq \frac{\pi_t(1-\pi_t)}{N} = \mathrm{Var}^*_{RDT}(\hat{\pi}_t)\ ,
\end{aligned}$$

  see comments to (13).

- the variance of $\hat{R}$

$$\begin{aligned}
\mathrm{Var}^*_{RCT}(\hat{R}) &\cong \frac{1}{N}\left(\sqrt{\frac{R(1-\pi_p)}{\pi_+}} + \sqrt{\frac{R^2(1-\pi_+)}{\pi_+}}\right)^2 \\
&\geq \frac{1}{N}\left(\frac{R(1-\pi_+R)}{\pi_+} + \frac{R^2(1-\pi_+)}{\pi_+}\right) \\
&= \frac{1}{\pi_+ N}R[1-R+2R(1-\pi_+)] \geq \frac{R(1-R)}{\pi_+ N} \cong \mathrm{Var}^*_{RDT}(\hat{R})\ ,
\end{aligned}$$

  compare with (11).

## 6. Hypothesis testing and sample size calculation

For relatively large $N$ and moderate $\pi_p$ and $\pi_t$, the central limit theorem [21] can guarantee asymptotic normality of all the above estimators. Let $\delta$ stand for either $\pi_t$ or $1 - R = \pi_t/\pi_+$, and

Table 1: Optimal randomization rates.

| Known parameter | Estimated parameter $R$ | Estimated parameter $\pi_t$ |
|---|---|---|
| | Randomization rates $\gamma$ for RDT | |
| $p_1$ | $\frac{\sqrt{(1-\zeta_{1p})\zeta_{1t}}}{\sqrt{(1-\zeta_{1p})\zeta_{1t}}+\sqrt{(1-\zeta_{1t})\zeta_{1p}}}$ | $\frac{\sqrt{\zeta_{1p}(1-\zeta_{1p})}}{\sqrt{\zeta_{1p}(1-\zeta_{1p})}+\sqrt{\zeta_{1t}(1-\zeta_{1t})}}$ |
| $p_2$ | $\frac{\sqrt{(1-\zeta_{1p})\zeta_{1t}}}{\sqrt{(1-\zeta_{1p})\zeta_{1t}}+\sqrt{(1-\zeta_{1t})\zeta_{1p}}}$ | $\frac{\sqrt{B^2\zeta_{1p}(1-\zeta_{1p})}}{\sqrt{B^2\zeta_{1p}(1-\zeta_{1p})}+\sqrt{C^2\zeta_{1t}(1-\zeta_{1t})}}$ |
| $q=\frac{p_2}{1-p_1}$ | $\frac{\sqrt{(1-\zeta_{1p})\zeta_{1t}}}{\sqrt{(1-\zeta_{1p})\zeta_{1t}}+\sqrt{(1-\zeta_{1t})\zeta_{1p}}}$ | $\frac{\sqrt{\left(1-\zeta_{1p}\right)\zeta_{1p}\left(1-\zeta_{1t}\left(1-q\right)\right)^2}}{\sqrt{\left(1-\zeta_{1t}\right)\zeta_{1t}\left(1-\zeta_{1p}\left(1-q\right)\right)^2}+\sqrt{\left(1-\zeta_{1p}\right)\zeta_{1p}\left(1-\zeta_{1t}\left(1-q\right)\right)^2}}$ |
| | Randomization rates $\kappa$ for RCT | |
| | $\frac{\pi_+\sqrt{\pi_p(1-\pi_p)}}{\pi_+\sqrt{\pi_p(1-\pi_p)}+\pi_p\sqrt{\pi_+(1-\pi_+)}}$ | $\frac{\sqrt{\pi_p(1-\pi_p)}}{\sqrt{\pi_p(1-\pi_p)}+\sqrt{\pi_+(1-\pi_+)}}$ |

Table 2: Best variances for RCT and RDT.

| Known parameter | Estimated parameter $R$ | Estimated parameter $\pi_t$ |
|---|---|---|
| | RDT | |
| $p_1$ | $\frac{\sqrt{(1-\zeta_{1p})\zeta_{1t}}}{\sqrt{(1-\zeta_{1p})\zeta_{1t}}+\sqrt{(1-\zeta_{1t})\zeta_{1p}}}$ | $\frac{\zeta_+\left[(\zeta_{1t}-\zeta_{1p})^2(1-\zeta_+)+\left(\sqrt{\zeta_{1p}(1-\zeta_{1p})}+\sqrt{\zeta_{1t}(1-\zeta_{1t})}\right)^2\right]}{N(1-p_1)}$ |
| $p_2$ | $\frac{\sqrt{(1-\zeta_{1p})\zeta_{1t}}}{\sqrt{(1-\zeta_{1p})\zeta_{1t}}+\sqrt{(1-\zeta_{1t})\zeta_{1p}}}$ | $\frac{\left[A^2\zeta_+^2(1-\zeta_+)+\left(\sqrt{B^2\zeta_{1p}(1-\zeta_{1p})}+\sqrt{C^2\zeta_{1t}(1-\zeta_{1t})}\right)^2\right]}{N\zeta_+}$ |
| $q$ | $\frac{1}{\zeta_+ N}\frac{\zeta_{1p}^2}{\zeta_{1t}^2}\left(\sqrt{\frac{1-\zeta_{1p}}{\zeta_{1p}}}+\sqrt{\frac{1-\zeta_{1t}}{\zeta_{1t}}}\right)^2$ | $\frac{q^2\left(\sqrt{\left(1-\zeta_{1t}\right)\zeta_{1t}\left(1-\zeta_{1p}\left(1-q\right)\right)^2}+\sqrt{\left(1-\zeta_{1p}\right)\zeta_{1p}\left(1-\zeta_{1t}\left(1-q\right)\right)^2}\right)^2}{\zeta_+ N(1-\zeta_{1t}+\zeta_{1t}q)^4}$ |
| | RCT | |
| | $\frac{\left(\sqrt{R(1-\pi_p)}+\sqrt{R(1-\pi_+)}\right)^2}{N\pi_+}$ | $\frac{\left(\sqrt{\pi_p(1-\pi_p)}+\sqrt{\pi_+(1-\pi_+)}\right)^2}{N\pi_+}$ |

suppose we are interested in testing the hypotheses, $\mathrm{H}_0:\delta=0;\quad \mathrm{H}_\mathrm{a}:\delta\geq\delta^*$, with the probabilities of Type I error and Type II error equal to $\alpha$ and $\beta$ respectively. Using the "normal" approximation, we have

$$z_{1-\alpha}+z_{1-\beta}=\delta^*/\sqrt{\mathrm{Var}[\hat{\delta}]}, \tag{15}$$

where $z_{1-\upsilon}$ is the $1-\upsilon$ quantile of the normal distribution $\mathcal{N}(0,1)$. Combining Table 2 and (15) we can calculate the required sample sizes for various scenarios.

### *6.1. Illustrative sample size calculations for different scenarios*

As an illustration, we consider two scenarios for testing $\pi_t$: $\pi_t=\delta\geq 0.2$ and $\pi_t=\delta\geq 0.1$, and two scenarios for testing $R$: $1-R=\pi_t/\pi_+=\delta\geq 0.5$ and $1-R=\pi_t/\pi_+=\delta\geq 0.2$. In both cases the variances of estimated parameters are calculated under the assumption that $\pi_t=0.2,\ \pi_p=0.2$ for the first scenario, and $\pi_t=0.1,\ \pi_p=0.4$ for the second one. We use two popular Type I/II error rates, 0.05/0.2 and 0.05/0.1, and three placebo assignment proportions for RDT ($\gamma$ =0.7, 0.5 and 0.3) in the sample size computations. For RDT with misclassifications, we consider values for false positive and false negative rates to be $p_1=p_2=0.10$. Table 3 lists the required sample sizes of RDT and RCT for various combinations of these design parameters.

As previously pointed out, when misclassifications are present in the open-label stage, RDT can still be a more efficient than RCT. Overall for the considered scenarios, RDT is more efficient than RCT in testing $R$; the required sample size can be reduced by more than 50%. The knowledge of both $p_1$ and $p_2$, or at least one of them can greatly increase the efficiency of RDT in testing and

Table 3: The total sample sizes ($N$) required for RCT and RDT for various combinations of design parameters.

| Design Parameters | | | | Sample Size | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | RCT | | RDT with known parameter shown in () | | | |
| | $\alpha$ | $\beta$ | $\gamma$ | $\pi_t$ | $R$ | $\pi_t$ ($p_1$) | $\pi_t$ ($p_2$) | $\pi_t$ ($q$) | $R$ |
| $\pi_p$=0.2 | 0.05 | 0.2 | 0.7 | 124 | 69 | 63 | 77 | 250 | 37 |
| $\pi_t$=0.2 | | | 0.5 | - | - | 61 | 77 | 172 | 45 |
| | | | 0.3 | - | - | 80 | 98 | 161 | 69 |
| | 0.05 | 0.1 | 0.7 | 172 | 95 | 87 | 107 | 346 | 51 |
| | | | 0.5 | - | - | 85 | 106 | 239 | 62 |
| | | | 0.3 | - | - | 110 | 135 | 222 | 96 |
| $\pi_p$=0.4 | 0.05 | 0.2 | 0.7 | 606 | 495 | 208 | 227 | 481 | 184 |
| $\pi_t$=0.1 | | | 0.5 | - | - | 206 | 227 | 377 | 198 |
| | | | 0.3 | - | - | 281 | 310 | 416 | 288 |
| | 0.05 | 0.1 | 0.7 | 840 | 686 | 288 | 314 | 667 | 254 |
| | | | 0.5 | - | - | 286 | 314 | 522 | 275 |
| | | | 0.3 | - | - | 389 | 686 | 576 | 399 |

estimating $\pi_t$. When the information about $p_1$ and $p_2$ is available only through the knowledge of $q$, the information "spent" for estimating misclassifications could lead to a significant drop in efficiency of RDT often making it inferior to RCT, see the corresponding column in Table 3. Let us emphasize that if both $p_1$ and $p_2$ are unknown one cannot test or estimate $\pi_t$ and has to resort to testing or estimating $\pi_t^- = \pi_t(1-p_1)$, see also the concluding part of Section 4. More can be found in [7, Section 6].

## 7. Simulations

We have compared RDT with RCT using the asymptotic variances for the parameter estimates. To confirm the quality of the approximation (i.e. to verify that a given sample size is sufficiently large) one can perform simulations similar to the following examples.

We consider the following design conditions as an illustration: $\pi_t$=0.1, $\pi_p$=0.4, $p_1$=0.10, $p_2$=0.10, and $\gamma$=0.5, and two different sample sizes of 50 and 200. For each sample size, 2000 Monte-Carlo simulations were run, and for each simulated "trial", the MLE for $\pi_t$ and $R$ were calculated using (12), (14), and (9). In the upper and lower blocks of Figure 7, respectively, for sample size 50 and 200, we generate histograms for those 2000 calculated MLEs. For comparison, the dashed lines in the plots give the density of the asymptotic distributions of the maximum likelihood estimators.

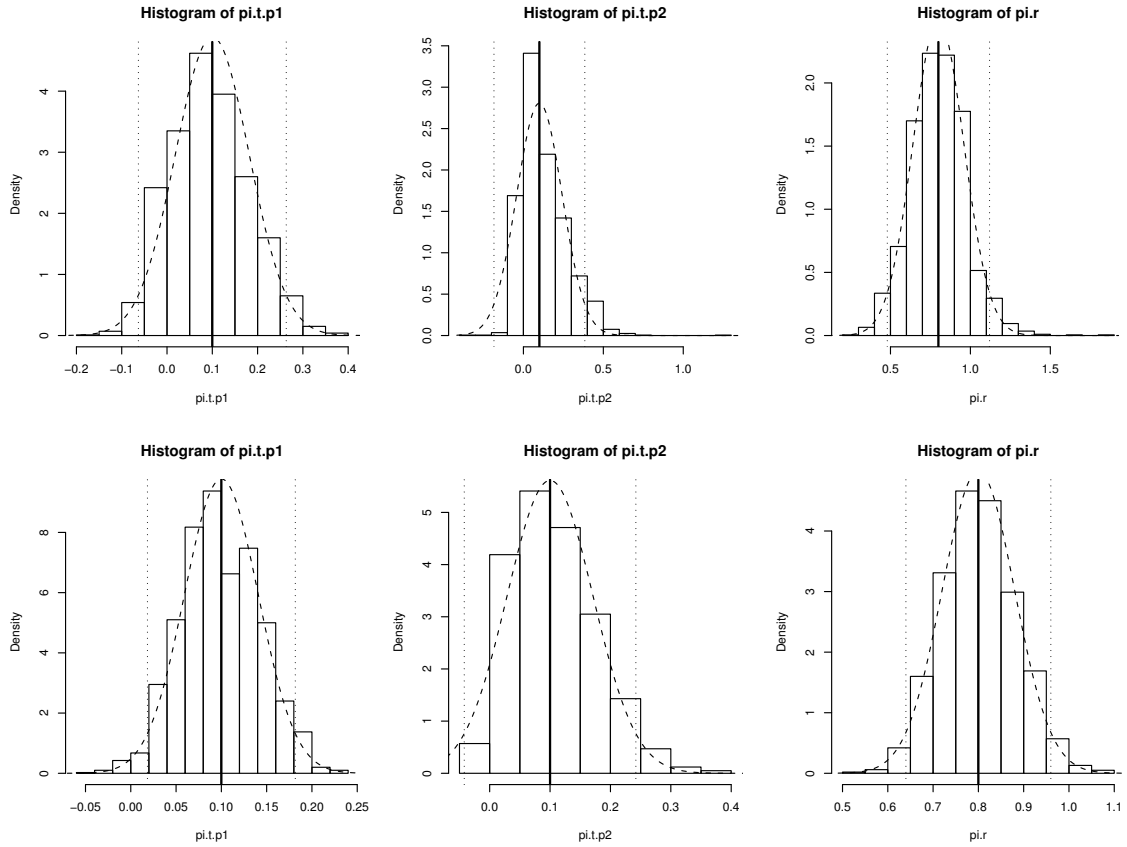
When the sample size is small, the estimator of $\pi_t$ given $p_2$ is slightly skewed to the right, suggesting their asymptotic variance should be used with some caution. The distribution of the estimators of $\pi_t$ (given $p_1$) and $R$ are well approximated by their asymptotic distributions even when the sample size is as few as 50.

One can certainly explore the finite sample properties under other design conditions than those we have shown here. The simulation program is available upon request.

## 8. Conclusions

A randomized discontinuation trial can alleviate an ethical concern by reducing the duration and degree of subjects' exposure to inactive placebo, and are often considered for evaluating therapies which require sustained administration. Statistical analysis of RDT is commonly carried out only on the outcomes of the enriched population of the second stage, while the information contained in the open-label stage is mostly wasted. Under certain circumstances when it is plausible to place additional assumptions on the population or the design, it is more sensible to incorporate the open-label stage

Figure 6: Parameter estimates based on 2,000 simulation repetitions. Upper: sample size=50; Lower: sample size=200 The first column: $\hat{\pi}_t$ given $p_1$; the second column: $\hat{\pi}_t$ given $p_2$; and the third column $\hat{R}$. Solid vertical line: true parameter value. Dotted lines: true parameter value $\pm$ asymptotic standard error for the estimate. Dashed line: asymptotic distribution.




outcomes to the analysis and maximize the total information we can obtain. In this paper we assume that the population can be partitioned into three groups of people: placebo responders, treatment-only responders, and non-responders. Placebo responders also respond to an active treatment. Treatment-only responders respond to the treatment but not to placebo. We propose to explore the efficiency of the RDT by examining the Fisher information matrices (or variances-covariances matrices) of the estimated parameters which takes into account the information from both the open-label and second stages.

We allow that the outcomes of the first (open-label) stage can potentially be misclassified. It reflects the fact that at the open-label stage some surrogate endpoints are usually used to detect responders and usually surrogate endpoints lead to less accurate measurements than primary endpoints. While the model is still simple (see Figure 1 for the model scheme), we find that the presence of misclassification makes it impossible to estimate the treatment effect ($\pi_t$) without additional assumptions. We propose to add some constraint to solve the non-identifiability issue and show that the information loss for estimating the misclassification rates can jeopardize the superiority of RDT for estimating $\pi_t$ under certain conditions.

We would like to emphasize that unlike $\pi_t$, the likelihood estimator of the probability ratio $R$ still exists even in the presence of misclassifications. Our result shows that under mild misclassifications,

letting $\gamma = 0.5$ secures superiority of RDT over RCT for a broad range of scenarios.

To have a comprehensive view of the RDT and RCT efficiencies, we compare the corresponding optimal allocations (i.e. minimizing the estimator variances). With the derived asymptotic variances for $\hat{\pi}_t$ and $\hat{R}$ and relying on central limit theorem, we also discuss hypothesis testing and sample size calculation.

An important issue when applying our results is the understanding of the meaning of individual response. There is always a debate on whether we should interpret the experiment results as "$100\pi_t\%$ of the population will respond to the treatment" or as "a subject of the population will response $100\pi_t\%$ of the time" [19, 20]. The evaluation of RDT in this paper relies on the first interpretation. In other words, we assume that the treatment (active treatment or placebo) effect is consistent over time (i.e., responders always respond while non-responders never respond), so that we can have a valid partition of population according to their responses.

When designing RDT, we assumed that there is no time trend of treatment, as we did in this paper. However, this may not be true in many applications. For example, when testing cytostatic agents, the size of tumor which changes over time can affect the effect of agents [17]. One way to deal with this issue is to incorporate a time-dependent response rate into the model. In order to have valid maximum likelihood estimator of unknown parameters, a practitioner will need more experiments to estimate a larger number of parameters. Another potential extension may be the introduction of dependence of misclassification rates $p_1(t)$ and $p_2(t)$ on time. Indeed, relying (similar to [23]) on mechanistic models of the disease progression one can derive functions that describe $p_1(t)$ and $p_2(t)$. Usually they are diminishing function of $t$. As soon as both of those functions are defined the design problem on selection of optimal duration ($t_1$ and $t_2$, $t_1 + t_2 \leq T$ ) of the open-label stage and the final stage can be considered.

Often, the outcomes can have more than two categories or can be continuous. For example, in the paper by Rosner et al. [17], the population is described by three categories: "high response", "low response" and "no response". The extension of our results to this scenario could be challenging and is definitely deserving further investigations.

In this paper, we defined the "efficiency" of either RCT or RDT as the estimation precision of the selected endpoints given a total number of subjects participating in a clinical trial. In general, the efficiency should be assessed using more general utility functions in which the estimator's variance is not the only contributor. For instance, RDT can be more expensive then RCT in terms of logistics, trial completion time, expenses associated with interim data analysis,it is based on more restrictive assumptions and more complicated statistical methods, etc.. Thus the analysis of the statistical efficiency is just the first step in the selection of the best design for a specific clinical trial.

**Acknowledgment**

We would like to thank Prof. Stephen J. Senn and Prof. Gary L. Rosner for their valuable comments on the earlier versions of this paper. Comments from the two anonymous reviewers were essential for the improvement the final version.

[1] Amery, W. and Dony, J. (1975). Clinical trial design avoiding undue placebo treatment. *Journal of Clinial Pharmacology*, October, 674–679.

[2] Burzykowski, T., Molenberghs, G. and Buyse, M. (2005). The evaluation of surrogate endpoints. New York: Springer.

[3] Capra, W. B. (2004). Comparing the power of the discontinuation design to that of the classic randomized design on time-to-event endpoints. *Controlled Clinical Trials* **25** 168–177.

[4] Chiron, C., Dulac, O. and Gram, L. (1996). Vigabatrin withdrawal randomized study in children. *Epilepsy Research* **25**, 209–215.

[5] Fedorov, V. V. and Atkinson, A. C. (1988) The optimum design of experiments in the presence of uncontrolled variability and prior information in "Optimal Design and Analysis of Experiments" Dodge, Y., Fedorov, V. V. and Wynn, H.P. 327-344. Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam] (New York; Amsterdam).

[6] Fedorov, V. V. and Hackl, P. (1997) Model-oriented design of experiments. Springer-Verlag Inc (Berlin; New York) 0-387-98215-9.

[7] Fedorov, V. V. and Liu, T. (2005) Randomized discontinuation trials: design and efficiency. GlaxoSmithKline Biomedical Data Science Technical Report, 2005-3.

[8] Fedorov, V. V. and Liu, T. (2007) Enrichment Design in "Wiley Encyclopedia of Clinical Trials." 1-8. John Wiley & Sons, Inc..

[9] Fedorov, V. V., Mannino F. and Zhang, R. (2008) Consequences of dichotomization. *Pharmaceutical Statistics* **8**, 50-61.

[10] Freidlin, B. and Simon, R. (2005) Evaluation of randomized discontinuation design. *Journal of Clinical Oncology* **23**(22), 5094-5098.

[11] Hallstrom, A. P. and Friedman, L. (1991). Randomizing responders. *Controlled Clinical Trials* **12**, 486–503.

[12] Kopec, J., Abrahamowicz, M. and Esdaile, J. (1993). Randomized discontinuation trials: Utility and efficiency. *Journal of Cinical Epidemiology* **46**, 959–971.

[13] Korn, E. L., Arbuck, S. G., Pulda, J. M., Simon, R., Kaplan, R. S. and Christian, M. C. (2001). Clinical trial designs for cytostatic agents: Are new approaches needed? *Journal of Clinical Oncology* **19**, 265–272.

[14] Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York.

[15] MacLehose, R. R., Reeves, B. C., Harvey, I. M., Sheldon, T. A., Russell, I.T. and Black, A. M. S. (2000). A systematic review of comparisons of effect sizes derived from randomized and non-randomized studies. *Health Technology Assessment* **4**(34), Chapter 7.

[16] Pablos-Méndez, A., Barr, R. G. and Shea, S. (1998). Run-in periods in randomized trials. *Journal of American Medical Association* **279**, 222–225.

[17] Rosner, G. L., Stadler, W. and Ratain, M. J. (2002). Randomized discontinuation design: Application to cytostatic antineoplastic agents. *Journal of Clinical Oncology* **20**, 4478–4484.

[18] Simon, R. and Maitournam, A. (2004) Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* **10**, 7659-6763.

[19] Senn, S. J. (2001). Individual therapy: New dawn or false dawn. *Drug Information Journal* **35**, 1479–1494.

[20] Senn, S. J. (2004). Individual response to treatment: Is it a valid assumption? *BMJ* **329**, 966–968.

[21] Shao, J. (1999). *Mathematical Statistics*. Springer, New York.

[22] Stadler, W. M., Rosner, G., Small, E., Hollis, D., Rini, B., Zaentz, S. D. and Mahoney, J. (2005). Successful implementation of the randomized discontinuation trial design: An application to the study of the putative antiangiogenic agent carboxyaminoimidazole in renal cell carcinoma - calgb 69901. *Journal of Clinical Oncology* **23**, 3726–3732.

[23] Trippa, L., Rosner G. L. (2012). Bayesian enrichment strategies for randomized discontinuation trials. *Biometrics* **68**, 203–225.