

Bayesian Response-Adaptive Designs for Basket Trials

Steffen Ventz,^{1,2,*} William T. Barry,^{1,3} Giovanni Parmigiani,^{1,4} and Lorenzo Trippa^{1,4}[Q1](#)

¹Dana-Farber Cancer Institute, Boston, Massachusetts

²University of Rhode Island, Kingston, Rhode Island

³Harvard Medical School, Boston, Massachusetts

⁴Harvard School of Public Health, Boston, Massachusetts

**email*: steffen@dfci.harvard.edu

SUMMARY. We develop a general class of response-adaptive Bayesian designs using hierarchical models, and provide open source software to implement them. Our work is motivated by recent master protocols in oncology, where several treatments are investigated simultaneously in one or multiple disease types, and treatment efficacy is expected to vary across biomarker-defined subpopulations. Adaptive trials such as I-SPY-2 (Barker et al., 2009) and BATTLE (Zhou et al., 2008) are special cases within our framework. We discuss the application of our adaptive scheme to two distinct research goals. The first is to identify a biomarker subpopulation for which a therapy shows evidence of treatment efficacy, and to exclude other subpopulations for which such evidence does not exist. This leads to a subpopulation-finding design. The second is to identify, within biomarker-defined subpopulations, a set of cancer types for which an experimental therapy is superior to the standard-of-care. This goal leads to a subpopulation-stratified design. Using simulations constructed to faithfully represent ongoing cancer sequencing projects, we quantify the potential gains of our proposed designs relative to conventional non-adaptive designs.

KEY WORDS: Adaptive randomization; Bayesian hierarchical models; Master protocols; Multi-arm clinical trials.

1. Introduction

We describe a general class of trial designs that are motivated by recent master protocols in oncology, where several treatments are investigated simultaneously for multiple diseases and efficacy is expected to vary across biomarker-defined subpopulations. Numerous advances in molecular biology have established that cancer is a heterogeneous disease, and that many malignancies share a subset of driver mutations in known oncogenes (Vogelstein et al., 2013). As the cost of sequencing assays decreases, genomic profiling has become standard practice in many cancer centers. For instance, every patient treated at the Dana-Farber Cancer Institute since 2011 has been offered the opportunity to participate in the PROFILE project (Lee et al., 2012) to determine the individual pattern of DNA alterations in a patient's cancer. The individual mutation data are then used to guide cancer therapies.

The development of anti-cancer treatments has focused increasingly on therapies that target genetic alterations common to multiple cancer types. Consequently, clinical investigations have shifted toward biomarker-driven studies that seek to match treatments to the subpopulations that benefit from them (Conley and Doroshow, 2014). This includes studies that are commonly referred to as umbrella and basket trials. Umbrella trials use a central infrastructure for molecular profiling to guide treatment assignment. These trials focus on a single malignancy and consist of multiple sub-studies for different biomarker subpopulations. Examples include BATTLE, Lung-MAP, and ALCHEMIST (Zhou et al. (2008), NCT02154490, NCT02194738). In contrast, basket

trials are designed to test therapies in multiple malignancies. Patients with different malignancies are assigned to treatments based on their biomarker profiles. NCI-MATCH (Conley and Doroshow, 2014) is an example, enrolling patients with multiple malignancies that range from brain cancer to melanoma. The fragmentation of the overall sample size into smaller subpopulations has stimulated the development of statistical designs that maintain adequate power at the subpopulation level (Zhou et al., 2008; Barry et al., 2015). This has also motivated the use of adaptive procedures in study design. Here, we focus on Bayesian designs, as seen in BATTLE and I-SPY-2 (Zhou et al., 2008; Barker et al., 2009), which adjust the randomization probabilities within subpopulations adaptively in favor of the most promising arms. Bayesian methods can model treatment effects across different cancer types and across biomarker-defined categories (Barry et al., 2015) with hierarchical priors.

The present work is motivated by the PROFILE project. We implement designs that learn using data from multiple malignancies, and assign patients adaptively to treatments based on their molecular profile. We generalize and build upon previous work on hierarchical models (Thall et al., 2003; Wathen et al., 2008; Lee et al., 2010) and response-adaptive randomization (Thall and Wathen, 2007; Trippa et al., 2012). We discuss a general class of designs for multi-arm trials with biomarker-defined subgroups and multiple malignancies. Patients are classified accordingly to biomarker measurements into predefined groups that remain fixed during the study. The unifying element of this class is the use of a simple

Bayesian model to borrow information across subpopulations and cancer types, coupled with the use of response-adaptive randomization.

We discuss the application of our model in two distinct settings: (i) *The subpopulation-stratified design*, in which multiple treatments are evaluated within a targeted biomarker-positive subpopulation. The aim is to identify treatments with positive effects, accounting for the possibility that treatment effects might be limited to a subset of malignancies. For instance, the BRAF inhibitor Vemurafenib has been shown to be effective in melanoma, high-grade gliomas and lung cancer, but not in colorectal cancer (Robinson et al., 2014).

(ii) *The subpopulation-finding design* evaluates experimental therapies in multiple biomarker-defined subgroups. One example would consist of a trial enrolling patients with abnormalities in the oncogenes PIK3CA, PIK3RI, PTEN, and mTOR, which activate the PI3K/Akt/mTOR pathway (Fruman and Rommel, 2014). Patients are treated with experimental therapies, and the primary goal of the study is to match treatments to biomarker-defined subgroups that are shown to be sensitive to the targeted therapies.

Biomarker designs for single cancer types have discussed in the literature. For example enrichment designs in (Wang et al., 2007, 2009; Brannath et al., 2009; Freidlin et al., 2010) test efficacy in the biomarker positive and in the overall population controlling the family wise type I error rate. Recently An et al. (2015) discuss alternative randomized designs in this context, and Mehta and Gao (2011), Mehta et al. (2014) studied group-sequential methods in biomarker studies and focused on the control of the family-wise type I error rates. Multi-arm biomarker-stratified designs have been proposed in (Zhou et al., 2008; Barker et al., 2009; Lee et al., 2010; Barry et al., 2015).

We combine adaptive randomization with an iterative procedure for tuning sequential stopping boundaries that maintain a pre-specified type I error level. This approach has connections with previous work for controlling type I error rates (Rosenberger and Hu, 1999). Our numerical illustrations show the potential gains over non-adaptive methods. An open-source R package is available at <http://bcb.dfc.harvard.edu/steffen/software.html>.

2. Bayesian Model

We introduce a model to describe response probabilities in multiple cancer types and biomarker subgroups. We use the probability model for response-adaptive randomization in Section 3.

2.1. Notation

We consider a clinical trial with experimental arms $a = 1, \dots, a^*$, evaluated in $d = 1, \dots, d^*$ cancer types, and $m = 1, \dots, m^*$ biomarker-defined subgroups. The biomarker subgroups are specified before the beginning of the study, and their definition remains fixed during the study. The index $a = 0$ will denote the control arm.

The random vector $(T_i, D_i, M_i, A_i, R_i)$ refers to the i -th patient. T_i is the enrollment time, $0 \leq T_i \leq T_{i+1}$, while D_i, A_i and M_i indicate the cancer type, treatment assignment, and biomarker subpopulation, respectively. The primary

outcome $R_i \in \{0, 1\}$ is binary, such as radiologic response after L weeks of treatment. In our simulation study, we will assume L identical across cancer types and subgroups. We use $N_{d,m,a}^i(i) = \sum_{n < i} I(D_n = d, M_n = m, A_n = a)$ to denote the number of patients with disease d in subpopulation m , which are assigned to arm a before T_i . Similarly, $N_{d,m,a}(i) = \sum_{n < i} I(D_n = d, M_n = m, A_n = a, T_n + L \leq T_i)$ is the number of patients with known outcome by time T_i , and Σ_i indicates the information available at the enrollment of the i th patient, which formally coincides with the σ -algebra generated by the random variables that are observable by time T_i . Lastly, $I_{d,m,a}(i) \in \{0, 1\}$, a function of Σ_i , defines whether or not patients with disease d in subpopulation m are randomized with positive probability to treatment a at time T_i .

2.2. Probability Model

We use a Bayesian prior for the response probabilities

$$\mathbb{P}(R_i = 1 | D_i = d, M_i = m, A_i = a) = p_{d,m,a} = g(\theta_{d,m,a}). \quad (1)$$

Here, $g(\cdot)$ denotes a link function mapping the real line into $[0, 1]$. The parameter $p_{d,m,0} = g(\theta_{d,m,0})$ is the probability of response under the control arm for that disease-marker combination. We use a multivariate normal prior for $\theta = \{\theta_{d,m,a}\}$.

To facilitate elicitation of the prior, we decompose $\theta_{d,m,0}$ into $\theta_{d,m,0} = \eta_d + \eta_{d,m}$, with independent normal components

$$\eta_d \sim N(0, \sigma_{\eta_d}^2) \quad \text{and} \quad \eta_{d,m} \sim N(\mu_{\eta_{d,m}}, \sigma_{\eta_{d,m}}^2). \quad (2)$$

The prior mean for $\theta_{d,m,0}$ equals $\mu_{d,m}$, and the correlation between $\theta_{d,m,0}$ and $\theta_{d,m',0}$ for markers $m \neq m'$ equals $\sigma_{\eta_d}^2 / \sqrt{(\sigma_{\eta_d}^2 + \sigma_{\eta_{d,m}}^2)(\sigma_{\eta_d}^2 + \sigma_{\eta_{d,m'}}^2)}$. With $\sigma_{\eta_{d,m}}^2 = \sigma_{\eta_{d,m'}}^2 = 0$ the correlations between $\theta_{d,m,0}$ and $\theta_{d,m',0}$ is one, while on the opposite extreme $\sigma_{\eta_d}^2 = 0$ makes these random variables independent.

If the control treatment for a given cancer type is identical across subpopulations and the biomarkers that define the subpopulations are not prognostic, then it is convenient to use identical prior means $\mu_{d,m}$ across subpopulations m . The data can subsequently inform the model of different response probabilities $p_{d,m,0}$ across markers m . On the other hand, if standard-of-care varies or the biomarkers are known to be prognostic, then it is convenient to set $\sigma_{\eta_d}^2 = 0$, in which case the probabilities $p_{d,m,0}, m \geq 1$ are independent.

To complete the model we add the treatment effects

$$\theta_{d,m,a} = \theta_{d,m,0} + \zeta_{d,m,a}, \quad a = 1, \dots, a^*,$$

and facilitate elicitation of the prior using the decomposition

$$\zeta_{d,m,a} = \beta_a + \beta_{m,a} + \beta_{d,a} + \beta_{d,m,a},$$

which are independent normal random variables

$$\begin{aligned} \beta_a &\sim N(0, \sigma_{\beta_a}^2), \quad \beta_{m,a} \sim N(0, \sigma_{\beta_{m,a}}^2), \\ \beta_{d,a} &\sim N(0, \sigma_{\beta_{d,a}}^2) \quad \text{and} \quad \beta_{d,m,a} \sim N(0, \sigma_{\beta_{d,m,a}}^2). \end{aligned} \quad (3)$$

Here, β_a can be interpreted as the mean treatment effect across (d, m) combinations. The random variables $\beta_{m,a}$ and $\beta_{d,a}$ represents marker-specific and disease-specific departures from β_a . We discuss in Subsections 2.3, 2.4, and the Web-based Supplementary Material the selection of the prior parameters in (3) and (2). For the subpopulation-finding and subpopulation-stratified designs, this will involve setting some of the parameters to zero. The parameters β are non-identifiable and only used to specify the normal prior for θ and to tune the degree of dependence across the treatment effects: $\text{Cov}(\zeta_{d,m,a}, \zeta_{d',m,a}) = \sigma_{\beta_a}^2 + \sigma_{\beta_{m,a}}^2$ if $d \neq d'$, and $\text{Cov}(\zeta_{d,m,a}, \zeta_{d,m',a}) = \sigma_{\beta_a}^2 + \sigma_{\beta_{d,a}}^2$ for $m \neq m'$.

The relative simplicity of the model allows us to graphically represent and tune key aspects of the prior in order to tailor a design to its specific biological and clinical context. For most settings, previous meta-analytic studies can guide elicitation of $\theta_{d,m,0}$. Then, for representative values of $\theta_{d,m,0}$ and $\theta_{d',m,a}$, we can plot the conditional distribution of the treatment effects $\zeta_{d',m,a}$ and the conditional distributions of $p_{d',m,a}$ for any $d' \neq d$ (see the examples in the Web-based Supplementary Material). The graphs illustrate the extent to which the adaptive algorithm learns and borrows information from patients with multiple diseases that share biomarker characteristics. If there is a treatment effect in one cancer type, it is often reasonable to hypothesize positive effects across malignancies. Nonetheless, the oncology literature points to both positive ($\zeta_{d,m,a}, \zeta_{d',m,a} > 0$) as well as negative ($\zeta_{d,m,a} > 0, \zeta_{d',m,a} \leq 0$) examples (Robinson et al., 2014). These examples motivate a careful elicitation of the prior. For most applications it is reasonable to simplify the selection of the σ^2 parameters to a choice of six values ($\sigma_1^2, \dots, \sigma_6^2$), $\sigma_{\eta_d}^2 = \sigma_1^2$, $\sigma_{\eta_{d,m}}^2 = \sigma_2^2$, $\sigma_{\beta_a}^2 = \sigma_3^2$, $\sigma_{\beta_{m,a}}^2 = \sigma_4^2$, $\sigma_{\beta_{d,a}}^2 = \sigma_5^2$, $\sigma_{\beta_{d,m,a}}^2 = \sigma_6^2$. We follow this approach in our simulations in Section 5. The Web-based Supplementary Material shows how to tune the prior parameters using simulations under several scenarios.

2.3. Subpopulation-Finding Design

For many phase II trials the primary goal is to identify subgroups with a positive treatment effect. The design randomizes patients with multiple cancer types $d = 1, \dots, d^*$ from multiple subgroups $m = 1, \dots, m^*$ to experimental treatments $a = 1, \dots, a^*$ and the aim is to find the subgroups that benefit from the respective therapies. For each arm $a = 1, \dots, a^*$ and subpopulation, $m = 1, \dots, m^*$, the goal is to test the null hypothesis

$$H_{m,a} = \{p_{d,m,a} \leq p_{d,m,0} \text{ for all malignancies } d = 1, \dots, d^*\}.$$

Studies in early drug development are frequently conducted without a control arm, and response rates $p_{d,m,a}$ are compared to estimates $\hat{p}_{d,m,0}$ from historical data. In this setting, we center the prior of $\theta_{d,m,0}$ at $\mu_{\eta_{d,m}} = g^{-1}(\hat{p}_{d,m,0})$, set $\sigma_{\eta_d}^2 = 0$, and select $\sigma_{\eta_{d,m}}^2$ to reflect the uncertainty level on the estimates $\hat{p}_{d,m,0}$ (Hobbs et al., 2012). Alternatively we can specify $\mu_{\eta_{d,m}} = \sigma_{\eta_d}^2 = \sigma_{\eta_{d,m}}^2 = 0$, so that $p_{d,m,a} = g(\zeta_{d,m,a})$ for $a > 0$. We can then tune a multivariate normal prior for ζ using the decomposition (3). For $d^* = 1$, the probability model is similar to the one used in BATTLE (Zhou et al., 2008). We will focus on subpopulation-finding designs without a control arm.

2.4. Subpopulation-Stratified Design

It is common to evaluate anti-cancer therapies with known targets using designs that enroll only patients from a biomarker-positive subgroup. Using our notation, the subpopulation-stratified design compares drugs $a = 1, \dots, a^*$ in cancers $d = 1, \dots, d^*$ to disease-specific control therapies $a = 0$. The goal is to find cancer types with positive effects under experimental therapies a in the subpopulation m . For instance, several drugs that target the EGFR pathway have been investigated in biomarker-defined subpopulations. In these cases, one can hypothesize positive treatment effects of the experimental arm a for patients across several cancer types d within the same biomarker group. We test the null hypotheses $H_{d,m,a} = \{p_{d,m,a} \leq p_{d,m,0}\}$ based on targeted type I error rates. In subpopulation-stratified designs, we borrow information on treatment effects across cancer types by setting $\sigma_{\beta_a}^2 = \sigma_{\beta_{d,a}}^2 = 0$ and tuning $\sigma_{\beta_{m,a}}^2, \sigma_{\beta_{d,m,a}}^2 > 0$. We focus on subpopulation-stratified designs with a control arm. The design can be easily modified for studies without a control arm.

3. Response-Adaptive Randomization

We define the probability of randomizing the i -th patient with disease-marker combination (d, m) to treatment a as

$$\mathbb{P}[A_i = a | D_i = d, M_i = m, \Sigma_i] \propto S_{d,m,a}(i) I_{d,m,a}(i), \quad (4)$$

where $S_{d,m,a}(i)$ is a non-negative function of the available data, and $I_{d,m,a}(i) \in \{0, 1\}$ is one if patients in the disease-marker combination (d, m) are eligible to receive treatment a at the enrollment of patient i and zero otherwise.

Several response-adaptive randomization strategies have been proposed in the literature. One of the earliest approaches that was introduced by Thompson (1933) specifies $S_{d,m,a}(i) = \mathbb{P}[p_{d,m,a} > p_{d,m,0} | \Sigma_i]$. More recent trials have used $S_{d,m,a}(i) = \mathbb{E}[p_{d,m,a} | \Sigma_i]$ (Zhou et al., 2008). Extensions of these methods have been proposed to control the exploration of the experimental arms during the early phase of the trial followed by the exploitation of acquired information to assign more patients to the most promising arms. Examples include

$$S_{d,m,a}(i) = \mathbb{P}[\zeta_{d,m,a} > 0 | \Sigma_i]^{h(i,m,d)}, \quad (5)$$

as suggested in Thall and Wathen (2007), where h is a monotone function of the number of observed outcomes for cancer d in subpopulation m , $N_{d,m}(i) = \sum_a N_{d,m,a}(i)$. Zhou et al. (2008) used $h(i,m,d) = \prod_m I(N_{d,m}(i) \geq N_{\min})$ to guarantee a minimum number of observed outcomes for each (d, m) combination before departing from balanced randomization. Thall and Wathen (2007) suggested $h(i,m,d) = 0.5 N_{d,m}(i) / N_{d,m}$, where $N_{d,m}$ is the expected number of patients in the subgroup at completion of the trial. Here, we will use $h(i,d,m) = \gamma_1 \times (N_{d,m}(i) / N_{d,m})^{\gamma_2}$. By tuning $\gamma_1 > 1$ and $\gamma_2 = -\log(\gamma_1) / \log u$, for some $u \in (0, 1)$, we specify h such that $h(i,d,m) = 1$ after a proportion u of the total expected number of outcomes is observed and $h(i,d,m) \approx \gamma_1$ at the end of the trial.

For the subpopulation-stratified design, randomization probabilities for the control arm are defined to approximately

1 match the number of patients assigned the control and the
 2 experimental arm with the highest number of patients. This
 3 sustains power and is achieved by defining

$$4 \quad S_{d,m,0}(i) = \exp \left\{ k \left[\max_{a>0} N'_{d,m,a}(i) - N'_{d,m,0}(i) \right] \right\}, \quad (6)$$

7 where k is a positive constant.

8 For a subpopulation-finding design, without a control arm,
 9 we define $S_{d,m,a}$ using the posterior probability of the event
 10 $P_{d,m,a} = \max_{a'} P_{d,m,a'}$,

$$12 \quad S_{d,m,a}(i) = \mathbb{P} [\cap_{a'} \{ \zeta_{d,m,a} \geq \zeta_{d,m,a'} \} | \Sigma_i]^{h(i,d,m)}. \quad (7)$$

14 It is important to design response-adaptive randomization
 15 such that it is not oversensitive to optimistic treatment effects
 16 estimates early during the trial. We multiply the right hand
 17 of equations [Q2](#) (5), (6), and (7) by a positive and increasing
 18 function of $(N_{d,m}^* - N'_{d,m,a}(i))_+$, where $x_+ = xI(x > 0)$, to
 19 guarantee a minimum of $N_{d,m}^*$ enrolled patients during the initial
 20 stage of the trial for each (d, m, a) combination that we
 21 investigate: $N_{d,m}^* > 0$ is a parameter of the adaptive design. In
 22 particular, we use the exponential transforms

$$25 \quad \exp \left\{ w \times (N_{d,m}^* - N'_{d,m,a}(i))_+ \right\}, \quad (8)$$

27 where $w > 0$. It has been shown that response-adaptive treat-
 28 ment allocation can be more variable compared to balanced
 29 designs (Thall et al., 2015). The factor (8) can be used to control
 30 the variability of treatment allocation for each (d, m, a)
 31 combination. In the extreme cases this correction yields either
 32 stratified balanced randomization or standard Bayesian adap-
 33 tive randomization.

34 Extensive simulation studies are necessary to evaluate
 35 adaptive designs. This makes the use of MCMC algorithms
 36 to compute $S_{d,m,a}(i)$ time consuming. We use a different
 37 approach to approximate posterior distributions. Under stan-
 38 dard regularity conditions (Pratt, 1981) our θ posterior is
 39 log-concave and hence unimodal. We compute the posterior
 40 mode $\hat{\theta}$ with a Newton-Raphson algorithm. Then we use a
 41 Bernstein-von-Mises approximation of the posterior with a
 42 normal distribution centered at the posterior mode and hav-
 43 ing covariance matrix equal to the inverse Hessian of the
 44 log-posterior at the mode. Computational details are out-
 45 lined in Web-based Supplementary Material together with a
 46 comparison to MCMC approximations.

48 4. Stopping Rules

49 As part of the response-adaptive design, we consider for each
 50 combination of disease, subpopulation, and experimental arm,
 51 early stopping due to sufficient evidence for efficacy or for
 52 futility. Stopping rules are applied sequentially before the
 53 assignment of each patient $i \geq 1$ using the data Σ_i available up
 54 to the i -th enrollment. Specifically, stopping rules are defined
 55 based on whether a test statistic for efficacy $V'_{d,m,a}(i)$ becomes
 56 larger than a pre-specified boundary $b'_{d,m,a}(i)$, or whether a
 57 test statistic for futility, $V''_{d,m,a}(i)$ becomes smaller than a pre-
 58 specified threshold $b''_{d,m,a}(i)$. When either event happens, we
 59 set the treatment availability indicator $I_{d,m,a}(j) = 0$ for all

$j \geq i$, and treatment a will not be assigned to the disease-
 marker combination (d, m) throughout the remainder of the
 study.

4.1. Stopping Rules for the Subpopulation-Stratified Design

Here, we describe test statistics and stopping boundaries for
 the subpopulation-stratified design. Stopping rules for the
 subpopulation-finding design are defined are chosen similarly,
 and details are given in the Web-based Supplementary Mate-
 rial. In the subpopulation-stratified design subpopulations
 are considered separately without borrowing of information
 across them, so we eliminate the index m from the notation
 for this subsection.

The set of null hypotheses is $\{H_{d,a}; d \in \mathcal{D}_a, a = 1, \dots, a^*\}$
 and treatment effects are evaluated across the diseases
 $\mathcal{D}_a = \{d : I_{d,a}(1) = 1\}$. We use standard efficacy statistics, with-
 out borrowing information across diseases,

$$V'_{d,a}(i) = \frac{\hat{p}_{d,a}(i) - \hat{p}_{d,0}(i)}{\sqrt{\widehat{\text{Var}}[\hat{p}_{d,a}(i)] + \widehat{\text{Var}}[\hat{p}_{d,0}(i)]}}, \quad (9)$$

where $\hat{p}_{d,a}$ is the empirical response rate for (d, a) and
 $\widehat{\text{Var}}[\hat{p}_{d,a}(i)] = \hat{p}_{d,a}(i)(1 - \hat{p}_{d,a}(i))/N_{d,a}(i)$. We use a decreasing
 boundary $b'_{d,a}(i)$, such that stronger evidence is required to
 stop for efficacy in the early stages of the study. We choose

$$b'_{d,a}(i) = \begin{cases} \lambda'_{d,a} \times (1 + s_1 \times s_2^{N_{d,a}(i) - N_{\min}}) & \text{if } N_{d,a}(i) \geq N_{\min}, \\ +\infty & \text{otherwise.} \end{cases} \quad (10)$$

Here, $s_1 \geq 0$ and $s_2 \in [0, 1]$ determine the shape of the bound-
 ary, while $N_{\min} \geq 0$ is a pre-selected minimum threshold for
 the number of observed outcomes $N_{d,a}(i)$, which is required
 before the arm can be recommended for a confirmatory study
 in disease d . The boundary decreases from $\lambda'_{d,a} \times (1 + s_1)$ to
 $\lambda'_{d,a}$. With $s_1 = 0$, $b'_{d,a}(\cdot)$ is identical to Pocock boundaries
 (Pocock, 1977). For large $s_2 \approx 0.95$, we can tune values of
 s_1 and λ' so that $b'_{d,a}(\cdot)$ has a shape similar to O'Brien-
 Fleming boundaries (O'Brien and Fleming, 1979). For the
 examples considered in Section 5 we found that values of
 (s_1, s_2) in $[2, 3.5] \times [0.85, 0.97]$ give efficacy boundaries that
 sacrifice minimal power ($\leq 4\%$) when compared to testing
 efficacy at the end of the trial without early stopping.

We also use the posterior probability of a positive treatment
 effect as futility statistics, $V''_{d,a}(i) = 1 - \mathbb{P}[H_{d,a} | \Sigma_i]$, borrowing
 information across diseases. We specify a monotone bound-
 ary $b''_{d,a}(i) = \lambda'' \times (1 - s_3^{N_{d,a}(i)})$ for futility, with $\lambda'', s_3 \in [0, 1]$
 to require again strong evidence to stop early during the trial.

When designing the adaptive trial, we first select
 s_1, s_2, s_3, N_{\min} , and λ'' using simulations of the study under
 a set of plausible scenarios. During the conduct of the trial,
 we then calibrate the parameters $\lambda' = \{\lambda'_{d,a}\}$ sequentially based
 on the accumulating data so that the type I error probabil-
 ities are controlled at a pre-specified α level. The Web-based
 Supplementary Material describes the algorithm that we use
 to calibrate λ' .

5. Examples and Simulation Studies

5.1. Subpopulation-Stratified Design: *PI3K* Inhibitor

Several *PI3K* inhibitors are currently under development in oncology. Preclinical and clinical studies suggest that this class of therapies might be effective for patients with *PI3K* abnormalities across multiple cancer types (Polivka and Janku, 2014). We consider a trial that restricts eligibility to patients with *PI3K* abnormalities and *breast, endometrium, colon/rectum, bladder, or ovarian cancer*. Based on data from our institute we assume accrual rates of (2.3, 1.3, 0.7, 0.4, 0.3) *PI3K* patients per week for the five cancer types. Response to treatment is measured $L = 8$ weeks after randomization. Analogous endpoints were used by Zhou et al. (2008). The trial compares $a^* = 3$ experimental arms to cancer-specific control regimens. For simplicity, we consider scenarios where the response probabilities under standard-of-care equal 0.3 for every cancer type.

Balanced randomization requires approximately 63 patients per arm to test $H_{d,a}$ in each cancer d with a 10% type I error and 85% power of detecting the treatment effect $p_{d,a} = 0.5$. To compare the operating characteristics of the adaptive design versus fixed randomization, we set the overall sample size per cancer type equal to $N_{d,m} = 240$.

We initially explore seven scenarios without early stopping, and evaluate the extent to which adaptive randomization increases the number of patients assigned to arms with positive effects, compared to balanced randomization. In all scenarios, we assume arms $a = 2, 3$ are ineffective with response rates equal to the control. Arm 1 has no effect for any cancer in scenario 1, a moderate (strong) effect for breast cancer in scenario 2 (scenario 5), a moderate (strong) effect for breast, endometrium, and ovarian cancer in scenario 3 (scenario 6), and a moderate (strong) effect for all five cancer types in scenarios 4 (scenario 7). Moderate and strong treatment effects correspond to response probabilities of 0.4 and 0.5, respectively.

We use a probit link function $g(\cdot)$ in the Bayesian model (1), and compare an adaptive design with strong borrowing of information across cancer types to an adaptive design without borrowing of information. Strong borrowing is achieved by setting the prior correlation $\text{Cor}(\zeta_{d,m,a}, \zeta_{d',m,a}) \approx 0.9$ for $d \neq d'$. No borrowing of information is specified with $\sigma_{\beta_{m,a}}^2 = 0$ so that $\text{Corr}(\zeta_{d,m,a}, \zeta_{d',a,m}) = 0$. We use the acronyms MAB and MAN to distinguish multi-arm adaptation with and without borrowing of information.

Table 1 shows the mean number of patients for each cancer that were randomized to arms $a = 0, \dots, 3$ across 5000 simulated trials under each of the seven scenarios. If all arms are ineffective (scenario 1), MAB and MAN assigned on average 54.4 patient with cancer $d = 1, \dots, 5$ to each experimental arm and 78 patients to the control. In scenarios 2 and 5, where arm 1 has treatment effect only for breast cancer, MAB borrowed information from the remaining cancer types for which the experimental treatment is ineffective. Thus, MAB randomized on average fewer breast cancer patients to arm 1 than MAN (67.5 vs. 69.6 and 75.6 vs. 76.8 in scenario 2 and 5). At the opposite extreme in scenarios 4 and 7, when arm 1 has a positive effect for all cancer types, MAB randomized more breast cancer patients to arm 1 than MAN

(74.7 vs. 69.6 and 79.3 vs. 76.8). In all scenarios allocation to the effective arm was increased between 10% and 32% compared to balanced randomization.

For ovarian cancer, which was the cancer type with lowest accrual rate in the study, the average number of patients randomized to arm 1 was higher than for breast cancer in scenarios 3 through 7. The higher proportions are due to the slower accrual rate, which gives the adaptive algorithm more time to accumulate information.

When there is a strong treatment effect for multiple cancer types, borrowing of information across cancer leads to assigning more patients to arm 1 for the cancer types that do not benefit from the treatment. For instance, in scenario 6 MAB assigns more patients with colon/rectum and bladder cancer to arm 1 than balanced randomization (76 compared to 60) because of the strong evidence of efficacy for three other cancer types. Supplementary Table S7 illustrates additional scenarios. Scenario 11 is nearly identical to scenario 6, but arm 1 is inferior to the control for colon/rectum and bladder cancer. In this case, MAB enrolled on average 52.7 and 48 patients to arm 1 in these cancer types compared to 60 with balanced randomization.

To further evaluate the adaptive approach, we set $\alpha = 0.1$ and calculated the power and type I error rates per cancer type for MAB, MAN, and balanced-randomization with $N_{d,m} = 240$ (Supplementary Table S8). Next, we used Monte-Carlo simulations to determine sample size requirements for MAN and balanced trial designs to achieve the same power as MAB, see Figure 1. For balanced designs, we consider both a single multi-arm study with three experimental treatments and one control arm (MB), and three separate two-arm studies (TB), each with an experimental arm and a control arm.

In scenarios 2 and 5, where arm 1 is only effective for breast cancer, MAB had 49% and 88% power with $N_{d,m} = 240$ patients if there was a moderate and strong treatment effect, respectively. MAN required 240 and 225 breast cancer patients to obtain the same power. In scenarios 3 and 6, where arm 1 has treatment effects for three cancer types, MAB has 50% and 91% power for ovarian cancer under moderate and strong effect respectively. Matching the power with MAN required 25 or 10 additional ovarian cancer patients (Figure 1), while MB required 75 or 90 additional patients respectively. In scenario 4, with moderate treatment effect across all cancers, MAN and MB required an additional 35 and 85 ovarian cancer patients to match MAB's power. In all scenarios, the redundant control arms in TB designs substantially increased sample size requirements to match the power of MAB.

Lastly, we implemented early stopping using the rules defined in Section 4.1 and compared adaptive (MAB) and balanced designs (MB and TB) under the same scenarios. Specifically, the combination (a, d) is dropped for futility if the corresponding posterior probability of a positive treatment effect is less than 5%, that is, $\lambda' = 0.05$ and $s_3 = 0$. For early stopping for efficacy, we choose $N_{\min} = 30$ and the parameters $(s_1, s_2) = (3.5, 0.8)$.

Figure 2 shows results obtained with 1000 simulated trials under scenario 6, where arm 1 has a strong effect (0.5 vs. 0.3) for three of five cancer types. The x-axis represents the

Table 1

Mean^{Q3} number of breast, endometrium, colon/rectum, bladder, and ovarian cancer patients randomized to each treatment for a subpopulation-stratified trial with five cancer types, three experimental arms, and cancer-specific control arms for a design with strong and no borrowing of information across cancer types

Cancer	Strong borrowing of information					No borrowing of information				
	Breast	Endometrium	Colon/rectum	Bladder	Ovarian	Breast	Endometrium	Colon/rectum	Bladder	Ovarian
Scenario 1	No effect for any cancer									
control	77.1	76.9	77.3	77.3	77.6	76.5	76.4	76.7	76.8	76.8
arm 1	54.4	54.4	54.5	54.4	54.3	54.4	54.5	54.4	54.4	54.3
$a \geq 2$	54.4	54.3	54.4	53.4	54.4	54.4	54.4	54.4	54.5	54.4
Scenario 2	Moderate effect for breast cancer									
control	77.2	76.4	76.6	76.7	76.8	76.9	76.8	76.8	76.8	76.8
arm 1	67.5	63.0	61.9	60.8	60.7	69.6	54.4	54.5	54.5	54.4
$a \geq 2$	47.5	50.0	50.7	50.9	50.9	46.6	54.5	54.3	54.4	54.4
Scenario 3	Moderate effect for breast, endometrium, and ovarian cancer									
control	77.7	78.0	77.2	77.2	79.0	76.9	77.1	76.7	76.7	77.7
arm 1	72.3	75.0	68.9	68.4	76.8	69.6	71.2	54.6	54.5	71.8
$a \geq 2$	44.9	43.3	46.0	46.0	42.1	46.6	45.9	54.1	54.2	45.4
Scenario 4	Moderate effect for all five cancer types									
control	78.1	78.7	79.3	79.8	79.9	76.9	77.1	77.4	77.7	77.7
arm 1	74.7	77.5	78.7	79.5	79.7	69.6	71.2	71.4	71.7	71.8
$a \geq 2$	43.6	41.9	41.0	40.3	40.3	46.6	45.9	45.4	45.3	45.4
Scenario 5	Strong effect for breast cancer									
control	78.5	76.8	76.8	76.6	76.9	78.3	76.4	76.7	76.8	76.8
arm 1	75.6	70.2	68.1	67.3	67.0	76.8	54.5	54.6	54.5	54.1
$a \geq 2$	42.9	46.4	47.5	47.9	47.9	42.3	54.5	54.0	54.2	54.9
Scenario 6	Strong effect for breast, endometrium, and ovarian cancer									
control	79.5	79.8	78.5	78.8	80.8	78.3	78.8	76.7	76.7	79.6
arm 1	78.7	79.8	77.8	78.1	80.9	76.8	78.2	54.6	54.5	79.1
$a \geq 2$	40.8	40.2	41.9	41.5	39.2	42.3	41.5	54.1	54.2	40.6
Scenario 7	Strong effect for all five cancer types									
control	79.9	80.1	80.4	80.6	80.6	78.3	78.7	79.2	79.6	79.6
arm 1	79.3	80.2	80.6	80.8	80.9	76.8	78.2	78.6	79.0	79.1
$a \geq 2$	40.4	39.8	39.4	39.3	39.2	42.3	41.5	41.2	40.7	40.6

Arms $a=2,3$ have no treatment effects in all seven scenarios. Arm 1 has no effect in scenario 1, a positive effect for breast cancer in scenarios 2 and 5; for breast, endometrium, and ovarian cancer in scenarios 3 and 6; and for all five cancer types in scenarios 4 and 7.

time in weeks since the beginning of the trial, and the y-axis represents the probability of having previously rejected the null hypothesis $H_{d,a}$. Each point (x,y) of the plotted curves for MAB, MB, and TB in panels (a) and (b) indicates the proportion y of simulated trials that declared arm 1 effective before time x. The power for MAB at the end of the trial is 0.09 higher compared to MB and TB (88-90% for cancer 1,2,5 with MAB compared to 77-83% for MB and TB). The vertical bars in Figure 2 show the time when the proportion of simulated trials that declared arm 1 effective for MAB, MB, and TB crosses 70%. For breast cancer, which has the highest accrual rate, MAB reached 70% power after 93 weeks, whereas MB and TB reaches this power 12 and 65 weeks later. For cancer types with lower accrual rates like ovarian cancer, this difference increases further. MAB reaches the 70% threshold 62 weeks earlier than MB.

We also compared the MAB design to a balanced design that uses group-sequential stopping rules and block random-

ization. For each cancer type d , blocks of four patients are randomized using standard block permutations. Arm $a > 0$ is dropped for futility in cancer d , if $V'_{d,a}(i)$ (see (9)) falls below the futility boundary $b'(N_{d,a}(i))$. Conversely, the treatment is declared effective for d if the same statistic crosses the efficacy boundary $b(N_{d,a}(i))$. We use truncated O'Brien-Fleming efficacy boundaries (O'Brien and Fleming, 1979), $b'(n) = c/\sqrt{n}$ if $n \geq 30$, and $+\infty$ otherwise, and futility boundaries are defined by the predictive power method (Betensky, 2000). The boundaries were tuned so that the type I error for the one-sided test $H_{d,a}$ equals $\alpha = 0.1$. We compared MAB to the group-sequential design under the assumptions of scenario 6, where the effective arm has a strong treatment effect for three of the five cancer types (Supplementary Figure S6). For breast and ovarian cancer, the group-sequential design had 83.8% power of detecting the treatment effect for cancer $d = 1, 5$ compared to 88.8% and 90% for MAB. Using the group-sequential design, 70% of the simulated trials

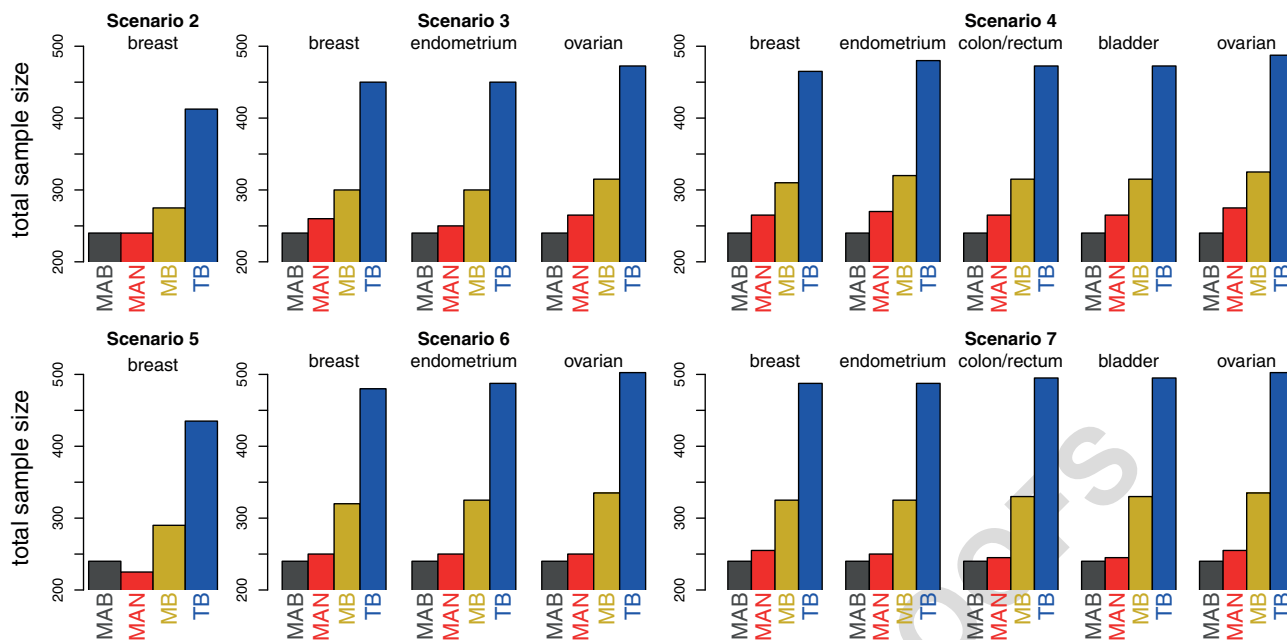


Figure 1. Sample size requirements to achieve equivalent levels of power. We display the total sample sizes for MAN, MB, and TB to achieve the same power as a MAB trial with 240 patients. MAB and MAN correspond to designs with and without borrowing of information across cancer types. MB and TB correspond to a balanced multi-arm design and three independent two-arm balanced designs. All scenarios refer to a trial with three experimental arms, five cancer types, and where only arm 1 has treatment effects, with response rates > 0.3 . Arm 1 has a positive effect for breast cancer in scenarios 2 and 5 (response rates 0.4 and 0.5), for breast, endometrium, and ovarian cancer in scenarios 3 and 6 (response rates 0.4 and 0.5), and for all cancer types in scenarios 4 and 7 (response rates 0.4 and 0.5), respectively.

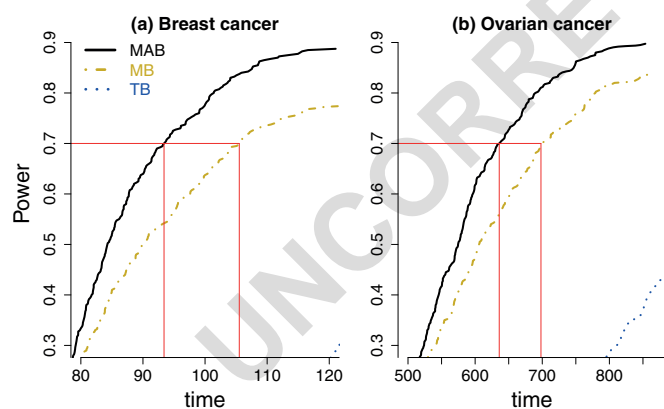


Figure 2. Probability of declaring arm 1 effective over the course of the trial. The x-axis represents the time in weeks since the beginning of the trial, and the y-axis denotes the probability of a positive finding. MAB correspond to a design with strong borrowing of information across cancer types, and MR and TB corresponds to a multi-arm and two-arm balanced design. Both MAB and MB have three experimental arms and only arm 1 has a positive treatment effect (PTE) of 0.5 versus 0.3 for 3 of the 5 cancer types (breast, ovarian and endometrium cancer). TB corresponds to an independent two-arm trial for each of the three experimental arms.

declared arm 1 effective for breast cancer during the first 96.4 weeks compared to 93 weeks with MAB. Similarly, for ovarian cancer the 70% threshold was reached by the group-sequential design approximately 50 weeks later than MAB.

5.2. Subpopulation-Finding Design: PI3K/Akt/mTOR Pathway

The *PI3K/Akt/mTOR* pathway signals several physiological functions, including cell survival/growth and mediates degradation of the tumor suppressor gene p53. Several genomic abnormalities activate the pathway and contribute to the genesis of multiple cancer types (Polivka and Janku, 2014). Multiple inhibitors of the pathway are currently in preclinical and clinical development (Fruman and Rommel, 2014) and are of potential use for different patient subpopulations. Here, we consider a trial with five experimental inhibitors without a control arm; subgroups are defined by abnormalities in the genes *PIK3CA*, *PIK3RI*, *PTEN*, and *mTOR*, $m^* = 4$. Patient eligibility is restricted to late stage *endometrial*, *colorectal* and *prostate cancer*, $d^* = 3$. Based on data from the Cancer Genome Atlas we used the patients accrual rates by cancer and biomarker profiles in Supplementary Table S9.

Powering a study to have high probability, say $\geq 80\%$, to detect a treatment effect for each combination (d, m, a) would require a large sample size and would result in long accrual periods for rare combinations (d, m) . The subpopulation-finding design aims to identify for each drug a the subgroups m with positive effect for at least one cancer type. Drug a is dropped early for futility in subgroup m if there is no evidence

Table 2

Scenarios for the subpopulation-finding design with PIK3CA, PIK3RI, PTEN, mTOR subpopulations and, endometrial, colorectal, and prostate cancer

	Subpopulations with positive effect	Cancer types with treatment effect
	$(p_{1,m,0}, p_{2,m,0}, p_{3,m,0}) = (0.1, 0.15, 0.05)$	
1	No subpopulation	No cancer type
2	PIK3CA	endometrial cancer
3	PIK3CA	All three cancer types
4	All 4 subpopulations	Endometrial cancer
5	All 4 subpopulations	All three cancer types
	$(p_{1,m,0}, p_{2,m,0}, p_{3,m,0}) = (0.1, 0.1, 0.1)$	
6	No subpopulation	No cancer type
7	PIK3CA	All three cancer types
8	PIK3RI	All three cancer types
9	PTEN	All three cancer types
10	mTor	All three cancer types
11	All four subpopulations	All three cancer types

An experimental arm with positive effect has a response probability equal to $p_{d,m,a} = p_{d,m,0} + 0.15$, where the response probability for the standard-of-care $p_{d,m,0}$ is independent of m as specified below.

of efficacy in any cancer type, that is, $\max_{d=1,2,3} \mathbb{P}[p_{d,a,m} > p_{d,0,m} | \Sigma_i] \leq b'_{a,m}(i)$; no early stopping rules for efficacy will be applied.

We consider several scenarios which are summarized in Table 2. The null hypotheses $H_{a,m}$ for the first five scenarios coincide with the set of response probabilities $\{p_{1,m,a} \leq 0.15, p_{2,m,a} \leq 0.1, p_{3,m,a} \leq 0.05\}$. In scenario 1, all arms are ineffective with response probabilities equal to 0.15, 0.1, and 0.05 for endometrial, colorectal and prostate cancer in all subpopulations. In scenarios 2 through 5 arm 1 has a positive effect in at least one subgroup and the remaining arms contain inhibitors without treatment effects. We consider response rates of arms with treatment effects equal to $p_{d,m,0} + 0.15$. Biomarkers are assumed to be mutually exclusive. We set $N_m = \sum_d N_{d,m} = 250$, that is, a total of 250 patients with mutation m are enrolled. The number of enrolled patients by cancer d , $N_{d,m}$, is a random variable whose expectation depends on the accrual rates of the combinations (d, m) .

Figure 3a shows the expected number of patients randomized to arm 1 by subpopulation and cancer type for all five scenarios. Supplementary Table S10 shows the corresponding type I error rates and the power for each scenario. Similar to Section 5.1, randomization is adapted with either strong (MAB) or no borrowing (MAN) of information across cancer types and subpopulations, or under a balanced design (MB). All three designs drop an arm for a subpopulation m early if the posterior probabilities of positive treatment effects fall below 0.1 for all cancer types, that is, $b'_{a,m}(i) = 0.1$.

In scenario 1, where treatments are ineffective for all combinations (d, m) , all three designs assigned on average approximately 50 patients per arm in the PIK3CA group, the PIK3RI group, the PTEN group, and the mTOR group;

in each group the distribution of cancer types was proportional to the accrual rate. When arm 1 is only effective for endometrial cancer patients with PIK3CA alterations (scenario 2), then MAB and MAN randomized on average 40.4 and 40.5 patients from this subgroup to the effective arm, compared to 24.9 patients for MB. The impact of borrowing information across cancer types is shown in scenario 3 (a positive effect for all cancer types in the PIK3CA population), where MAB randomized on average 97.1 patients with PIK3CA to the effective arm. The MAN and MB designs randomized on average 77.5 and 50 patients from the PIK3CA group to arm 1.

Similarly, the consequences of borrowing information across subpopulations can be seen in scenario 4 where the inhibitor in arm 1 is effective for all endometrial cancers in the study population. In this case MAB assigned on average 5 additional endometrial cancer patients from the PIK3CA group to the effective arm compared to scenario 2. Since MAN and MB do not borrow information across subpopulations, patients allocation remained identical to scenario 2 for both designs.

As in subpopulation-stratified designs, when there are treatment effects in several (d, m) combinations, borrowing of information resulted in more patients being assigned to arm 1 in the other (d, m) combinations without treatment effects. For instance, in scenario 4 the MAB design assigned on average 3 additional prostate cancer patients to arm 1 compared to MB. Supplementary Figure S9 shows the allocation of patients for additional scenarios where therapy 1 is superior to the control for some (d, m) combinations and inferior for other (d, m) combinations. For instance, in Scenario 14 arm 1 has a positive effect for patients with PIK3RI, PTEN, and mTOR alterations, but it is inferior to the historical control for the PIK3CA group. In this case, MAB assigned on average 14.9, 19.5, and 6.3 colorectal, endometrial and prostate cancer patients to arm 1 compared to 18.0, 24.9, and 7.1 patients with balanced randomization.

We also considered the precision of the estimated response probabilities at the end of the trial. Figure 3b and S7 show the mean and interquartile range of the estimated probabilities for the effective arms 1 for each (d, m) combination across 1000 simulated trials under MAB, MAN, and MB. While it is known that adaptation can yield biased estimates, the maximum bias that we observed was 0.01 under MAB and MB, and 0.03 for MAN.

With MAB, which randomized more patients to the effective arm than MB, we observed smaller interquartile ranges for the response rate estimates of arm 1. For prostate cancer, which is the malignancy with lowest accrual rate in the simulation study, the interquartile range of the estimated response rate for arm 1 was up to 40% smaller with MAB than seen with MB (PIK3RI in Scenario 4).

The original design of NCI-MATCH defines independent sub-studies for each (m, a) combination, and uses a two-stage analysis plan to evaluate efficacy (Simon, 1989). To facilitate a comparison of this design to the subpopulation-finding designs, we consider additional scenarios (6 through 11 in Table 2) where the historical response rates does not vary by cancer types, that is $H_{a,m} = \{p_{1,m,a} \leq 0.1, p_{2,m,a} \leq 0.1, p_{3,m,a} \leq 0.1\}$. A two-stage Simon design targeting 10% types I type II

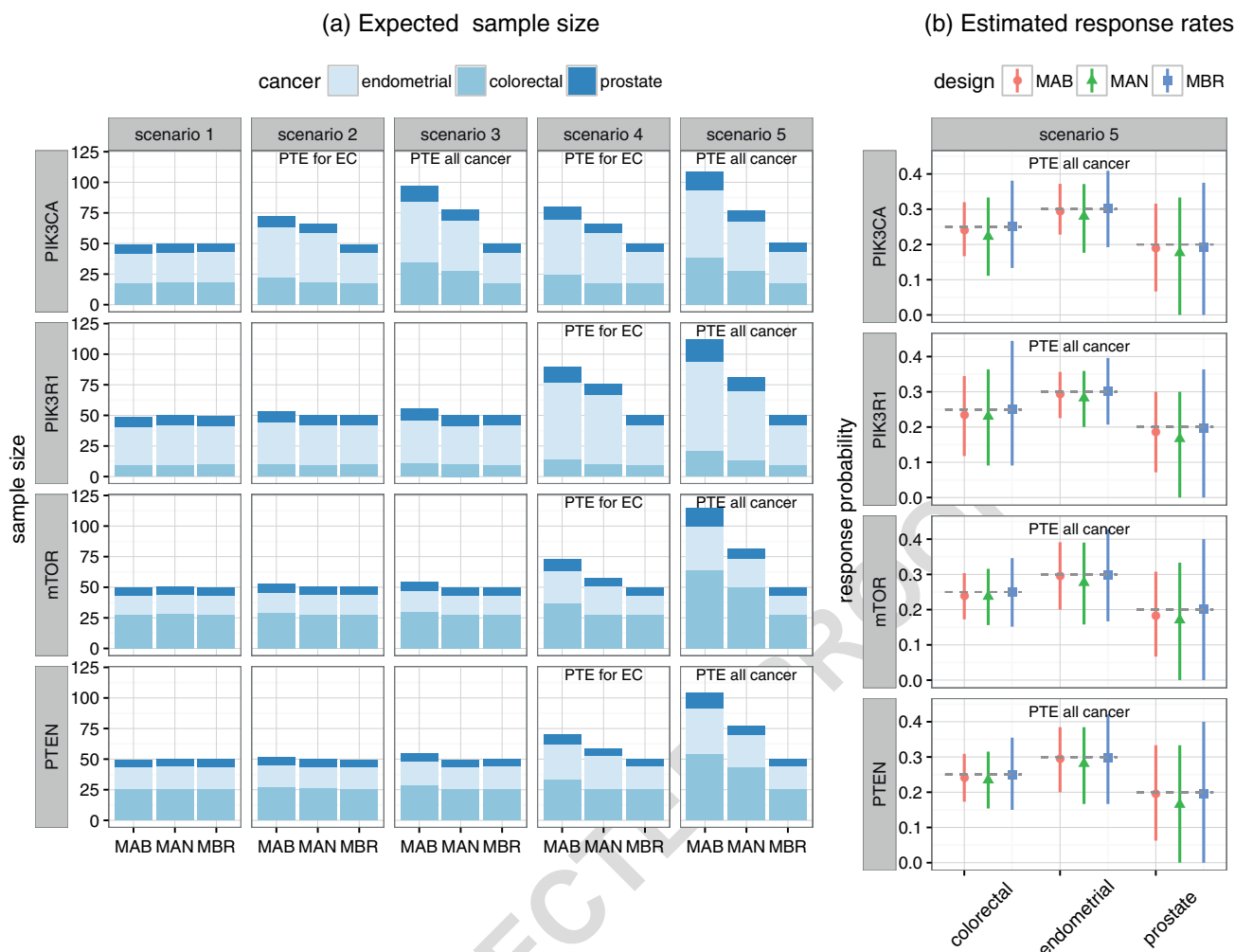


Figure 3. Mean sample size and precision of the point estimates. Results are based on 1000 simulations. Multiple scenarios are defined for a subpopulation-finding design with five experimental arms, four subpopulations, three cancer types, and a maximum sample size of 250 patients per subpopulations. Panel a: Expected number of patients treated with the effective therapy for MAB, MAN, and MB. Panel b: Mean response probability estimates for the effective therapy (dots, triangles, and squares represent MAB, MAN, and MB) and inter-quartile range (vertical bar) of the estimates across 1000 simulations. The dotted horizontal lines indicate the true response probabilities.

errors bounds, with response rates 0.1 under $H_{a,m}$ and 0.25 under the alternative, would initially allocate 21 patients to each combination (m, a) . If three or more of the initial patients in (m, a) respond, then 29 additional patients from subpopulation m are allocated to arm a . With one effective therapy and four ineffective therapies, the design has an expected sample size of 173 patients per subpopulation; 47.8 patients for the effective therapy; and 31.2 for each of the remaining arms. Thus, in scenarios 6–11 we set the sample size under MAB equal to $N_m = 173$ for each subpopulation m . Figure S8 and Table S11 in the Web-based Supplementary Material summarizes the comparison. In scenario 6, where no therapy is effective for any disease-marker combination, MAB randomized in each biomarker group on average three more patients to arm 1 compared to Simon’s design.

In scenarios 7 through 10, the effective therapy has a positive effect only in a single biomarker subpopulation.

For instance in scenario 8, where the treatment effect is restricted to RIK3RI, MAB assigns 36% more RIK3RI patients to the effective treatment compared to Simon’s design. When the effective therapy has positive effects across all subpopulations (scenario 11), MAB randomized between 35% more PTEN patients and 61% more mTOR patients to the effective arm compared to the two-stage Simon design.

6. Discussion

The development of anti-cancer therapies focuses increasingly on compounds which target genomic pathways that are connected with multiple malignancies. In this work, we proposed a broad class of designs for basket trials, which facilitates the exploration of several treatments in multiple disease types across biomarker-driven subgroups. Each design combines a Bayesian hierarchical model, with a response-adaptive

1 treatment assignment, and a set of sequential stopping
 2 rules. As illustration, we examined two types of studies,
 3 the subpopulation-finding and the subpopulation-stratified
 4 design. The subpopulation-finding design aims to identify a
 5 subgroup of patients, within a set of subpopulations, that
 6 benefits from the experimental therapy. In contrast, the
 7 subpopulation-stratified design identifies cancer types within
 8 a biomarker-homogeneous population which respond to a
 9 therapy.

10 There has been an extensive debate about the merits and
 11 drawbacks of Bayesian adaptive randomization. Opponents
 12 argue that adaptive randomization increases the overall sam-
 13 ple size, especially in the two-arm setting (Korn and Freidlin,
 14 2011). Thus, while the relative number of patients receiving
 15 the best treatment option may increase, a larger overall sam-
 16 ple size might also result in more patients being exposed to
 17 an inferior arm (Korn and Freidlin, 2011). Compared to bal-
 18 anced randomization, adaptive randomization requires more
 19 resources dedicated to the design and implementation of the
 20 trial (Korn and Freidlin, 2011). Ethical concerns have been
 21 recently discussed by Hey and Kimmelman (2015) and in the
 22 subsequent letters (Berry, 2015; Joffe and Ellenberg, 2015;
 23 Korn and Freidlin, 2015; Lee, 2015; Saxman, 2015). Most
 24 advocates of adaptation agree that benefits are small in the
 25 two-arm settings and more attractive in multi-arm trials.
 26 Wason and Trippa (2014) compared group-sequential designs
 27 with outcome adaptive randomization, and quantified gains
 28 in power under adaptive randomization when a few superior
 29 treatments exists, as well as a slight increase in the aver-
 30 age sample size when none of the experimental arms has a
 31 treatment effect.

32 To reduce the variability of treatment assignment under
 33 outcome adaptive randomization, we introduce a correction
 34 factor (8) that enforce a minimum enrollment to each com-
 35 bination of cancer type, biomarker, and treatment. In the
 36 extreme cases this correction yields either stratified balanced
 37 randomization or standard Bayesian adaptive randomization.
 38 Potential future directions would be to extend our method to
 39 incorporate clustering of treatment effects across cancer types
 40 and biomarker subgroups, accounting for the possibility that
 41 treatments may show strong effects for some disease-marker
 42 combinations, but remain ineffective for other combinations.
 43 A Bayesian nonparametric model, such as a Dirichlet prior,
 44 could be utilized for the treatment effects distribution across
 45 patients subgroups.

46 Our testing procedures and stopping rules satisfy frequen-
 47 tist constraints on type I errors, as expected in the regulatory
 48 process of new drugs' development. The Bayesian compo-
 49 nent of the proposed designs uses a hierarchical model that
 50 drives patient allocation. In the early phase of development
 51 on new treatments, where signal seeking is a major goal, a
 52 Bayesian testing procedure could also be considered. How-
 53 ever, the majority of recent phase II basket trials are designed
 54 with targeted types I and II error rates, including both
 55 NCI-MATCH and CUSTOM (Conley and Doroshow, 2014;
 56 Lopez-Chavez et al., 2015). When the conclusions of the trial
 57 are reported to the medical community, p-values and hypothe-
 58 sis testing based on type I error rates are de facto the accepted
 59 standard when communicating results. This is the main moti-
 60 vation for frequentist analysis after Bayesian randomization.

Examples of Bayesian designs which contain frequentist
 hypothesis testing procedures are discussed in (Trippa et al.,
 2012; Wason and Trippa, 2014; Ventz and Trippa, 2015).

We follow the practice of recent basket trials, such as
 NCI-MATCH and Lung-MAP, and do not adjust for mul-
 tiplicity when testing several therapies in multiple subgroups
 and cancer types. There is no general agreement on whether
 one should correct for multiplicity in multi-arm trials
 (Proschan and Waclawiw, 2000). The algorithm for type I
 error control can in principle be extended to the control of
 the FDR, or Bonferroni corrections can be applied.

Many cancer centers now routinely measure the genomic
 profile of their patients. With the decreasing cost of genomic
 profiling, this is likely to become standard in the foreseeable
 future. Several ongoing basket trials implement multi-
 cancer studies with biomarker-defined subgroups (Conley and
 Doroshow, 2014). Our Bayesian model uses information from
 all subgroups and cancer types and randomizes patients with
 higher probability to the most effective treatments. This
 approach has the potential to accelerate drug development
 and to provide faster access to effective treatments for cancer
 patients.

7. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections
 2.2, 3, 4.1, 5, and an R package that implements the designs
 are available with this article at the Biometrics website on
 Wiley Online Library.

ACKNOWLEDGMENTS

This work was supported by funding from the Wong Family.
 In addition Giovanni Parmigiani was supported by the NCI
 grant 5P30 CA006516-46.

REFERENCES

- An, M.-W., Lu, X., Sargent, D. J., and Mandrekar, S. J. (2015).
 The direct assignment option as a modular design compo-
 nent: an example for the setting of two predefined subgroups.
Computational and Mathematical Methods in Medicine^{Q4}.
- Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry,
 D. A., and Esserman, L. J. (2009). I-SPY 2: An adap-
 tive breast cancer trial design in the setting of neoadjuvant
 chemotherapy. *Clinical Pharmacology & Therapeutics* **86**,
 97–100.
- Barry, W. T., Perou, C. M., Marcom, P. K., Carey, L. A., and
 Ibrahim, J. G. (2015). The use of Bayesian hierarchical mod-
 els for adaptive randomization in biomarker-driven phase ii
 studies. *Journal of Biopharmaceutical Statistics* **25**, 66–88.
- Berry, D. A. (2015). Commentary on Hey and Kimmelman. *Clin-
 ical Trials* **12**, 107–109.
- Betensky, R. A. (2000). Alternative derivations of a rule for early
 stopping in favor of H₀. *American Statistician* **54**, 35–39.
- Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch,
 M., and Racine-Poon, A. (2009). Confirmatory adaptive
 designs with Bayesian decision tools for a targeted therapy
 in oncology. *Statistics in Medicine* **28**, 1445–1463.
- Conley, B. A. and Doroshow, J. H. (2014). Molecular analysis
 for therapy choice: NCI MATCH. *Seminars in Oncology* **41**,
 297–299.

- 1 Freidlin, B., Jiang, W., and Simon, R. (2010). The cross-validated
2 adaptive signature design. *Clinical Cancer Research* **16**,
3 691–698.
- 4 Fruman, D. A. and Rommel, C. (2014). PI3K and cancer: lessons,
5 challenges and opportunities. *Nature Reviews Drug Discov-*
6 *ery* **13**, 140–156.
- 7 Hey, S. P. and Kimmelman, J. (2015). Are outcome-adaptive allo-
8 cation trials ethical? *Clinical Trials* **12**, 102–106^{Q5}.
- 9 Hobbs, B. P., Sargent, D. J., and Carlin, B. P. (2012). Commensu-
10 rate priors for incorporating historical information in clinical
11 trials using general and generalized linear models. *Bayesian*
12 *Analysis* **7**, 639.
- 13 Joffe, S. and Ellenberg, S. S. (2015). Commentary on Hey and
14 Kimmelman. *Clinical Trials* **12**, 116–118.
- 15 Korn, E. L. and Freidlin, B. (2011). Outcome-adaptive randomiza-
16 tion: Is it useful? *Journal of Clinical Oncology* **29**, 771–776.
- 17 Korn, E. L. and Freidlin, B. (2015). Commentary on Hey and
18 Kimmelman. *Clinical Trials* **12**, 122–124.
- 19 Lee, J., van Hummelen, P., Go, C., Palescandolo, E., Jang, J., Park,
20 H. Y., Kang, S. Y., Park, J. O., Kang, W. K., MacConaill, L.,
21 and Kim, K.-M. (2012). High-throughput mutation profil-
22 ing identifies frequent somatic mutations in advanced gastric
23 adenocarcinoma. *PLoS ONE* **7**, e38892.
- 24 Lee, J. J. (2015). Commentary on hey and kimmelman. *Clinical*
25 *Trials* **12**, 110–112.
- 26 Lee, J. J., Xuemin, Gu, and Suyu, Liu (2010). Bayesian adap-
27 tive randomization designs for targeted agent development.
28 *Clinical Trials* **7**, 584 – 596.
- 29 Lopez-Chavez, A., Thomas, A., Rajan, A., and et al (2015). Molec-
30 ular profiling and targeted therapy for advanced thoracic
31 malignancies: A biomarker-derived, multiarm, multihistol-
32 ogy phase ii basket trial.^{Q6} *JCO* **33**, 1000–1007.
- 33 Mehta, C., Schäfer, H., Daniel, H., and Irl, S. (2014). Biomarker
34 driven population enrichment for adaptive oncology trials
35 with time to event endpoints. *Statistics in Medicine* **33**,
36 4515–4531.
- 37 Mehta, C. R. and Gao, P. (2011). Population enrichment designs:
38 case study of a large multinational trial. *Journal of Biophar-*
39 *maceutical Statistics* **21**, 831–845.
- 40 O’Brien, P. C. and Fleming, T. R. (1979). A multiple testing
41 procedure for clinical trials. *Biometrics* **35**, 549–556.
- 42 Pocock, S. J. (1977). Group sequential methods in the design and
43 analysis of clinical trials. *Biometrika* **64**, 191–199.
- 44 Polivka, J. and Janku, F. (2014). Molecular targets for cancer
45 therapy in the PI3K/AKT/mTOR pathway. *Pharmacology*
46 *& Therapeutics* **142**, 164–175.
- 47 Pratt, J. W. (1981). Concavity of the log likelihood. *Journal of*
48 *the American Statistical Association* **76**, 103–106.
- 49 Proschan, M. A. and Waclawiw, M. A. (2000). Practical guide-
50 lines for multiplicity adjustment in clinical trials. *Controlled*
51 *Clinical Trials* **21**, 527–539.
- 52 Robinson, S. D., O’Shaughnessy, J. A., Cowey, C. L., and Konduri,
53 K. (2014). BRAF V600E-mutated lung adenocarcinoma with
54 metastases to the brain responding to treatment with vemu-
55 rafenib. *Lung Cancer* **85**, 326–330.
- 56 Rosenberger, W. F. and Hu, F. (1999). Bootstrap meth-
57 ods for adaptive designs. *Statistics in Medicine* **18**,
58 1757–1767.
- 59 Saxman, S. B. (2015). Commentary on hey and kimmelman. *Clin-*
60 *ical Trials* **12**, 113–115.
- Simon, R. (1989). Optimal 2-stage designs for phase II trials.
Controlled Clinical Trials **10**, 1–10.
- Thall, P., Fox, P., and Wathen, J. (2015). Statistical contro-
versies in clinical research: Scientific and ethical problems
with adaptive randomization in comparative clinical trials.
Annals of Oncology **26**, 1621–1628.
- Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker,
L. H., and Benjamin, R. S. (2003). Hierarchical Bayesian
approaches to phase II trials in diseases with multiple sub-
types. *Statistics in Medicine* **22**, 763–780.
- Thall, P. F. and Wathen, K. J. (2007). Practical Bayesian adaptive
randomisation in clinical trials. *European Journal of Cancer*
5, 859–866.
- Thompson, W. R. (1933). On the likelihood that one unknown
probability exceeds another in view of the evidence of two
samples. *Biometrika* **25**, 285–294.
- Trippa, L., Lee, E. Q., Wen, P. Y., Batchelor, T. T., Clough-
esy, T., Parmigiani, G., and Alexander, B. M. (2012).
Bayesian adaptive randomized trial design for patients with
recurrent glioblastoma. *Journal of Clinical Oncology* **30**,
3258–3263.
- Ventz, S. and Trippa, L. (2015). Bayesian designs and the control of
frequentist characteristics: A practical solution. *Biometrics*
71, 218–226.
- Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Jr.,
L. A. D., and Kinzler, K. W. (2013). Cancer genome land-
scapes. *Science* **339**, 1546–1558.
- Wang, S.-J., James Hung, H., and O’Neill, R. T. (2009). Adaptive
patient enrichment designs in therapeutic trials. *Biomedical*
Journal **51**, 358–374.
- Wang, S.-J., O’Neill, R. T., and Hung, H. (2007). Approaches
to evaluation of treatment effect in randomized clinical
trials with genomic subset. *Pharmaceutical Statistics* **6**,
227–244.
- Wason, J. M. S. and Trippa, L. (2014). A comparison of
Bayesian adaptive randomization and multi-stage designs
for multi-arm clinical trials. *Statistics in Medicine* **33**,
2206–2221.
- Wathen, J. K., Thall, P. F., Cook, J. D., and Estey, E.
H. (2008). Accounting for patient heterogeneity in
phase II clinical trials. *Statistics in Medicine* **27**,
2802–2815.
- Zhou, X., Liu, S., Kim, E. S., Herbst, R. S., and Lee, J. J.
(2008). Bayesian adaptive design for targeted therapy devel-
opment in lung cancer - A step toward personalized medicine.
Clinical Trials **5**, 181–193.

Received February 2016. Revised November 2016.

Accepted January 2017.

AUTHOR QUERY FORM

JOURNAL: BIOMETRICS

Article: BIOM12668

Dear Author,

During the copyediting of your paper, the following queries arose. Please respond to these by annotating your proofs with the necessary changes/additions using the E-annotation guidelines attached after the last page of this article.

We recommend that you provide additional clarification of answers to queries by entering your answers on the query sheet, in addition to the text mark-up.

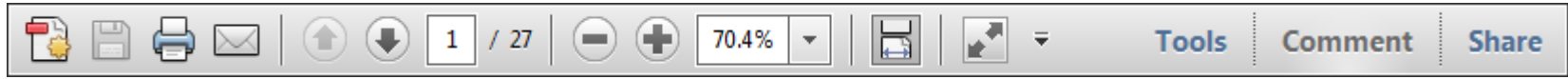
<i>Query No.</i>	<i>Query</i>	<i>Remak</i>
Q1:	Please confirm that given names (red) and surnames/family names (green) have been identified correctly.	
Q2:	Please check the change made.	
Q3:	Please check the presentation of all the Tables.	
Q4:	Please provide volume number and page range for this reference.	
Q5:	Please check the presentation of journal title for this reference.	
Q6:	As per the style of the journal et al. is allowed after six author. Please provide list of all the authors for this reference then add et al.	

UNCORRECTED PROOFS

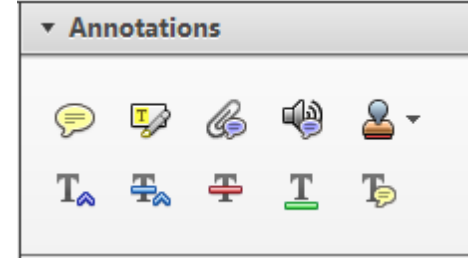
Required software to e-Annotate PDFs: Adobe Acrobat Professional or Adobe Reader (version 8.0 or above). (Note that this document uses screenshots from Adobe Reader X)

The latest version of Acrobat Reader can be downloaded for free at: <http://get.adobe.com/reader/>

Once you have Acrobat Reader open on your computer, click on the [Comment](#) tab at the right of the toolbar:



This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the [Annotations](#) section, pictured opposite. We've picked out some of these tools below:



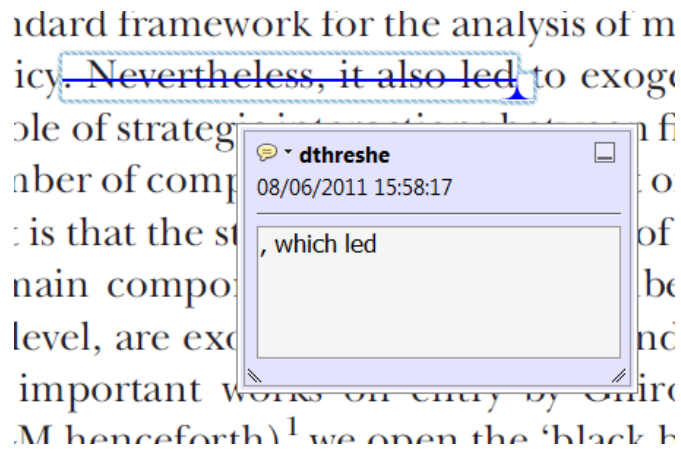
1. Replace (Ins) Tool – for replacing text.



Strikes a line through text and opens up a text box where replacement text can be entered.

How to use it

- Highlight a word or sentence.
- Click on the [Replace \(Ins\)](#) icon in the Annotations section.
- Type the replacement text into the blue box that appears.



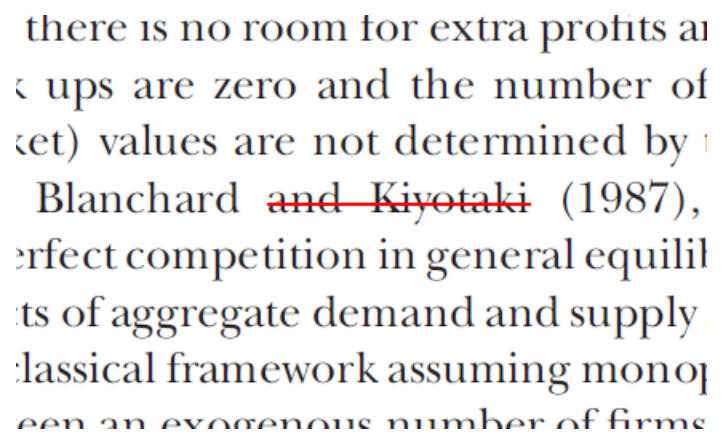
2. Strikethrough (Del) Tool – for deleting text.



Strikes a red line through text that is to be deleted.

How to use it

- Highlight a word or sentence.
- Click on the [Strikethrough \(Del\)](#) icon in the Annotations section.



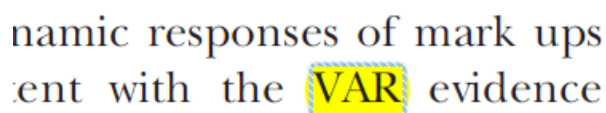
3. Add note to text Tool – for highlighting a section to be changed to bold or italic.



Highlights text in yellow and opens up a text box where comments can be entered.

How to use it

- Highlight the relevant section of text.
- Click on the [Add note to text](#) icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.



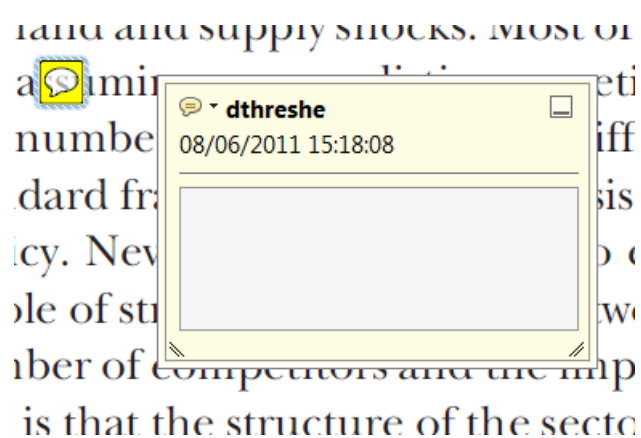
4. Add sticky note Tool – for making notes at specific points in the text.



Marks a point in the proof where a comment needs to be highlighted.

How to use it

- Click on the [Add sticky note](#) icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.



USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

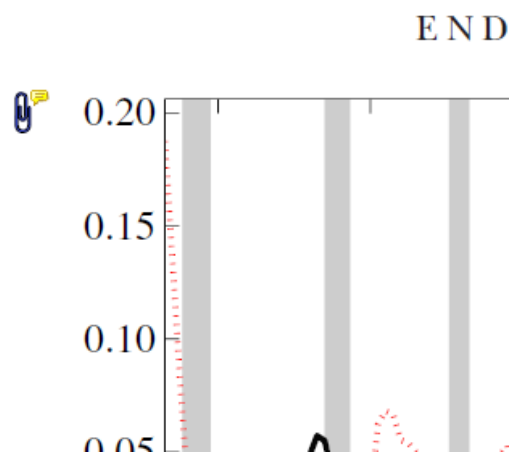
5. Attach File Tool – for inserting large amounts of text or replacement figures.



Inserts an icon linking to the attached file in the appropriate place in the text.

How to use it

- Click on the [Attach File](#) icon in the Annotations section.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.



6. Add stamp Tool – for approving a proof if no corrections are required.



Inserts a selected stamp onto an appropriate place in the proof.

How to use it

- Click on the [Add stamp](#) icon in the Annotations section.
- Select the stamp you want to use. (The [Approved](#) stamp is usually available directly in the menu that appears).
- Click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).

of the business cycle, starting with the
 on perfect competition, constant ret
 production. In this environment goods
 extra profits and the market for marke
 he market for goods is determined by the model. The New-Key
 otaki (1987), has introduced produc
 general equilibrium models with nomin
 and market-clearing. Most of this literat

APPROVED

Drawing Markups

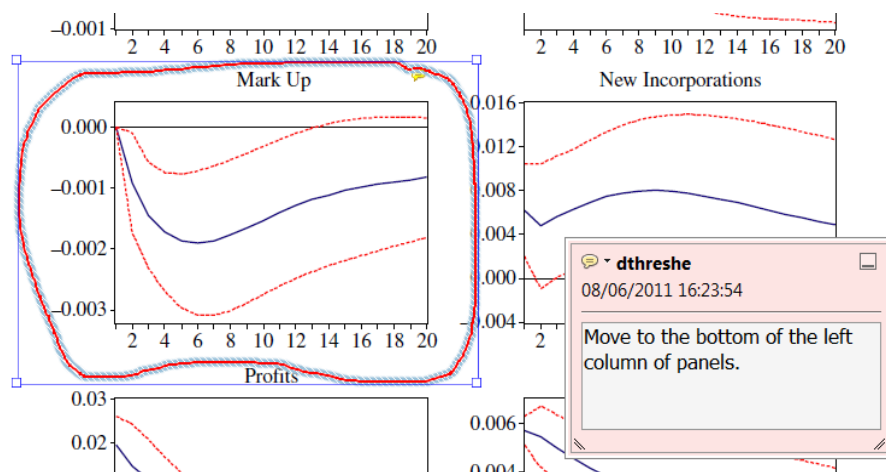


7. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.

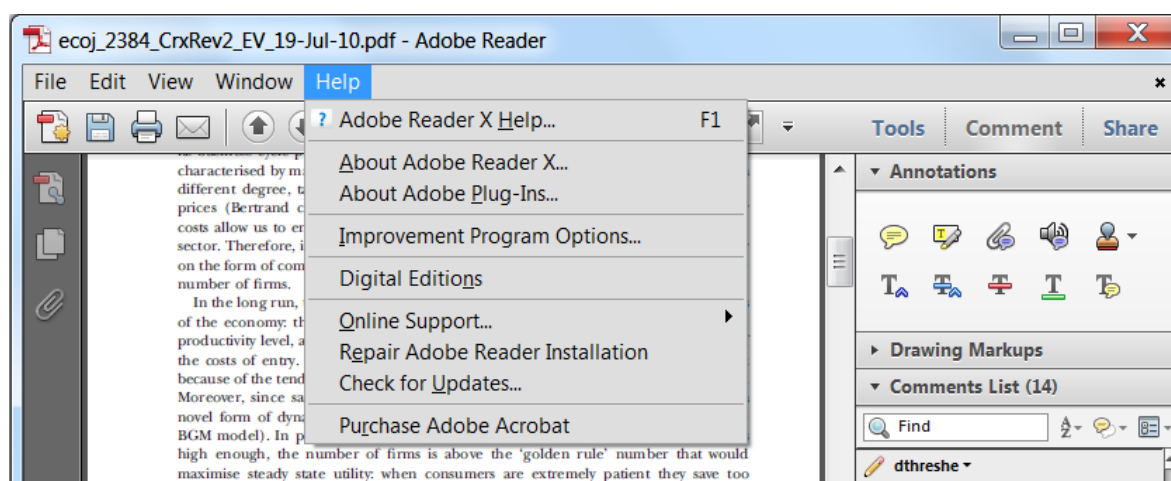
Allows shapes, lines and freeform annotations to be drawn on proofs and for comment to be made on these marks..

How to use it

- Click on one of the shapes in the [Drawing Markups](#) section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.





For further information on how to annotate proofs, click on the [Help](#) menu to reveal a list of further options:



Proof Correction Marks

Please correct and return your proofs using the proof correction marks below. For a more detailed look at using these marks please reference the most recent edition of The Chicago Manual of Style and visit them on the Web at: <http://www.chicagomanualofstyle.org/home.html>

<i>Instruction to typesetter</i>	<i>Textual mark</i>	<i>Marginal mark</i>
Leave unchanged	... under matter to remain	<i>stet</i>
Insert in text the matter indicated in the margin	^	^ followed by new matter
Delete	Ʒ through single character, rule or underline or Ʒ through all characters to be deleted	<i>del</i>
Substitute character or substitute part of one or more word(s)	Ƨ through letter or — through characters	new character Ƨ or new characters Ƨ
Change to italics	— under matter to be changed	<i>ital</i>
Change to capitals	≡ under matter to be changed	<i>Caps</i>
Change to small capitals	≡ under matter to be changed	<i>sc</i>
Change to bold type	~ under matter to be changed	<i>bf</i>
Change to bold italic	~ under matter to be changed	<i>bf+ital</i>
Change to lower case	Ɔ	<i>lc</i>
Insert superscript	√	√ under character e.g. √
Insert subscript	^	^ over character e.g. ^
Insert full stop	⊙	⊙
Insert comma	↵	↵
Insert single quotation marks	↙ ↘	↙ ↘
Insert double quotation marks	↵ ↶	↵ ↶
Insert hyphen	=	=
Start new paragraph	¶	¶
Transpose	┌┐	┌┐
Close up	linking  characters	
Insert or substitute space between characters or words	#	#
Reduce space between characters or words	˘	˘

WILEY

Additional reprint and journal issue purchases

Should you wish to purchase additional copies of your article, please click on the link and follow the instructions provided:
<https://caesar.sheridan.com/reprints/redir.php?pub=10089&acro=BIOM>

Corresponding authors are invited to inform their co-authors of the reprint options available.

Please note that regardless of the form in which they are acquired, reprints should not be resold, nor further disseminated in electronic form, nor deployed in part or in whole in any marketing, promotional or educational contexts without authorization from Wiley. Permissions requests should be directed to mailto: permissionsus@wiley.com

For information about 'Pay-Per-View and Article Select' click on the following link: <http://wileyonlinelibrary.com/ppv>