Published in final edited form as:

Stat Biopharm Res. 2019 ; 11(3): 210–219. doi:10.1080/19466315.2018.1518259.

Incorporating biomarkers to improve statistical power of immunotherapeutic neoadjuvant clinical trials in patients with triple-negative breast cancer

Feng Gao^{1,2}, Guoqiao Wang², Jingqin Luo^{1,2}, Jingxia Liu^{1,2}, Ling Chen², Chengjie Xiong^{2,*}

¹Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine, St. Louis, MO, USA

²Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA

Abstract

Neoadjuvant (preoperative) approach to breast cancer treatment has become widely accepted. Traditionally the primary objective of neoadjuvant treatment is to improve subsequent surgical intervention and the effectiveness are often evaluated by its ability achieving complete pathological response (pCR), the eradication of the malignant disease in the breast and axillary lymph nodes. More recently, neoadjuvant treatment has also become recognized as an in vivo, preoperative ‘window of opportunity’ to explore the efficacy of novel agents such as immunotherapies where a wealth of tumor biomarkers are also routinely collected to quantify antitumor immunity. However, one challenge to combine the traditional pCR and efficacy biomarkers is that these tumor biomarkers are only partially available and cannot be measured in patients who have achieved pCR. In this article, a stepwise hypothesis testing procedure is proposed to combine a continuous tumor biomarker with the conventional binary endpoint in a two-arm randomized phase II superiority trial to improve statistical power. The operating characteristics of proposed procedure is illustrated with a real-world example and the performance is also evaluated numerically.

Keywords

multiple hypotheses; multiple endpoints; neoadjuvant clinical trials; cancer immunotherapy; immune biomarkers

1. Introduction

Triple negative breast cancer (TNBC) is characterized by the lack of expression of estrogen, progesterone and HER2 receptors. It accounts for 15% to 20% of invasive breast cancers diagnosed in the United States and affects approximately 200,000 women annually (Berry et al 2006). Patients with TNBC suffer from a worse initial prognosis among all breast cancer subtypes, and chemotherapy remains standard treatment due to the lack of established

*Corresponding Author: Chengjie Xiong, PhD, Division of Biostatistics, Campus Box 8067, Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, MO 63110, chengjie@wustl.edu.

targeted therapies against a nuclear or a surface receptor in TNBC. In recent years, neoadjuvant (pre-operative) chemotherapy has become a widely accepted therapeutic modality to locally advanced TNBC. It offers the advantages of downstaging the disease and makes patients possible candidates for breast-conserving surgery rather than mastectomy. The treatment efficacy in the neoadjuvant setting is often measured by the pathological complete remission (pCR), defined as the eradication of the malignant disease in the breast and in the axillary lymph nodes. Numerous studies have consistently demonstrated improved survival outcomes in patients achieved pCR (Liedtke et al 2008).

Neoadjuvant approach also provides prognostic information that could guide the choice of adjuvant (post-operative) treatments (Thompson and Thompson 2012). Early identification of poor response or disease progression is important to spare patients of receiving ineffective therapy. The neoadjuvant platform creates an excellent model to assess pathologic tumor response and analyze tissue upon completion of systemic chemotherapy. Because the primary tumour remains intact during therapy, it avoids the issue of adaptive resistance and pretreatments that can be problematic in the advanced disease setting and provides an in-vivo “window of opportunity” to study the activity of novel agents. This allows for the identification of molecular changes induced by treatment, and permits determination of predictors of chemotherapy resistance/sensitivity to better individualize management. Consequently, in recent years the neoadjuvant clinical trials have emerged as a powerful platform to assess novel therapeutic modalities such as checkpoint blockade therapy or immunotherapy in general.

Since ipilimumab entered the treatment landscape for advanced melanoma in 2011, immunotherapy has continued to revolutionize cancer therapy and the number of neoadjuvant trials exploring the combination of chemotherapy with immunotherapy is rapidly increasing in triple-negative breast cancer (Stagg and Allard 2013). One unique feature of these immunotherapeutic trials is that a wealth of biomarkers are routinely collected to monitor antitumor efficacy or understand its mechanism. There is a dire need to incorporate such immunogenicity biomarker(s) under the neoadjuvant setting for several reasons.

- A neoadjuvant trial intends to select promising regimens for further evaluation in a timely manner. Unlike the conventional adjuvant confirmatory trials to seek a definitive answer regarding an intervention, the goal of a neoadjuvant trial is to screen new treatment via a preoperative platform and the trial is often conducted at relatively small scale.
- The conventional endpoint of pCR alone may be inadequate to measure the therapeutic efficacy for a neoadjuvant immunotherapeutic trial. Unlike traditional cytotoxic anti-tumor agents that work by directly killing existing tumor cells, an immunotherapeutic agent usually works on selective molecular targets and modulates tumor environment. Hence, it often provides the benefits of prolonged duration of response or survival rather than tumor shrinkage.
- pCR alone may require a relatively large sample size. Historically pCR can be achieved in one-fifth or one-third breast cancer patients and a neoadjuvant trial is

often designed to target a large effect size. The recent studies have demonstrated 50% pCR rate in TNBC patients (Sikov et al. 2015). It may no longer be realistic to assume an effect size of 20% or 30% difference in the pCR.

- There exist biomarkers known to be predictive of overall survival or progression-free survival. Although survival endpoints themselves are rarely available under a neoadjuvant trial due to inadequate follow-up, some biomarkers may serve as surrogate endpoints for survival outcomes. The tumor-infiltrating lymphocyte (TIL), for example, has been confirmed in multiple independent studies to be associated with survival outcomes in TNBC following adjuvant or neoadjuvant chemotherapy (Loi et al 2013; Denkert et al 2015).

Incorporating TIL into pCR is therefore expected to enrich information regarding treatment efficacy and increase statistical power in neoadjuvant immunotherapeutic trials. However, one potential challenge in combining these two endpoints is that sometimes TIL is only partially available. It cannot be measured if patients have achieved pCR or in patients whose core biopsies fail to catch adequate tumor cells. In the present study, a stepwise hypothesis testing procedure is proposed to combine TIL as a secondary endpoint with the conventional primary endpoint (pCR) to improve statistical power in neoadjuvant trials for TNBC patients. The remainder of this paper is structured as follows. In Section 2, we first review commonly used approaches for multiple hypotheses and then describe a procedure to incorporate a partially available continuous biomarker into a clinical trial as a secondary efficacy endpoint. In Section 3, the operating characteristics of the proposed procedure are illustrated by re-visiting a real-world example and the performance is also assessed numerically with a simple simulation. We finally conclude with a discussion in Section 4.

2. Method

2.1 Methods to address multiple hypothesis tests in clinical trials

Randomized clinical trials have become the gold standard in clinical research to evaluate an experimental treatment against a control (either placebo or standard of care). A traditional clinical trial is often designed based on the primary endpoint with a single hypothesis test that will be performed at the end of study. The final sample size for such a design is known in advance and calculated to control the Type I (α) and Type II (β) error rates in detecting scientifically meaningful effect size at specific levels. Multiple hypothesis tests arise when investigators intend to plan interim analyses during trial or to pursue multiple endpoints, objectives, etc. However, testing multiple hypotheses may increase the probability to erroneously reject true null hypothesis and result in a type I error rate beyond the pre-specified level α . In the past few decades, many statistical methods have been developed to address multiple hypothesis tests and to control the familywise error rate (FWER).

Group sequential (GS) design is one of the most frequently used approaches for multiple hypothesis tests induced by planned interim analyses (Jennison and Turnbull 2000). A group sequential design is often defined in terms of stopping boundaries regarding its decision of early stopping or continuation. Pocock (1977) sets constant stopping boundaries across all interim analyses and this typically leads to the final analysis having a critical value greater

than the fixed-sample counterpart (i.e., claiming significance at a level less than the common 0.05 threshold). O'Brien and Fleming (1979) developed a sequential design that scales the stopping boundaries by the proportion of information at each analysis time and results in a final boundary nearly identical to that of the fixed-sample design. Wang and Tsiatis (1987) unified the Pocock and O'Brien-Fleming boundaries by varying the user-specified tuning parameter one can move between the Pocock and the O'Brien-Fleming designs.

Joint modeling provides another attractive approach to address multiple hypotheses for trials with multiple endpoints. For example, Xiong and colleagues (2005) propose an intersection-union test (IUT) approach to assess treatment efficacy in the case of a clinical trial with two continuous primary endpoints. The proposed test is based on bivariate normal distribution with known variance–covariance matrix. Sozu *et al.* (2006) extend Xiong's power formula to the case of an unknown variance–covariance matrix using a Wishart distribution. Sozu *et al.* (2010) develop a formula of power and sample size calculation for clinical trials with multiple binary primary endpoints, while Sozu *et al.* (2012) extend the methods to trials with mixed continuous and binary endpoints. Most of these methods focus on the issue of multiple co-primary endpoints. That is, they assume all the endpoints are mutually exchangeable and an intervention is deemed efficacious only if it improves on all of the multiple primary endpoints.

Various approaches have also been proposed for hypothesis tests involving both primary and secondary endpoints. Moye (2000) introduces a prospective alpha allocation scheme (PAAS) to split the type I error across endpoints (both primary and secondary) in a trial. The proposed PAAS preserves the type I error of primary endpoint and determines sample size at a pre-specified alpha level based on the primary endpoint. It relaxes the overall family-wise error slightly to accommodate secondary endpoint(s), and makes it possible to claim a trial “positive” when the primary endpoint is not statistically significant but one or more secondary endpoints are significant. The adaptive alpha allocation approach (Li and Mehrotra 2008a, Li *et al.* 2013) take a similar strategy for multiple testing procedure except that the alpha to the secondary endpoints is allocated adaptively depending on the result of primary endpoint. Dmitrienko *et al.* (2003) propose a “gatekeeping” procedure for clinical trials with multiple endpoints and Bretz *et al.* (2009) develop a graphical approach for its implementation. Under this framework, the primary endpoint(s) will serve as gatekeeper in that the secondary endpoints are assessed only if the primary endpoint has been statistically significant. Li and Mehrotra (2008b) and Hamasaki *et al.* (2015) among others extend gatekeeping procedures to group-sequential design setting.

In the section follows, we propose a stepwise testing procedure to incorporate a continuous biomarker as a secondary endpoint into an immunotherapeutic neoadjuvant clinical trial. The hypothesis tests will be performed in a hierarchical manner and the secondary hypothesis will be tested only if the primary hypothesis is not rejected.

2.2. Stepwise testing procedure to incorporate immunogenicity biomarker

We assume that (X, Y) is a vector of two efficacy variables where X is considered the primary efficacy outcome which is binary data (i.e., with $X=1$ for success and $X=0$ for failure), and Y is considered a secondary efficacy outcome which is continuous and

measured only on individuals who failed the primary efficacy endpoint (i.e., those with $X=0$). Let (X^c, Y^c) and (X^t, Y^t) represent the specification of (X, Y) in the control group and the treatment group, respectively. Further we assume that, for $i = c, t$, the probability of success in the primary efficacy outcome is p^i , i.e., the probability of $X^i = 1$. Hence, $\Pr(X^i = 0) = 1 - p^i$.

Further, for $i = c, t$, and $k=0$ or 1 , conditioning on $X^i = k$, we assume that Y^i follows an unspecified distribution with mean μ_k^i and variance σ^2 (i.e., assuming common to both treatment groups).

For simplicity, we always assume that a larger mean in Y corresponds to a bigger treatment effect. Suppose that for $i = c, t$, X^i is observed in a sample of size n_i , resulting data $X_1^i, X_2^i, \dots, X_{n_i}^i$. For the secondary outcome Y , however, because it is designed to be observed on only those who fail in the primary outcome X , i.e., subjects with $X_j^i = 0$, we use $Y_{10}^i, \dots, Y_{N_{i0}}^i$ to denote the observed Y 's for treatment arm $i = c, t$, where N_{i0} 's are random variables, representing the number of subjects in the original sample who fail in the primary efficacy outcome.

Let $n = n_c + n_t$ be the total sample size, and λ denote the allocation ratio with $\lambda = \frac{n_t}{n_c}$, i.e., $\lambda=1$ for equal allocation ($n_t = n_c$) and $n_t = \lambda n_c$ for $\lambda \neq 1$.

We consider a decision process by declaring the efficacy of an active treatment in a sequential manner from the following decision rules:

Step 1: the active treatment is efficacious if it outperforms the control group by the primary efficacy;

Step 2: if the treatment fails to be declared as efficacious at Step 1, it can still be declared efficacious if it outperforms the control arm by the secondary endpoint among those who fail the primary endpoint if N_{i0} 's are larger than 0.

Statistically, this decision process requires two statistical tests to be tested in a sequential manner, $H_{01}: p^t = p^c; H_{a1}: p^t > p^c$, and then $H_{02}: \mu_0^t = \mu_0^c; H_{a2}: \mu_0^t > \mu_0^c$. The overall rejection region (RR) is defined as either $\hat{p}^t - \hat{p}^c > \delta_x$, or $\hat{\mu}_0^t - \hat{\mu}_0^c > \delta_y$ given $\hat{p}^t - \hat{p}^c \leq \delta_x$, where

$$\hat{p}^i = \frac{\sum_{j=1}^{n_i} X_j^i}{n_i}, \hat{\mu}_0^i = \frac{\sum_{j=1}^{N_{i0}} Y_{j0}^i}{N_{i0}}, \text{ and } Y_{j0}^i \text{ for those } Y \text{'s whose } X=0.$$

2.2.1 Type II Error and Power Function—The power function of the RR is given by, $f(p^t, p^c, \mu_0^t, \mu_0^c) = \Pr(\hat{p}^t - \hat{p}^c > \delta_x) + \Pr(\hat{p}^t - \hat{p}^c \leq \delta_x, N_{c0} > 0, N_{t0} > 0, \hat{\mu}_0^t - \hat{\mu}_0^c > \delta_y)$.

Type I error is the supreme of the power function when $p^t = p^c$ and $\mu_0^t = \mu_0^c$.

If both sample sizes n_i are large, $\hat{p}^t - \hat{p}^c$ follows a normal distribution with mean $p^t - p^c$ and standard deviation $\sqrt{p^t(1-p^t)/n_t + p^c(1-p^c)/n_c}$. Conditioning on N_{c0} and N_{t0} , and assuming N_{j0} 's are also large, with certain regularity conditions, $\hat{\mu}_0^t - \hat{\mu}_0^c$ follows another normal distribution with mean $\mu_0^t - \mu_0^c$ and variance $\sigma^2(1/N_{t0} + 1/N_{c0})$.

Let $N_0 = (N_{c0}, N_{t0})$ denote the set of all pairs of patients who do not achieve pCR, such that $N_{c0} > 0, N_{t0} > 0$ and $\hat{p}^t - \hat{p}^c \leq \delta_x$ (or equivalently $\lambda N_{c0} - N_{t0} \leq \lambda n_c \delta_x$). Hence, the power function for the proposed procedure is

$$\begin{aligned} f(p^t, p^c, \mu_0^t, \mu_0^c) &= \Pr(\hat{p}^t - \hat{p}^c > \delta_x) + \Pr(\hat{\mu}_0^t - \hat{\mu}_0^c > \delta_y, \lambda N_{c0} - N_{t0} \leq \lambda n_c \delta_x, N_{c0} \geq 0, N_{t0} \geq 0) \\ &= \Pr(\hat{p}^t - \hat{p}^c > \delta_x) + \sum_{(u,v) \in N_0} \Pr(\hat{\mu}_0^t - \hat{\mu}_0^c > \delta_y | (N_{c0}, N_{t0}) = (u, v)) \Pr((N_{c0}, N_{t0}) = (u, v)) \end{aligned} \quad (1)$$

The first term can be calculated as following when the sample sizes are large,

$$\Pr(\hat{p}^t - \hat{p}^c > \delta_x) = 1 - \Phi\left(\frac{\sqrt{\lambda n_c}(\delta_x - p^t + p^c)}{\sqrt{p^t(1-p^t) + \lambda p^c(1-p^c)}}\right), \quad (2)$$

Where Φ is the distribution function of standard normal distribution,

$\delta_x = z_{1-\alpha_x} \sqrt{\frac{(\lambda+1)p^c(1-p^c)}{\lambda n_c}}$, $z_{1-\alpha_x}$ is the 100(1- α_x) percentile of standard normal distribution, and α_x is determined as in the section follows.

For the second term given any pair of $(u, v) \in N_0$, we have

$$\begin{aligned} &\sum_{(u,v) \in N_0} \Pr(\hat{\mu}_0^t - \hat{\mu}_0^c > \delta_y | (N_{c0}, N_{t0}) = (u, v)) \bullet \Pr((N_{c0}, N_{t0}) = (u, v)) \\ &= \sum_{(u,v) \in N_0} \binom{n_c}{u} (1-p^c)^u (p^c)^{n_c-u} \binom{n_t}{v} (1-p^t)^v (p^t)^{n_t-v} \bullet \Pr(\hat{\mu}_0^t - \hat{\mu}_0^c > \delta_y | u, v) \end{aligned} \quad (3)$$

Since the probabilities of $\Pr(N_{c0} = 0, N_{t0} > 0)$, $\Pr(N_{c0} > 0, N_{t0} = 0)$, or $\Pr(N_{c0} = 0, N_{t0} = 0)$ are almost ignorable, the above formula could be approximated as,

$$(1 - \Pr(\hat{p}^t - \hat{p}^c > \delta_x)) \bullet (1 - \Phi\left(\frac{\delta_y - (\mu_0^t - \mu_0^c)}{\sigma\sqrt{1/m_c + 1/m_t}}\right)), \quad (4)$$

where $\delta_y = z_1 - \alpha_y \sigma\sqrt{1/m_c + 1/m_t}$, $m_c = n_c(1 - p^c)$ and $m_t = n_t(1 - p^t)$ represent the expected number of patients who will not achieve pCR in the control and experimental groups, respectively. In the special scenario where the data of secondary endpoint are available in all the participants (e.g., biomarkers from specimens of tumor biopsy during treatment or from peripheral blood samples), then the first term in the formula (1) remains unchanged, but the second term can be readily extended as,

$$(1 - \Pr(\hat{p}^t - \hat{p}^c > \delta_x)) \bullet (1 - \Phi\left(\frac{\delta_y - (\mu_0^t - \mu_0^c)}{\sigma\sqrt{1/n_c + 1/n_t}}\right)). \quad (5)$$

Hence the entire power function is readily computable.

2.2.2 Type I Error and Alpha-spending—Assuming a clinical trial is designed so that the primary efficacy endpoint at Step 1 will have a probability of type I error of α_x , then δ_x

can be decided such that $\alpha_x = \max[1 - \Phi\left(\frac{\sqrt{(1+\lambda)n_c}\delta_x}{\sqrt{\bar{p}(1-\bar{p})}}\right)]$, with $\bar{p} = (\lambda p^t + p^c)/(1 + \lambda)$. Assuming the clinical trial is designed so that the secondary efficacy endpoint at Step 2 will have type I error of α_y , then δ_y can be decided such that $\alpha_y = \max(1 - \Phi\left(\frac{\delta_y}{\sigma\sqrt{1/m_c + 1/m_t}}\right))$, with

$m_c = n_c(1 - p^c)$ and $m_t = n_t(1 - p^t)$ representing the expected number of patients who will not achieve pCR in the control and experimental groups, respectively.

Since hypothesis of the secondary endpoint will be tested only if the trial fails to reject the hypothesis of primary endpoint, the Type I error for the overall trial can be determined as, $\alpha = \alpha_x + (1 - \alpha_x) \bullet \alpha_y$. The choice of α_x and α_y can have significant effect on the overall power function of the proposed sequential testing. Analogous to O'Brien-Fleming (1979), we allocate majority of type I error (e.g., 80% of the total α) to the primary endpoint throughout this paper, but the impact of different alpha-allocation schemes on statistical power is also evaluated in the subsequent section.

3. An Illustration with a Real-world Example

We illustrate the sample size determination based on the CALGB 40603 which is one of the largest neoadjuvant trials in patients with TNBC. CALGB 40603 was a 2×2 factorial, open-label, randomized phase II trial, and evaluated the impact of adding carboplatin and/or

bevacizumab to neoadjuvant weekly paclitaxel followed by dose-dense doxorubicin and cyclophosphamide. The primary endpoint was pCR rate and the study assumed that one third of patients achieve pCR under standard of care neoadjuvant chemotherapy. The sample size ($N=105/\text{arm}$) was designed to detect 20% improvement in pCr (55% vs 35%) with 90% power at 1-sided $\alpha=0.05$ (Sikov et al 2015). The results showed that adding carboplatin to the standard of care significantly increased the pCR rate (54% vs 41%, $p=0.003$).

In the following subsections, suppose that an immunotherapeutic agent is added to the experimental arm. For the primary endpoint (binary variable), the effect size (δ_x) is defined as the difference of pCR rates between experimental and control arms. For the secondary endpoint (continuous variable), the effect size (δ_y) is defined as the difference of TIL means and scaled by the standard deviation (σ , assuming common to the two treatment groups). We assume an equal assignment of subjects in the two treatment groups ($\lambda=1$). Unless otherwise specified, the overall hypotheses are tested at 1-sided $\alpha=0.10$ throughout this paper, and we allocate $\alpha_x=0.08$ to the primary endpoint (e.g., indicating an $\alpha_y=0.0217$ for the secondary endpoint). The performance of the proposed stepwise testing procedure is first evaluated in terms of power and sample size gain, as comparing to the traditional approach that uses only the primary efficacy endpoint. We then assess the impact of different alpha-allocation schemes on the statistical power. A simple simulation study is also performed to compare analytical and empirical powers given a few scenarios with relatively small sample sizes.

3.1 The improvement of overall power upon incorporation of secondary endpoint

Figure 1 shows the statistical power as a function of the effect sizes in both primary endpoint (δ_x) and secondary endpoint (actually δ_y/σ), given a combination of different sample sizes (total $N=100$ or 200 patients) and different pCR rates in the control arm ($\pi_0=0.3$ or 0.4). The power surface reveals that in general the secondary endpoint helps to improve the overall power only if there exists a relatively strong effect size in the secondary endpoint. It also shows that the impact of secondary endpoint not only depends on the effect size of secondary endpoint itself, but also depends on the effect size of primary endpoint. When the study is slightly under powered for the primary endpoint, for example, adding secondary endpoint helps to improve the overall power to the targeted levels (80% or 90%). When the primary endpoint has a very small effect size, however, adding secondary endpoint will be unlikely to reach designed power unless the effect size of secondary endpoint is very large.

3.2 The reduction of sample size with incorporation of secondary endpoint

Suppose a new therapeutic neoadjuvant trial will be designed after CALGB 40603. Now that a pCR rate of 40% or even 50% can be achieved in the standard of care, it may be unrealistic to target a 20% or 30% improvement in the primary endpoint. Table 1 illustrates the reduction in the total sample size after incorporating the secondary endpoint. We do so by comparing the sample size from the traditional approach that uses only the primary efficacy endpoint with $\alpha=10\%$ to these from the proposed sequential testing (with $\alpha=0.08$ and 0.0217 for the primary and secondary endpoints, respectively). For an easy comparison, the effect size of secondary endpoint (δ_y/σ) is further standardized using following formula so that the power of the secondary endpoint is approximately 1-, 1.5-, and 3-time that of the primary endpoint (Sozu et al 2012),

$$\frac{\Delta_Y}{\sigma} = \frac{(z_{1-\alpha} + K \cdot z_{1-\beta}) \cdot \Delta_X}{z_{1-\alpha} \sqrt{(\pi_0 + \pi_1)(\theta_0 + \theta_1)} + z_{1-\beta} \sqrt{2 \cdot (\pi_0 \theta_0 + \pi_1 \theta_1)}}$$

where $Z_{1-\alpha}$ and $Z_{1-\beta}$ are the 100(1- α) and 100(1- β) percentiles of the standard normal distribution, and π_0 and π_1 are respectively the pCR rates for control and experimental arms, with $\theta_0=1-\pi_0$, $\theta_1=1-\pi_1$, and $K=1, 1.5$ and 3 , respectively.

Table 1 shows the minimum of total sample sizes required to achieve 80% and 90% power for the proposed stepwise testing procedure. The corresponding sample sizes from the traditional approach are also calculated using Chi-square test with a normal approximation. The sample size gain heavily depends on the effect size of secondary endpoint. For clinical trials designed to achieve 80% power, for example, the incorporation of secondary endpoint will not improve overall study power when the effect sizes of primary and secondary endpoint are similar. The impact of secondary endpoint is trivial when the effect size of the secondary endpoint is only 1.5 times of that in the primary endpoint. When the effect size of the secondary endpoint is 3 times of that in the primary endpoint, however, there is 15–20% reduction of total sample size after incorporating secondary endpoint. As a comparison, Table 1 also presents the potential sample size gain if the secondary endpoint can be measured in all patients. Under this scenario, adding secondary endpoint can lead to nearly 40% reduction of total sample size when the effect size of the secondary endpoint is 3 times of that in the primary endpoint.

3.3 The impact of alpha-allocation on statistical power

In all aforementioned power calculations, a majority of type I error is allocated to the primary endpoint. However, the choice of α_x and α_y may have significant impact on the power function of the proposed sequential testing. In this section, the impact of different alpha-allocation schemes on overall study power is evaluated in several scenarios given a combination of different sample sizes (total $N=100$ or 200 patients) and the control-arm pCR rates ($\pi_0=0.3$ or 0.4). In each study, the effect size of primary endpoint is chosen such that 80% power (at 1-sided $\alpha=0.10$) can be achieved under the traditional approach that uses only the primary endpoint. In other words, the study will be under powered to some extent for the primary endpoint in the proposed design because a smaller alpha (i.e., $\alpha_x=0.08$ rather than 0.10) will be allocated to the primary endpoint.

Figure 2 shows the overall study power as a function of the effect size in the secondary endpoint given different alpha-allocation schemes. As long as at least half of total alpha is assigned to the primary endpoint, the influence of alpha-allocation scheme is trivial and the incorporation of secondary endpoint will increase the overall power in the presence of even moderate γ . In contrast, when majority of alpha is assigned to the secondary endpoint, this will seriously reduce the power for primary endpoint under the proposed design and the overall power can be improved only in the presence of very strong γ .

3.4 Comparing analytical and empirical power

The performance of the proposed procedure is also assessed numerically using a simple simulation study. In this section, the analytical and empirical powers are compared in several studies given a combination of different sample sizes (total $N=50, 100, 150$ or 200 patients) and different pCR rates in the control arm ($\pi_0=0.3, 0.4$, or 0.5). In each study, the effect size of primary endpoint is chosen such that 80% power (at 1-sided $\alpha=0.10$) can be achieved under the traditional approach that uses only the primary endpoint.

Figure 3 shows the empirical and analytical powers for studies with the control arm pCR rate $\pi_0=0.4$, the primary endpoint effect sizes $\delta_x=0.294, 0.211, 0.172, 0.149$, and total sample sizes $N=50, 100, 150, 200$, respectively. In each simulation, the data of primary endpoint (pCR) in the control and treatment arms are generated from binomial distributions $B(N/2, \pi_0)$ and $B(N/2, \pi_0 + \delta_x)$, respectively. The data of secondary endpoints (TIL) in the control and treatment arms are generated from normal distributions $N(0, 1)$ and $N(\delta_x, 1)$, with the sample size of $N/2 \cdot (1 - \pi_0)$ and $N/2 \cdot (1 - \pi_0 - \delta_x)$, respectively. The between-group differences for the primary and secondary endpoints are compared by Chi-square test (with Yates' continuity correction) and two-sample Student t-test (assuming equal variance), respectively, and the secondary endpoint is tested (at 1-sided $\alpha=0.0217$) only if the primary endpoint is not significant (at 1-sided $\alpha=0.08$). The empirical power for each simulation is based on 10,000 datasets. The analytical and empirical powers agree very well in general, though the analytical power is slightly higher than empirical one when the sample size is very small. Similar patterns are observed for $\pi_0=0.3$ or $\pi_0=0.5$ and the results are not presented here.

The control of Type I error is also evaluated using simulated data. Figure 4 shows the empirical α for studies with a total sample size of $N=50, 100, 150, 200$, respectively. Data are generated and analyzed exactly the same way as in the above simulations except that $\delta_x = \delta_y = 0$ in all trials. The results reveal that type I error can be well controlled at the designed level $\alpha=0.10$ under a wide range of pCr rates.

4. Discussion

The statistical design of clinical trials is conducted traditionally based on a single outcome that is designated as the primary endpoint. A primary endpoint describes how the most important features of a disease are affected by treatment. However, many human diseases are characterized by multi-dimension impairment, including signs, symptoms, quantitative measurements, and patient-reported outcomes. Alzheimer's disease, for example, is characterized by poor cognition and disorderly behavior of daily living. Accordingly, the assessment of new intervention needs multiple measures for a comprehensive understanding on its therapeutic effects. One of magnificent examples is a vaccine study which often requires up to 10 primary endpoints (including incidence/severity of disease and immunogenicity biomarkers) to demonstrate efficacy to obtain marketing authorization (Offen et al 2007). Multiple primary endpoints have also been adopted to assess the benefits of new interventions in other complex diseases such as cancer, rheumatoid arthritis, chronic kidney disease, and cardiovascular disease.

In the present study, a stepwise testing procedure is proposed to deal with some specific challenges currently faced by immunotherapeutic neoadjuvant clinical trials for patients with TNBC (Stagg et al 2013; Thompson et al 2012). Namely, 1). the trials need to be conducted at a relatively small scale; 2). though pathological complete remission rate (pCr) remains the widely used standard endpoint for neoadjuvant trials, pCr alone may be inadequate for measuring antitumor efficacy by immunotherapeutic agents; and 3). the current standard of care can achieve very impressive pCr and thus leave limited room for further improvement in terms of pCr. Since there have been increasing data confirming the association between tumor-infiltrating lymphocyte (TIL) and survival outcomes in TNBC following adjuvant or neoadjuvant chemotherapy (Loi et al 2013; Denkert et al 2015), it is a natural choice to incorporate TIL as an additional metric for treatment efficacy. Operating characteristics of the proposed procedure are illustrated using simulated data and a real-world example. Results show that the incorporation of secondary endpoint will improve the overall study power only in the presence of relatively large effect size in the secondary endpoint. When the effect size of the secondary endpoint is 3 times that of the primary endpoint, for example, the incorporation of secondary endpoint can achieve 15–20% reduction of total sample size for a study with 80% power, as comparing to a traditional design that uses only the primary endpoint. As the goal of a neoadjuvant trial is to screen for promising treatment, we therefore expect that those new therapies with either strong effect on pCR alone or those with moderate pCR but very strong immunogenicity effect would be more likely to be selected by the proposed procedure for further evaluation.

The proposed procedure has some limitations. Its operating characteristics are only assessed under a few scenarios with limited range of specifications. The performance could also be influenced by other parameters such as the correlation between primary and secondary endpoints. Because TIL is only available in these patients who have not achieved pCR, it is difficult to have an empirical estimate regarding the strength of correlation between them. Consequently, our stepwise testing procedure is performed assuming independence, similar to that of Moye's (2000) prospective alpha allocation scheme (PAAS). Although we only consider a special scenario where the tumor biomarker is only partially available, the concept of proposed design can be readily extended to more general cases such as the availability of biomarkers from intermediate tumor biopsies or peripheral blood samples, or the combination of multiple biomarkers.

Acknowledgements

The authors would like to thank the Associate Editor and two anonymous referees for their insightful comments. This study is partially supported by grants from the *NCI Cancer Center Support Grant P30 CA091842* (PI: Eberlein), *Komen Foundation Grant KG111025* (PI: Gillanders/Mardis), *US Army Medical Research Acquisition Activity Grant BC140654* (PI: Gillanders), and *National Institute on Aging (NIA) Grant R01 AG053550* (PI: Xiong).

References

1. Berry DA, Cirincione C, Henderson IC, Citron ML, Budman DR, Goldstein LJ, Martino S, Perez EA, Muss HB, Norton L, Hudis C, Winer EP. Estrogen-receptor status and outcomes of modern chemotherapy for patients with node-positive breast cancer. *JAMA*. 2006; 295(14):1658–1667. [PubMed: 16609087]

2. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; 28:586–604. [PubMed: 19051220]
3. Denkert C, von Minckwitz G, Brase JC, Sinn BV, Gade S, Kronenwett R, Pfitzner BM, Salat C, Loi S, Schmitt WD et al.: Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers. *Journal of clinical oncology*. 2015, 33(9):983–991. [PubMed: 25534375]
4. Dmitrienko A, Offen WW, Westfall PH. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* 2003; 22:2387–2400. [PubMed: 12872297]
5. Dmitrienko A, and Tamhane AC. Gatekeeping Procedures with clinical trial applications. *Pharmaceutical Statistics*. 2007; 6: 171–180. [PubMed: 17583553]
6. Hamasaki T, Asakura K, Evans SR, Sugimoto T, and Sozu T. Group-Sequential Strategies in Clinical Trials with Multiple Co-Primary Outcomes. *Statistics in Biopharmaceutical Research*. 2015; 7(1): 36–54.
7. Jennison C, and Turnbull BW. Group sequential methods with applications to clinical trials. Chapman & Hall, Boca Raton, 2000.
8. Li J, and Mehrotra DV. An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine*. 2008a; 27:5377–5391. [PubMed: 18759248]
9. Li J, and Mehrotra DV. Gatekeeping testing via adaptive alpha allocation. *Biometrical Journal* 2008b; 50:704–715. [PubMed: 18932133]
10. Li H, Sankoh AJ, and D'Agostino RB. Extension of adaptive alpha allocation methods for strong control of the family-wise error rate. *Statistics in Medicine*. 2013; 32: 181–195 [PubMed: 22806685]
11. Liedtke C, Mazouni C, Hess KR, Andre F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M et al.: Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *Journal of clinical oncology*. 2008, 26(8):1275–1281. [PubMed: 18250347]
12. Lin DY, Shen L, Ying Z, Breslow NE. Group sequential designs for monitoring survival probabilities. *Biometrics*. 1996; 52(3):1033–41. [PubMed: 8805766]
13. Loi S, Sirtaine N, Piette F, Salgado R, Viale G, Van Eenoo F, Rouas G, Francis P, Crown JP, Hitre E et al.: Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase III randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: BIG 02–98. *Journal of clinical oncology*. 2013, 31(7):860–867. [PubMed: 23341518]
14. Moye LA. Alpha calculus in clinical trials: considerations and commentary for the new millennium. *Statistics in Medicine*. 2000; 19:767–779. [PubMed: 10734280]
15. O'Brien PC, and Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35(3):549–556. [PubMed: 497341]
16. Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, Muirhead R, Stryszak P, Boddy A, Chen K, Copley-Merriman K, Dere W, Givens S, Hall D, Henry D, Jackson JD, Krishen A, Liu T, Ryder S, Sankoh AJ, Wang J, and Yeh CH. Multiple co-primary endpoints: medical and statistical solutions. *Drug Information Journal*. 2007;41, 31–46.
17. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977; 64(2):191–199.
18. Sikov WM, Berry DA, Perou CM, Singh B, Cirincione CT, Tolaney SM, Kuzma CS, Pluard TJ, Somlo G, Port ER et al.: Impact of the addition of carboplatin and/or bevacizumab to neoadjuvant once-per-week paclitaxel followed by dose-dense doxorubicin and cyclophosphamide on pathologic complete response rates in stage II to III triple-negative breast cancer: CALGB 40603 (Alliance). *Journal of clinical oncology*. 2015, 33(1):13–21. [PubMed: 25092775]
19. Sozu T, Kanou T, Hamada C, and Yoshimura I. Power and sample size calculations in clinical trials with multiple primary variables. *Japanese Journal of Biometrics*. 2006; 27, 83–96.
20. Sozu T, Sugimoto T, and Hamasaki T. Sample size determination in clinical trials with multiple coprimary binary endpoints. *Statistics in Medicine*. 2010; 29, 2169–2179. [PubMed: 20687162]

21. Sozu T, Sugimoto T, and Hamasaki T. Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometrical Journal*. 2012; 54(5): 716–729. [PubMed: 22829198]
22. Scharfstein DO, Tsiatis AA, and Robins JM. Design and analysis of group sequential studies. *Journal of the American Statistical Association*. 1997; 92: 1342–1350.
23. Stagg J, and Allard B. Immunotherapeutic approaches in triple-negative breast cancer: latest research and clinical prospects. *Therapeutic Advances in Medical Oncology*. 2013; 5(3): 169–181 [PubMed: 23634195]
24. Thompson AM, and Moulder-Thompson SL. Neoadjuvant treatment of breast cancer. *Annals of Oncology*. 2012; 23: x231–x236 [PubMed: 22987968]
25. Wang SK, and Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*. 1987; 43(1): 193–199. [PubMed: 3567304]
26. Xiong C, Yu K, Gao F, Yan Y, and Zhang Z. Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to an Alzheimer's treatment trial. *Clinical Trials*. 2005; 2, 387–393. [PubMed: 16317808]

Figure 1: the overall power (values ranged 0 – 1) as a function of the effect size in primary endpoint (δ_x , values ranged from 0 – 0.4) and the effect size in secondary endpoint (actually δ_y/σ , values ranged from 0 – 0.6). The power is calculated given a combination of different sample sizes (total $N=100$ or 200 patients) and different pCR rates in the control arm ($\pi_0=0.3$ or 0.4).

Figure 2:

the overall power as a function of the effect size in the secondary endpoint (γ) given different alpha-allocation schemes. The overall power is calculated given a combination of different sample sizes (total N=100 or 200 patients) and the control-arm pCR rates ($\pi_0=0.3$ or 0.4), where the effect size of primary endpoint (δ) in each study is chosen to achieve 80% power at 1-sided $\alpha=0.10$ under the traditional approach.

Figure 3:

the empirical power from 10000 simulated trials and each trial with a total sample size of N=50, 100, 150, and 200 patients, respectively. The pCR rate in the control arm is 0.4 in each study and the effect size of primary endpoint (Δ_x) is chosen to achieve 80% power at 1-sided $\alpha=0.10$ under the traditional approach.

Figure 4:
the empirical alpha from 10000 simulated trials and each trial with a total sample size of N=50, 100, 150, and 200 patients, respectively. Data in all trials are simulated with both the primary and secondary effect sizes $\alpha_x = \alpha_y = 0$.

Table 1:

Percentages of sample size reduction due to the incorporation of secondary outcome, where the effects (γ) of secondary endpoint are standardized respectively as 1-, 1.5-, and 3-time of the corresponding effect size (χ) in the primary endpoint.

Scenarios	π_0	χ	γ	80% Power			90% Power		
				N_{pCR}	N_{new}	% [*]	N_{pCR}	N_{new}	% [*]
Scenario A: secondary endpoint partially available (e.g., TIL only measured in non-pCR cases)	0.4	0.10	0.123	438	444	-1.4	638	640	-0.3
			0.154		424	3.2		612	4.1
			0.247		348	20.5		496	22.3
		0.15	0.185	194	198	-2.1	284	286	-0.7
			0.231		190	2.1		274	3.5
			0.370		158	18.6		226	20.4
	0.5	0.10	0.123	448	462	-3.1	652	664	-1.8
			0.154		444	0.9		638	2.1
			0.247		376	16.1		538	17.5
		0.15	0.189	198	204	-3.0	286	292	-2.1
Scenario B: secondary endpoint fully available (e.g., TIL measured in all cases)	0.4	0.10	0.123	438	414	5.5	638	598	6.3
			0.154		380	13.2		546	14.4
			0.247		272	37.9		384	39.8
		0.15	0.185	194	184	5.2	284	266	6.3
			0.231		170	12.4		244	14.1
			0.370		122	37.1		172	39.4
	0.5	0.10	0.123	448	424	5.4	652	610	6.4
			0.154		390	12.9		556	14.7
			0.247		276	38.4		390	40.2
		0.15	0.189	198	186	6.1	286	266	7.0
			0.236		170	14.1		244	14.7
			0.378		120	39.4		170	40.6

π_0 : pCR rate in the control arm; χ : effect size in the primary endpoint

γ : standardized effect size in the secondary endpoint

N_{pCR} : total sample sizes based on pCR alone

N_{new} : total sample sizes based on the proposed procedure (pCR+TIL)

* : % of sample size reduction = $(N_{pCR} - N_{new}) / N_{pCR} * 100$