January 2013

Prevalence, Sample Size, And Power Investigation In Neoadjuvant Therapy For Breast Cancer Clinical Trials

Jennifer So
Yale University, email.jso@gmail.com

Prevalence, Sample Size, and Power Investigation in Neoadjuvant Therapy for Breast Cancer Clinical Trials

By

Jennifer So

A Thesis Presented to

The Faculty of Yale School of Public Health

In Candidacy for the Degree of

Master of Public Health in Biostatistics

May 2013

ACKNOWLEDGMENTS

I would like to thank my faculty advisor and thesis readers, Dr. Peter Peduzzi and Dr. Haiqun Lin for the patience and support throughout the year leading to the completion of the thesis requirement for the Master of Public Health at the Yale School of Public Health. I have gained an immense amount of knowledge in insight to biostatistics and clinical trials from coursework and my experiences with professors. I am truly grateful to have had an opportunity to work with esteemed biostatisticians. I would also like to thank Dr. Lajos Pusztai, who is a Professor of Medicine and the Director of the Breast Medical Oncology Section, for the proposition of the study and support in the clinical aspect of the study.

ABSTRACT

Common practice in neoadjuvant therapy clinical trials for breast cancer that use pathological complete response rate (pCR) as an endpoint are conducted within a Human Epidermal Growth Factor Receptor 2 (HER2) indication group for both Estrogen Receptor – positive (ER+) and Estrogen Receptor – negative (ER–) cancers. Given the clinical background and trends of breast cancer therapy trials, this study aims to demonstrate in cases where the observed prevalence and response rates may be so different from the expected values that priori sample size calculation and power analyses were based on, that dangers may arise in producing an insufficiently powered study with unreliable results. Critiques of the widespread practice of underpowered clinical trials are long-standing and such related ethical issues have been substantially debated in biostatistics and medicine. However, an overwhelming prevalence of underpowered studies even recently and studies failing to do a priori sample size and power calculation is still found.

This study uses simple statistical methods to show the effects of not accounting for proportional differences and also detect power differences in pCR rates between two arms in a randomized study. To demonstrate how the overall pCR rate can change for the same effect size in a particular HER2 group study based on changing the proportion of ER+ and ER– patients, pCR rates are calculated over a series of hypothetical studies with varying proportions of cases. The power needed to detect an absolute difference in pCR rates between the two arms could vary greatly depending on the actual trial accrual by ER status.

If a study is designed with specified prevalence rates for ER+ and ER- groups and the observed pCR rates are different than hypothesized, this situation could result in an underpowered study.

TABLE OF CONTENTS

INTRODUCTION

The burden of cancer in general is heavy in terms of mortality, incidence, and treatment. Among cancer cases in the United States, breast cancer is the leading cause of cancer among women and account for 29% of cancer cases and 14% of cancer deaths annually. Since 1975, breast cancer incidence has been fluctuating and increasing and mortality has steadily and shallowly declined. A reasonable guess in describing these statistics is that breast cancer has been extensively funded for more research to be done since the realization of its impact on public health (Toriola & Colditz, 2013). Treatment of breast cancer is highly dependent on post-menopausal hormones and other factors that depend on personal cases.

Although the factors involved in breast cancer identification and targeted therapy are many and complicated, this study is interested in the population with known human epidermal growth factor receptor 2 (HER2) and estrogen receptor (ER) status. Some are sensitive to the hormone estrogen, which stimulates the tumor to grow. The cancers that have estrogen receptors on the surface of their cells are called estrogen receptor–positive (ER+) cancers, and those without detected receptors are estrogen receptor–negative (ER–). Another factor of breast cancer involves the human epidermal growth factor receptor 2 (HER2) that generally helps cells, grow, divide, and repair themselves. Women with HER2–positive cancer (HER2+) were found to have higher risk of recurrence due to the fast growth of cancer cells (Gonzalez-Angulo et al., 2009). Many recent studies group ER-, HER2-, and progesterone receptor (PR)-negative together into a so called “triple-negative” phenotype. Tumor markers are increasingly important in breast cancer research and studies thus far have found patients with such a diagnosis to react similarly and effectively to certain treatments for their basal-like characteristics (Bauer, Brown, Cress, Parise,

& Caggiano, 2007). Despite the additional PR marker discussed that has been used in recent clinical trials, many trials focus on ER and HER2 statuses to determine the risk of recurrence or predict the success of treatment (Liedtke et al., 2008; Tischkowitz etal.,2007). ER status is a significant prognostic factor for predicting treatment effects because it can identify patients who may benefit from endocrine therapy (Berry et al., 2006). The results from early stage breast cancer clinical trials for endocrine therapy strongly have long suggested that this type of neoadjuvant therapy such as "letrozole", may benefit patients with ER+ disease disproportionately to ER– disease (Ellis et al., 2001). In addition, patients who receive tamoxifen, which was initially tested as an adjuvant therapy, have found to be most receptive to those with ER+ and high expression of HER2 (Shou et al., 2004).

Neoadjuvant chemotherapy, which is treatment given prior to the surgical procedure introduced early by Rosen in 1982, is performed in patients with early breast cancer when an indication for chemotherapy is given by a physician. It aims to reduce the burden of the tumor prior to a procedure, and also allows the option of lumpectomy to be available (Rosen et al., 1982). These benefits of neoadjuvant chemotherapy are supported by current research, which suggest firmly establishing it as an option for women with breast cancer. Though it is not necessary for much of this study to be specific to neoadjuvant chemotherapy studies, it is critical to understand the benefits of having the ability to observe how the body reacts to certain drugs before operation and perform presurgical genetic testing to locate the markers, and design such a study in a statistically accurate manner.

Pathologic complete response (pCR) is the endpoint of choice for this study; it is defined as the absence of any residual invasive cancer on hematoxylin and eosin evaluation of the resected breast specimen and all sampled ipsilateral lymph nodes following completion of

neoadjuvant systemic therapy, which is coded ypT0 ypN0 in the current AJCC staging system (Prowell & Pazdur, 2012). Pathologic complete response has been used as an endpoint in numerous trials of neoadjuvant systemic therapy for breast cancer. Other popular methods of analysis in breast cancer clinical trials include disease free survival and overall survival. A uniform definition of pCR has not been determined in conducting current breast cancer studies even throughout its use as an endpoint. For example, some investigators have defined pCR as the absence of both in situ and invasive cancer following neoadjuvant chemotherapy, whereas others have considered only the invasive component in the definition. Other definitions of pCR include the absence of residual cancer in the breast and regional lymph nodes at the time of definitive surgery, and as a complete response in the breast, irrespective of axillary nodal involvement (Buzdar et al. 2005; von Minckwitz et al. 2010; Bear et al. 2006; Wolmark et al. 2001). Furthermore, pathology outcomes in neoadjuvant trials have been termed not only pCR, but near pCR, quasi pCR, comprehensive pCR, strict pCR, and pCRinv (Kuroi, Toi, Tsuda, Kurosumi, & Akiyama, 2006). Such variation in the definition of pCR has made clinical interpretation of data from neoadjuvant trials challenging. Though there is need for an adoption of a single term with a standard definition for future proposed trials, a review of more current studies within that past ten years have argued for use of strictly complete and comprehensive pCR, which has been labeled type ypT0 ypN0 (von Minckwitz et al., 2012).

The effectiveness of adjuvant therapy for breast cancer is well–established, but certain subpopulations of breast cancer patients continue to be at risk for recurrence and death, even with the best adjuvant therapy. Because even novel postoperative systemic therapies can only be assessed in multiyear trials, it is difficult to assess the potential effectiveness early. However, the potential clinical benefits of preoperative systemic therapy can be assessed and predicted early

using a pCR endpoint (Faneyte et al., 2003). Based on previous studies that use pCR as an endpoint, those patients who receive pCR are part of a nonrandomized patient subset determined by outcome subsequent to randomization. It has been expected that a large difference in pCR rate between treatment arms will be needed to produce a statistically significant difference in Disease Free Survival or Overall Survival for further analyses to show clinical significance (Donahue et al., 2009; Untch et al., 2011).

Ongoing breast cancer trials have explored the effect of certain treatments on various subsets of possible groups using pCR as an endpoint. For example, Trastuzumab (Herceptin) has been found to be the most effective treatment in combination with neoadjuvant chemotherapy for HER2-positive cancers from comparing this group with HER2-negative cancers along with other subsets of identification. Though many of the treatment recommendations have come from similar results from repeated trials, many studies are underpowered due to changes in actual values from what was expected. Because sample size, power, and effect size calculations depend heavily on previous assumptions, it is important to consider how much power a study can lose from deviating enough from expectations.

In addition, some previous randomized adjuvant chemotherapy clinical trials have shown that ER+ cancers were found to be more sensitive to chemotherapy and to have higher pathologic complete response rates than ER- cancers. Consequently, a combined analysis of several studies that examined chemotherapy hazard and recurrence reduction among patients with ER–positive and ER–negative cancers showed that the ER–negative group performed much better in these trials for similarly targeted treatments (Pusztai et al., 2008).

Given the clinical background and trends of breast cancer treatment trials, this study aims to provide examples for cases in clinical trials where the observed prevalence and response rates

may be so different from the expected values that priori sample size calculation and power analyses were based on, that dangers may arise in producing an insufficiently powered study with unreliable results. This study uses simple statistical methods to show how changes in certain components may affect the power of the study for tests for treatment effect and group comparisons.

MOTIVATION FOR STUDY

Trials must have sufficient statistical power to detect differences of clinical interest. However, critiques of the widespread practice of underpowered clinical trials are long-standing and the ethical issues associated with such situations have been substantially debated in medical journals alone (Halpern, Karlawish, & Berlin, 2002; Newell, 1978). However, an overwhelming prevalence of underpowered studies even recently and studies failing to do a priori sample size and power calculation is still found (Vogel et al., 2006; Coombes et al., 2004). Priori power analysis is essential in the planning of clinical trials because it determines the chance of detecting a true-positive result. Continuing a study with insufficient sample size or power will not produce reliable results, rendering it futile and undermining its clinical value.

A priori sample size calculation and power analysis, using standard formula calculations and statistical software, can determine the sample size required to get a significant result with adequate power, characterize the power of a study to detect a meaningful effect, and conduct sensitivity analyses of power or required sample size to other factures such as how the prevalence rate affects the power for the particular interests of this study (SAS Institute Inc ., 2004).

STATISTICAL BACKGROUND

For binary endpoints such as the pCR rate, the required sample size depends on the desired level of significance and power, clinically relevant difference or effect size, and the overall event rate or proportions. As a consequence of the overall event rate varying considerably between studies, determination of this parameter is significant in obtaining the necessary statistical significance and power to have meaningful study results (Friede & Kieser, 2004).

Because we are studying counts of success from pathological complete rates, we are concerned with the proportion of times that an event occurs rather than the number of times. We are interested in whether the discrepancy between proportions in each group that can be seen without statistical analysis is due to chance alone. Counts and proportions follow a binomial distribution, which provides the foundation for the analysis of proportions. Because with a large enough sample, the binomial distribution increasingly resembles that of a normal distribution, we are able to calculate standard normal probabilities for subsequent sample size calculation and power analyses (Pagano, Gauvreau, & Pagano, 2000).

In addition, we are able to make statistical inference about the value of the population proportion from the central limit theorem that shows that the mean of the sampling distribution is the population mean p and the standard error is $\sqrt{p(1-p)/n}$. Such inference contributes to the derivation of the sample size and power formula which is a function of the standard error of the sample size for treatment effect and difference of two proportions.

There are multiple hypotheses that can be tested using the attributes of statistics that simple sample size and power formulas have been derived from. The two hypotheses tested in this study is one of testing equivalency of a treatment and control groups and another of testing equivalency in proportions. Because the variances are a function of the sample size, the

calculated value simultaneously satisfies the equality P ($Z > Z_\alpha$) = α if the null hypothesis is true and P ($Z > Z_\alpha$) = 1-β if the alternative hypothesis is true (Lachin, 1981).

The first approach we use is to determine the sample size in terms of risk or treatment difference between an experimental group and a control group (Donner, 1984). The null hypothesis of no difference between the treatment group rate and the control group rate being zero is compared to the alternative hypothesis of the difference not being zero.

$$H_o : \mathrm{p_T} - \mathrm{p_C} = 0$$

$$H_a : \mathrm{p_T} - \mathrm{p_C} \neq 0$$

The z statistic that is used for comparing such responses is:

$$Z = \frac{\left(\mathrm{p_T} - \mathrm{p_C}\right)}{\sqrt{\bar{\mathrm{p}}\,(1-\bar{\mathrm{p}})\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}} \qquad (1)$$

Z	Test statistic for hypothesis test with mean 0 and variance 1
n_T	Number of participants in treatment group
n_C	Number of participants in control group
p_T	pCR rate for treatment group
p_C	pCR rate for control group
$\bar{p}$	Estimated average of the pCR rates $\left(\bar{p} = \frac{p_T + p_C}{2}\right)$

Sample Size:

When designing a study, investigators must determine a sample size that will be necessary to provide a specified power of a test of hypothesis, which is the probability that we will reject the null hypothesis given that it is false. For example, with equal sized groups and a significance level set at α = 0.05, we calculate the sample size needed to maintain the power of

the test at .8. This is assuming that we are willing to risk a 20% chance of failing to reject the null hypothesis.

In order to calculate sample size, the estimated parameters required are z-value corresponding to the desired type I error (α) for a two-sided test, z-value corresponding to the desired type II error (β), rate for experimental treatment group, rate for control group, and the estimated average of the two weights. The sample size formula is a derivation of the Z statistic formula, which is algebraically equivalent to the chi-square statistic that may be used, as well (Friedman, Furberg, & DeMets, 2010). The conventional α and β values are 0.05 and 0.2, respectively, but may vary closely and still be acceptable. Statisticians have used these values as a threshold, but sometimes may not be as conservative as to rejecting based on a rigid margin and accept α values even up to 0.1. Such decisions depend solely on the primary investigator and the statistician that considers how conservative they must be to generate reliable results.

Given a case where two populations are stratified further into two groups and information regarding the prevalence and response rate is given for each group, the control rate for that particular population is calculated by summing the product of the prevalence and the response rate for one group.

$$rate_{control} = (prevalence_1 * rate_1) + (prevalence_2 * rate_2) \qquad (2)$$

Sample size formula, assuming equal sized treatment groups:

$$N = \frac{4(Z_{\alpha/2} + Z_{\beta})^2 \, \bar{p} \, (1 - \bar{p})}{|p_T - p_C|^2} \qquad (3)$$

N	Total sample size for HER2 group
$z_{a/2}$	z-value corresponding to the type I error (α) for a two-sided test
z_{β}	z-value corresponding to the type II error (β)
p_T	pCR rate for experimental treatment group

p_C pCR rate for control group

$\bar{p}$ Estimated average of the pCR rates $\left(\bar{p} = \frac{p_T + p_C}{2}\right)$

where z_β is 0.84 for 80% power and $z_{a/2}$ is 1.96 for α=0.05, both of which can be looked up in a z-table.

The second approach is to use the sample size that is calculated for the given parameter estimates and observe how varied prevalence rates in each of the stratified groups changes the power of the study. This is a case where certain prevalence and response rates were expected from previous findings of small studies, but the real rates that were found in the actual trial were different from expected. Having recruited based on the sample size calculation made before the study commenced, a difference in the prevalence rates per group will affect the overall response rate for the control group and therefore the power. Not only prevalence rates, but also pCR rates can be found to be different than expected. However, the difference may be a negligible component that can be observed by simulation.

Power:

Formula for power derived from the sample size formula:

$$z_\beta = \frac{-z_{\frac{a}{2}}\sqrt{2\bar{p}(1-\bar{p})} + \sqrt{n}\,|p_T - p_C|}{\sqrt{p_T(1-p_T) + p_C(1-p_C)}} \qquad (4)$$

n Sample size per group ($\frac{N}{2}$)

$z_{a/2}$ z-value corresponding to the type I error (α) for a two-sided test

z_β z-value corresponding to the type II error (β)

p_T pCR rate for experimental treatment group

p_C pCR rate for control group

$\bar{p}$ Estimated average of the pCR rates $\left(\bar{p} = \frac{p_T + p_C}{2}\right)$

METHODS

This study aims to address and explore mainly two questions:

Question 1: How does the complete response rate vary according to the prevalence of each group?

Question 2: How does the prevalence rate affect the power of a study?

More recently, much neoadjuvant chemotherapy for breast cancer trials have been focused on investigating whether a particular treatment in addition to the chemotherapy leads to better outcomes than with chemotherapy only. In addition, studies also attempt to compare how effective the treatment in addition to chemotherapy is to one subset of the breast cancer population to another in order to be able to provide targeted and personalized treatments for patients as variable as those with breast cancer.

The data presented by a clinician and verified by several studies is shown in Table 1 and is consistently referred to throughout. According to clinical practice and judgment, HER2- and HER2+ groups should be studied separately (Untch et al., 2010; von Minckwitz et al., 2008; Iwata et al., 2011). Therefore, when we vary parameters to observe effects on a certain component of interest regarding power and prevalence, we do so for either HER2 group. The same methodology is used for both groups.

Prevalence estimates and pathologic complete response rates are provided for each ER group within an HER2 group. Overall, the prevalence of HER2- is 80% and that of HER2+ is 20%. The prevalence of breast cancer patients with HER2- and ER+ markers is estimated to be 70%; and, the prevalence for those with HER2- and ER- markers is estimated to be 30%, which is 1 – (HER2-, ER+ Prevalence). The corresponding pCR rates are 5% and 15% for HER2-,ER+ and HER2-,ER-, respectively. Among those with the HER2+ marker, 60% are ER+ with a pCR rate of 25%, and 40% are ER- with pCR rate 45%.

In the following sample size calculations and power analyses, the ER+ prevalence explored will range from 0.1 to 0.9 by 0.05. In addition, we use hypothetical varied treatment effects on pCR rates, which are 10%, 30%, 50%, and 70%. These are not the effect sizes, but are 10%, 30%, 50%, and 70% increases from the control pCR rate. The effect sizes depend on the prevalence that is varied for these four cases.

Question 1:

First, the rate for the control group must be estimated from known information about the prevalence and rates of the group from previous studies. The endpoint in determining sample size and power is dependent on the prevalence of each group, where total pCR= (prevalence ER+)(pCR rate for ER+) + (prevalence ER-)(pCR rate for ER-). Based on the values presented in Table 1 for the HER2- group, the total pCR is (0.70)(.05) + (0.30)(.15) = .08, which is 8%. We are interested in varying the prevalence in order to see how the pCR rate differs from this value. In order to observe the trend of total pCR as prevalence of ER+ is increased, we calculate the total pCR with fixed pCR rates at 5% and 15% and plot against the prevalence. For further insight, we can also overlay plots of different pCR rates, which are hypothetical deviations in actuality from expected rates. The values chosen for investigatory pCR rates are arbitrary positive and negative deviations from the actual rates.

Question 2:

The Z value for beta formula that is derived from the sample size formula is used to calculate the power given all fixed parameter estimates. In order to investigate the effects of varied prevalence rate on the power of a study given the fixed pCR rates, we calculate a sample size to detect an effect between two treatment groups under a variety of scenarios. One scenario is to assume that the pCR rate for one group is based on Table 1, and assume that the new

treatment increases the pCR rate by a certain percentage. An initial example based on the values presented by Table 1, with a total pCR for HER2- of 0.08 and assumed rate increase by 50% to 0.12 is used for later comparisons. We calculate the sample size needed in order to detect this difference with a type I error of 5% for a two-sided test and 80% power. This procedure can be repeated for varying each parameter and for each HER2 group.

RESULTS

As the α, β, and difference in response values are varied, the resulting effect on the magnitude of the sample size can be seen. When holding α and β constant, as the difference in response increases, the resulting sample size must increase in order to guarantee a high probability of detecting the real difference. If a particular trial is able to obtain a sample size much greater than what is necessary to detect an effect that is expected, then the study can be better powered and more accurate. However, most studies are not able to recruit enough participants to detect the expected difference and result in an underpowered study. In addition, the prevalence of participants recruited for a particular group may not be the expected prevalence and have a significant impact on the power. It is important to consider reasonable estimates for the parameters given previous clinical findings and acceptable significance and power constraints.

Question 1:

Upon exploration of how the complete response rates vary according to the prevalence of each group, we observe HER2+ rates drop more sharply as prevalence of ER+ is increased than HER2- rates. Based on the values presented in Table 1 for the HER2- group, the total pCR is 8%. The same formula is used in order to calculate the total pCR of the HER+ case, which consists of different parameters. The two HER2 cases are independent of each other, and rely on different assumed pCR rates. The slope of the HER2+ group is sharper with equal increment increases of ER+ prevalence because the given pCR rate difference is larger; pCR rates are 25% for ER+ and 45% for ER- in the HER2+ group, as opposed to 5% and 15%, respectively, in the HER2- group.

Figure 1 shows the plot of varied prevalence from 0.1 to 0.9 by 0.05 for ER+ and the corresponding total pCR rates, given pCR rates for ER+ and ER- and that the prevalence rate of ER- is 1-prevalence of ER+. The pCR rates used to calculate the data points in Figure 1 are as given by reference in Table 1. The total pCR axis was tailored to the minimum and maximum values of the output for better viewing of data points. In the case of HER- where pCR rate of ER+ is 0.05 and ER- is 0.15, the total pCR rate is maximum at 0.145 when the prevalence of ER+ is 0.05 and minimum at 0.055 when the prevalence is 0.95. Similarly, the maximum total pCR rate for HER2+ is 0.44 when ER+ prevalence is 0.05 and minimum is 0.26 with ER+ prevalence at 0.95.

In addition, we also illustrate the curves for observed changes in pCR rates for the ER group. Because the trend of HER2+ and HER2- are similar, the inference of varying pCR rates in the case of HER2+ is analogical to that of HER2-. Figure 2 elucidates the idea that the range of the total pCR, when varied across ER+ prevalence, is restricted to the values of the pCR rates. If both the rates are increased or decreased by the same amount, the plotted points would shift up and down with the same slope. If the difference in the pCR rates change, the slope changes and the prevalence change has a lesser effect on the total pCR rate, which is used as the control rate for hypothesis testing and sample size calculation.

Question 2:

We use the total pCR rate, calculated with formula 2 and previously explored with varied ER+ prevalence, as a the control rate then to observe how the change in prevalence rate affects the power of a study. This is the case where the actual observed prevalence rate once a study is completed is different from what the previous sample size calculation was based on. We simulate

such a situation by calculating a sample size to detect an effect between the treatment and control groups initially under the scenario given by Table 1.

The first example with a total pCR for HER2- of 0.08 and assumed rate increase by 50% to 0.12 resulted in a sample size of 1764. 882 participants are required in each treatment arm in order to detect a 4% difference between the experimental treatment and control group, assuming that the control pCR rate of the control group is 8%.

Using the same method to calculate sample size for the HER2+ case with the values presented in Table 1, the control pCR is 33% and the corresponding sample size to detect a 50% increase in pCR rate from treatment effect to 49.5% is 280 with 140 participants in each group. This calculation assumes that the prevalence of ER+ is 60% with a pCR rate of 25%, and the prevalence of ER- is 40% with a pCR rate of 45%.

Given these sample sizes, we recalculated the pCR rate for when the prevalence of the ER+ and ER- groups are different. For if the ER+ prevalence is 0.8 instead of 0.7 given the same known pCR rates for the HER2-, then the control pCR rate would be 0.07. A 50% increase in pCR rate for the treatment group to 10.5% yields for the difference between groups to be 0.035. Then, we determined the power necessary to detect the 0.035 increase in pCR rate for the treatment group given the previously calculated sample size of 1764 and new $\bar{p}$ of 0.0875.

Applying the formula for investigatory values yields a Z_β value of 0.6424 which corresponds to about 96% power for a study within the HER2- group. The same is method is used for a study pertaining to the HER2+ group, given a sample size of 280 and pCR rates of 25% and 45% for ER+ and ER-, respectively. Table 3 displays the resulting values of Z_β per prevalence of ER+ and the corresponding power for both HER2+ and HER2- groups. For these

particular cases, the power can range from 66.73% to 97.24% for HER2+ and 66.75% to 95.42% for HER2-.

In addition, we found that no matter how extreme of a treatment effect we observe, the power of the study does not significantly change distinct from the varied prevalence. Figure 3 shows the power plots for varied ER+ Prevalence for 10%, 30%, 50%, and 70% increase in treatment pCR rate for both HER2+ and HER2-. For a treatment only expected to increase the pCR rate, the power is sufficient given prevalence rates similar to what was expected regardless of how drastic of an increase is expected in the pCR rate. Overall, however, we can make inferences based on all of the analyses that there exists a danger of having underpowered studies with unexpected prevalence rates of ER+ and ER- within an HER2 group.

DISCUSSION

The analyses in this study depended on roughly estimated response rates and prevalence. Numerous study designs exist to experiment and test the effect of neoadjuvant chemotherapy treatment on any possible subset of the breast cancer population. Due to the inability to obtain accurate and consistent values to present more impactful results, this study demonstrates trends for possible situations and provides guidelines for considerations in future sample size calculation and power analysis for similar studies.

Uncertainty in the power and sample size estimates arise because the ER+ and ER- breast cancers have very different sensitivities to chemotherapy. In addition, there are many other factors that can be considered in determining an inclusion criteria that it is difficult to know which subset of the population the treatment should target. Differences in patient composition have substantial impact on what the overall pCR rate is, and a comprehensive literature review of any particular subset of the population was difficult to obtain. In addition, pCR rates that were used as the base were provided by a consulting clinician and verified in the literature subjectively. Similar rates were accepted to use those values as the control example, even though the exact procedure and study design were not the same as any proposed here.

The major advantage to this study is the scope of its applicability. It generalizes several cases for clinical investigators to be able to apply their own study into such a framework towards the goal of maintaining enough power for their particular study. It also poses the hypothetical situations from changing parameters from control calculation with values that should be close to the unknown actual values. This study was aimed to emphasize an important concept of the

dangers of resulting in an underpowered study due to lack of priori power analysis and sample size calculation or observed values that deviate too far from the expected parameters.

Investigators should understand the concepts and relationships between parameters of sample size, power, and probability that were presented with statistical background in planning future breast cancer clinical trials.

REFERENCES

Bauer, K. R., Brown, M., Cress, R. D., Parise, C. A., & Caggiano, V. (2007). Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype. Cancer, 109(9), 1721-1728.

Bear, H. D., Anderson, S., Smith, R. E., Geyer, C. E., Mamounas, E. P., Fisher, B., ... & Wolmark, N. (2006). Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer: National Surgical Adjuvant Breast and Bowel Project Protocol B-27. Journal of Clinical Oncology, 24(13), 2019-2027.

Berry, D. A., Cirrincione, C., Henderson, I. C., Citron, M. L., Budman, D. R., Goldstein, L. J., ... & Winer, E. P. (2006). Estrogen-receptor status and outcomes of modern chemotherapy for patients with node-positive breast cancer. JAMA: the journal of the American Medical Association, 295(14), 1658-1667.

Buzdar, A. U., Ibrahim, N. K., Francis, D., Booser, D. J., Thomas, E. S., Theriault, R. L., ... & Hortobagyi, G. N. (2005). Significantly higher pathologic complete remission rate after neoadjuvant therapy with trastuzumab, paclitaxel, and epirubicin chemotherapy: results of a randomized trial in human epidermal growth factor receptor 2–positive operable breast cancer. Journal of clinical oncology, 23(16), 3676-3685.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Routledge Academic.

Coombes, R. C., Hall, E., Gibson, L. J., Paridaens, R., Jassem, J., Delozier, T., ... & Bliss, J. M.

(2004). A randomized trial of exemestane after two to three years of tamoxifen therapy in postmenopausal women with primary breast cancer. New England Journal of Medicine, 350(11), 1081-1092.

Donahue, J. M., Nichols, F. C., Li, Z., Schomas, D. A., Allen, M. S., Cassivi, S. D., ... & Deschamps, C. (2009). Complete pathologic response after neoadjuvant chemoradiotherapy for esophageal cancer is associated with enhanced survival. The Annals of thoracic surgery, 87(2), 392.

Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials—a review. Statistics in medicine, 3(3), 199-214.

Ellis, M. J., Coop, A., Singh, B., Mauriac, L., Llombert-Cussac, A., Jänicke, F., ... & Borgs, M. (2001). Letrozole is more effective neoadjuvant endocrine therapy than tamoxifen for ErbB-1–and/or ErbB-2–positive, estrogen receptor–positive primary breast cancer: evidence from a phase III randomized trial. Journal of Clinical Oncology, 19(18), 3808-3816.

Faneyte, I. F., Schrama, J. G., Peterse, J. L., Remijnse, P. L., Rodenhuis, S., & Van de Vijver, M. J. (2003). Breast cancer response to neoadjuvant chemotherapy: predictive markers and relation with outcome. British journal of Cancer, 88(3), 406-412.

Fleiss, J. L. 198 1. Statistical methods for rates and proportions.

Friede, T., & Kieser, M. (2004). Sample size recalculation for binary data in internal pilot study designs. Pharmaceutical Statistics, 3(4), 269-279.

Friedman, L. M., Furberg, C., & DeMets, D. L. (2010). Fundamentals of clinical trials. Springer.

Gonzalez-Angulo, A. M., Litton, J. K., Broglio, K. R., Meric-Bernstam, F., Rakkhit, R., Cardoso,

F., ... & Hortobagyi, G. N. (2009). High risk of recurrence for patients with breast cancer who have human epidermal growth factor receptor 2–positive, node-negative tumors 1 cm or smaller. Journal of clinical oncology, 27(34), 5700-5706.

Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. JAMA: the journal of the American Medical Association, 288(3), 358-362.

Iwata, H., Sato, N., Masuda, N., Nakamura, S., Yamamoto, N., Kuroi, K., ... & Toi, M. (2011). Docetaxel followed by fluorouracil/epirubicin/cyclophosphamide as neoadjuvant chemotherapy for patients with primary breast cancer. Japanese Journal of Clinical Oncology, 41(7), 867-875.

Kuroi, K., Toi, M., Tsuda, H., Kurosumi, M., & Akiyama, F. (2006). Issues in the assessment of the pathologic effect of primary systemic therapy for breast cancer. Breast Cancer, 13(1), 38-48.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. Controlled clinical trials, 2(2), 93-113.

Liedtke, C., Mazouni, C., Hess, K. R., André, F., Tordai, A., Mejia, J. A., ... & Pusztai, L. (2008). Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. Journal of Clinical Oncology, 26(8), 1275-1281.

Newell, D. J. (1978). Type II errors and ethics. British Medical Journal, 2(6154), 1789.

Norman, G. R., & Streiner, D. L. (1994). Principal components and factor analysis. Biostatistics. The Bare Essentials, 163-177.

Pagano, M., Gauvreau, K., & Pagano, M. (2000). Principles of biostatistics (pp. 196-213). Pacific Grove, CA: Duxbury.

Prowell, T. M., & Pazdur, R. (2012). Pathological complete response and accelerated drug approval in early breast cancer. New England Journal of Medicine, 366(26), 2438-2441.

Pusztai, L., Broglio, K., Andre, F., Symmans, W. F., Hess, K. R., & Hortobagyi, G. N. (2008). Effect of Molecular Disease Subsets on Disease-Free Survival in Randomized Adjuvant Chemotherapy Trials for Estrogen Receptor–Positive Breast Cancer. Journal of Clinical Oncology, 26(28), 4679-4683.

Rice, W. R. (1989). Analyzing tables of statistical tests. Evolution, 43(1), 223-225.

Rosen, G., Caparros, B., Huvos, A. G., Kosloff, C., Nirenberg, A., Cacavio, A., ... & Urban, C. (1982). Preoperative chemotherapy for osteogenic sarcoma: selection of postoperative adjuvant chemotherapy based on the response of the primary tumor to preoperative chemotherapy. Cancer, 49(6), 1221-1230.

SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2000-2004.

Shou, J., Massarweh, S., Osborne, C. K., Wakeling, A. E., Ali, S., Weiss, H., & Schiff, R. (2004). Mechanisms of tamoxifen resistance: increased estrogen receptor-HER2/neu cross-talk in ER/HER2–positive breast cancer. Journal of the National Cancer Institute, 96(12), 926-935.

Tischkowitz, M., Brunet, J. S., Bégin, L. R., Huntsman, D. G., Cheang, M. C., Akslen, L. A., ... & Foulkes, W. D. (2007). Use of immunohistochemical markers can refine prognosis in triple negative breast cancer. BMC cancer, 7(1), 134.

Toriola, A. T., & Colditz, G. A. (2013). Trends in breast cancer incidence and mortality in the United States: implications for prevention. Breast cancer research and treatment, 1-9.

Untch, M., Fasching, P. A., Konecny, G. E., Hasmüller, S., Lebeau, A., Kreienberg, R., ... &

Loibl, S. (2011). Pathologic complete response after neoadjuvant chemotherapy plus trastuzumab predicts favorable survival in human epidermal growth factor receptor 2–overexpressing breast cancer: Results from the TECHNO trial of the AGO and GBG study groups. Journal of Clinical Oncology, 29(25), 3351-3357.

Untch, M., Rezai, M., Loibl, S., Fasching, P. A., Huober, J., Tesch, H., ... & von Minckwitz, G. (2010). Neoadjuvant treatment with trastuzumab in HER2-positive breast cancer: results from the GeparQuattro study. Journal of Clinical Oncology, 28(12), 2024-2031.

Vogel, V. G., Costantino, J. P., Wickerham, D. L., Cronin, W. M., Cecchini, R. S., Atkins, J. N., ... & Wolmark, N. (2006). Effects of tamoxifen vs raloxifene on the risk of developing invasive breast cancer and other disease outcomes. JAMA: the journal of the American Medical Association, 295(23), 2727-2741.

Von Minckwitz, G., Eidtmann, H., Rezai, M., Fasching, P. A., Tesch, H., & Eggemann, H. (2010). Neoadjuvant chemotherapy with or without bevacizumab: primary efficacy endpoint analysis of the GEPARQUINTO study (GBG 44). Cancer Res, 70(24 suppl), S4-6.

Von Minckwitz, G., Kaufmann, M., Kümmel, S., Fasching, P., Eiermann, W., Blohmer, J. U., ... & Untch, M. (2008, December). Integrated meta-analysis on 6402 patients with early breast cancer receiving neoadjuvant anthracycline-taxane+/-trastuzumab containing chemotherapy. In San Antonio Breast Cancer Symposium, San Antonio, TX.

von Minckwitz, G., Untch, M., Blohmer, J. U., Costa, S. D., Eidtmann, H., Fasching, P. A., ... & Loibl, S. (2012). Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. Journal of Clinical Oncology, 30(15), 1796-1804.

Wolmark, N., Wang, J., Mamounas, E., Bryant, J., & Fisher, B. (2001). Preoperative chemotherapy in patients with operable breast cancer: nine-year results from National Surgical Adjuvant Breast and Bowel Project B-18. JNCI Monographs, 2001(30), 96-102.

APPENDIX

Tables

Table 1: Data presented and ranges explored (reference italicized)

	HER 2 (-)		**HER 2 (+)**	
Prevalence	*80%*		*20%*	
	ER (+)	**ER (-)**	**ER (+)**	**ER (-)**
Fixed prevalence	*70%*	*30%*	*60%*	*40%*
ER+ prevalence explored	0.1 to 0.9 by 0.05			
Fixed pCR rates:	*5%*	*15%*	*25%*	*45%*
Treatment effects on pCR rates explored	10%, 30%, 50%, and 70% increase from control pCR rate			
Effect Size	varied by proportions			

Table 2: Total pCR rates for varied prevalence of ER+ and ER-

HER2-				
prev ER+	pCR ER+	prev ER-	pCR ER-	pCR
0.05	0.05	0.95	0.15	0.145
0.1	0.05	0.9	0.15	0.14
0.15	0.05	0.85	0.15	0.135
0.2	0.05	0.8	0.15	0.13
0.25	0.05	0.75	0.15	0.125
0.3	0.05	0.7	0.15	0.12
0.35	0.05	0.65	0.15	0.115
0.4	0.05	0.6	0.15	0.11
0.45	0.05	0.55	0.15	0.105
0.5	0.05	0.5	0.15	0.1
0.55	0.05	0.45	0.15	0.095
0.6	0.05	0.4	0.15	0.09
0.65	0.05	0.35	0.15	0.085
0.7	0.05	0.3	0.15	0.08
0.75	0.05	0.25	0.15	0.075
0.8	0.05	0.2	0.15	0.07
0.85	0.05	0.15	0.15	0.065
0.9	0.05	0.1	0.15	0.06
0.95	0.05	0.05	0.15	0.055

HER2+				
prev ER+	pCR ER+	prev ER-	pCR ER-	pCR
0.05	0.25	0.95	0.45	0.44
0.1	0.25	0.9	0.45	0.43
0.15	0.25	0.85	0.45	0.42
0.2	0.25	0.8	0.45	0.41
0.25	0.25	0.75	0.45	0.4
0.3	0.25	0.7	0.45	0.39
0.35	0.25	0.65	0.45	0.38
0.4	0.25	0.6	0.45	0.37
0.45	0.25	0.55	0.45	0.36
0.5	0.25	0.5	0.45	0.35
0.55	0.25	0.45	0.45	0.34
0.6	0.25	0.4	0.45	0.33
0.65	0.25	0.35	0.45	0.32
0.7	0.25	0.3	0.45	0.31
0.75	0.25	0.25	0.45	0.3
0.8	0.25	0.2	0.45	0.29
0.85	0.25	0.15	0.45	0.28
0.9	0.25	0.1	0.45	0.27
0.95	0.25	0.05	0.45	0.26

Table 3: Power calculation for varied prevalence rates of ER+

HER2+			HER2-		
ER+ Prevalence	Z_β	Power (1- β)	ER+ Prevalence	Z_β	Power (1- β)
0.1	1.916906	0.972375	0.1	1.687489	0.9542453
0.2	1.746893	0.9596721	0.2	1.503327	0.9336226
0.3	1.574272	0.9422877	0.3	1.329533	0.9081639
0.4	1.398373	0.9189995	0.4	1.164603	0.8779101
0.5	1.218388	0.8884617	0.5	1.007251	0.843093
0.6	1.033313	0.8492712	0.6	0.856363	0.8041016
0.7	0.841873	0.8000705	0.7	0.710954	0.7614436
0.8	0.642392	0.7396907	0.8	0.570135	0.715707
0.9	0.43258	0.6673401	0.9	0.433088	0.6675245

Figures

Figure 1: Plot of pCR control rate against prevalence of ER+ for both HER2+ and HER2-

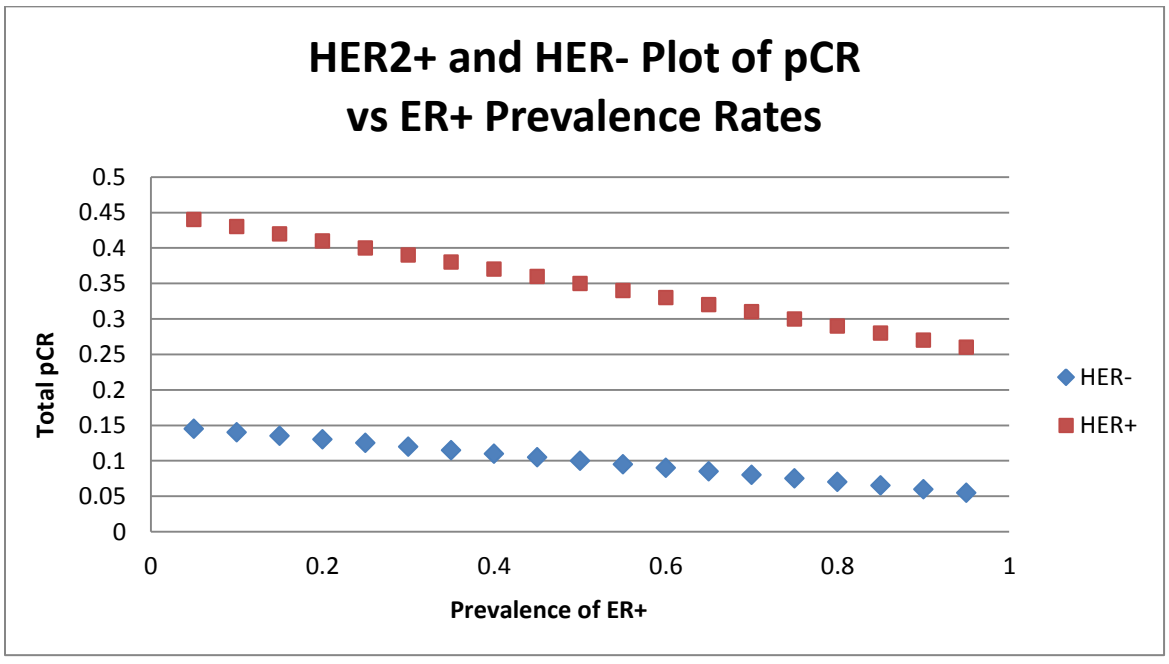


Figure 2: Plot of pCR control rate against prevalence of ER+ for HER2+

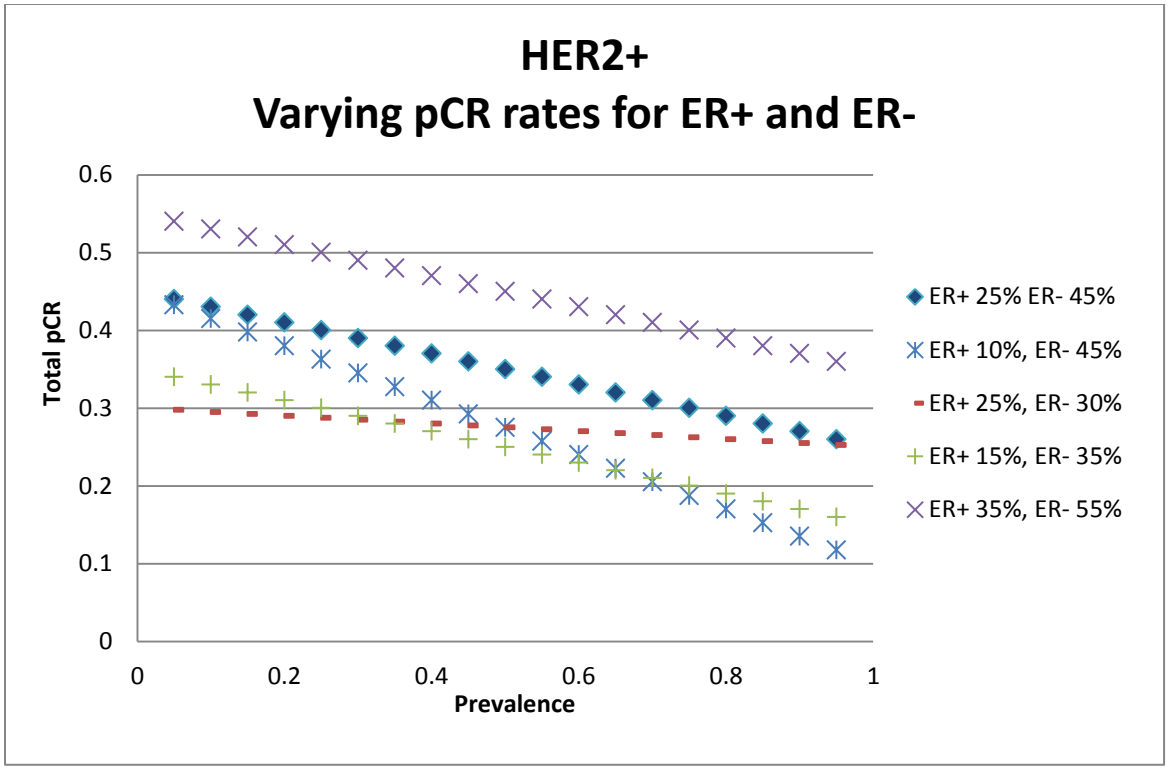
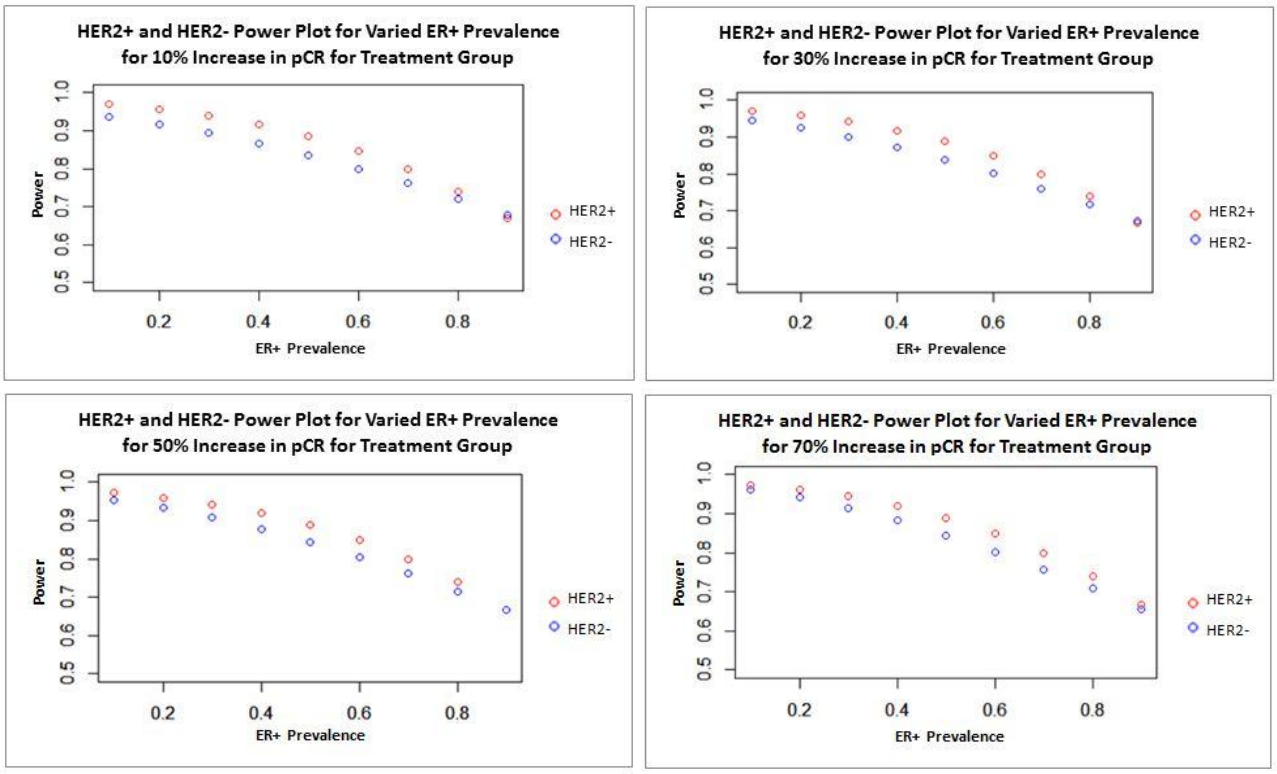


Figure 3: Power plots for varied ER+ Prevalence for 10%, 30%, 50%, and 70% increase



Calculations

1) Total pCR rate: (0.70)(.05) + (0.30)(.15) = .08

2) Question 2:

HER2-

$$N = \frac{4 * (1.96 + 0.84)^2 * 0.1 * 0.9}{|0.12 - 0.08|^2} = 1764$$

HER2+

$$N = \frac{4 * (1.96 + 0.84)^2 * 0.4125 * 0.5875}{|0.495 - 0.33|^2} = 279.15 => 280$$

3) Power calculation for HER-:

$$Z_\beta = \frac{-1.96\sqrt{2 * 0.0875 * (1 - 0.0875)} + \sqrt{1764}\,(0.105 - 0.07)}{\sqrt{0.105(1 - 0.105) + 0.07(1 - 0.07)}}$$

$$Z_\beta = 0.64$$