Published in final edited form as:
Clin Trials. 2013 ; 10(6): 886–895. doi:10.1177/1740774513506618.

Multicenter trials using FDG PET to predict chemotherapy response : Effects of differential measurement error and bias on power calculations for unselected and enrichment designs

proxy Brenda F Kurland,
University of Pittsburgh - Biostatistics, Pittsburgh, Pennsylvania, United States

proxy Robert K Doot,
University of Pennsylvania - Radiology, Philadelphia, Pennsylvania, United States

proxy Hannah M Linden,
University of Washington - Medicine, Seattle, Washington, United States

proxy David A Mankoff, and
University of Pennsylvania - Radiology, Philadelphia, Pennsylvania, United States

proxy Paul E Kinahan
University of Washington - Radiology, Seattle, Washington, United States

Keywords

predictive biomarker; FDG PET imaging; multicenter; measurement error; bias; power; design

Introduction

Predictive biomarkers have great potential to personalize cancer therapies, but their clinical validation may be complicated and resource-intensive [1]. Retrospective analysis of banked samples from randomized clinical trials are an ideal setting for timely and cost-effective assessment of potential predictive biomarkers identified in smaller studies [2, 3]. However, unlike novel assays performed on formalin-fixed tissue samples, molecular imaging biomarkers cannot be derived from banked samples. Without access to retrospective analysis of large datasets, functional imaging biomarkers and biomarkers requiring fresh blood or tissue must be developed in a more incremental fashion. A prospective multicenter observational study is a sensible step in order to reproduce earlier findings in a multicenter setting, and design considerations for these studies are the focus of this manuscript.

A multicenter replication study may be used to evaluate potential cutpoints for the predictive biomarker for use in guiding therapy, and in general will inform the design of future studies. However, the principal analytic goal is merely to show a relationship between the predictive biomarker and the clinical endpoint. Since this is only a small step toward marker validation, investigators will want to conduct the trial quickly and efficiently. If new targeted therapies

Fred Hutchinson Cancer Research Center - Clinical Statistics, 1100 Fairview Ave N, D5-360 Seattle Washington 98109, United States, T: 206-667-2804

emerge or the standard of care changes, development of a biomarker may be stunted [2], even if the marker still has potential under the new treatment paradigm. Enrolling from a large number of centers can speed accrual, but adds startup costs at each center, and may introduce bias or increased variability (error).

Enriched samples are used to speed early drug testing by including only patients who are most likely to benefit [2, 4, 5]. Later trials may test the therapy in an unselected patient population, to prevent bias introduced by the selection process, especially in case the mechanism of action is different from expected [6, 7].

Rapid enrollment (with many study sites) and enrichment designs are both strategies that may facilitate development of a predictive biomarker. The objective of this manuscript is to demonstrate the use of simulation studies to select a design that will optimize study power, under realistic scenarios derived from the substantive literature. The case study for illustration is a multicenter study using ^{18}F-fluorodeoxyglucose (FDG) PET to monitor response to breast cancer neoadjuvant chemotherapy. Our prior work demonstrated that PET measurement error and quantification methods may have a great impact on study power [8]. This simulation study extends that work by examining bias and measurement error for individual measurements. Under this realistic scenario for generation of study data, we can examine the study power implications for different trial designs: recruitment from primary sites versus accelerating enrollment with secondary sites, and unselected designs versus enrichment for patients considered more likely to show an association between the biomarker and clinical response.

Case study: Early FDG PET response as a predictive biomarker for response to breast cancer neoadjuvant chemotherapy

Several studies have shown that early metabolic response measured by FDG PET after 1-4 cycles of neoadjuvant chemotherapy for locally advanced breast cancer (LABC) predicts subsequent histopathologic response (Figure 1(a)) [9-13]. These results suggest that early changes in FDG uptake could be used as a predictive biomarker, with potential to direct therapy by discontinuing ineffective therapies and continuing effective therapies (Figure 1(b)). Testing of this treatment paradigm could be accomplished using a marker-based strategy design [14]. However, while published evaluations of FDG PET response were relatively consistent in demonstrating an association between early metabolic response and histopathologic response, most were single-center observational studies. Breast cancer patient populations and neoadjuvant chemotherapy regimens varied, as did FDG PET patient preparation, scanning protocol, and quantification of FDG uptake. Since FDG PET scans must be performed prospectively, this is a scenario where a multicenter observational study is a viable intermediate step toward validation of serial FDG PET as a predictive biomarker.

The most common method of quantifying FDG uptake is the standardized uptake value (SUV), which normalizes the measured tissue activity in a region of interest (identified within a tumor site) by injected dose and patient mass. Following the observation that cell kill by cytotoxic chemotherapy is expected to occur as a percentage of tumor cells rather than absolute numbers [15], metabolic response (decrease in FDG uptake as a result of effective therapy) is measured as the percentage change in FDG SUV:

$$SUV\ change = 100 * \frac{(\mathrm{SUV}_{1} - \mathrm{SUV}_{0})}{SUV_{0}}$$

where SUV_0 is baseline (pre-therapy) FDG SUV, and SUV_1 is follow-up (i.e., after 2 cycles of neoadjuvant chemotherapy) FDG SUV. Several percentage change thresholds have been proposed as reflecting a minimal clinically meaningful metabolic response, or a response that is unlikely to be due to chance [16-18]. However, factors affecting the apparent change in FDG SUV for breast cancer treatment response may limit the utility of standardized criteria. Perhaps most important is that untreated breast tumors may have lower FDG uptake than other cancers routinely imaged with FDG PET, but are still detectable due to the low average uptake of normal breast tissue [19]. For low-uptake tumors, circulating FDG in the bloodstream and uptake by surrounding normal tissue may account for a relatively large fraction of the measured SUV; since these fluctuating background levels are not expected to change predictably with treatment, they may obscure decreases in tumor FDG metabolism that could accompany treatment response. Prior studies have shown that FDG PET SUV is not effective in measuring response when pre-therapy SUV is < 3 [10, 20, 21], and that dynamic PET imaging provides superior precision in measuring response in low-uptake breast tumors [12]. While results from dynamic PET imaging inform our simulation design, this multicenter study will only examine the standardized uptake value (SUV), which requires less scanner time, is measured by clinical scanner software, and is more well-developed toward broad use in clinical trials [17].

Response to pre-surgical breast cancer chemotherapy is evaluated by pathologic complete response (pCR) at surgery, which has a strong association with long-term clinical prognosis [22]. As assessed by the local imaging study used for data simulation, pCR has occurred when no invasive tumor is seen by microscopic inspection of post-therapy surgical breast tissue specimens. The primary analysis will be logistic regression, with percent change in SUV predicting pCR and the C-statistic examined as a preliminary measure of classification accuracy. Known clinical predictors of pCR should be included as covariates in logistic regression models evaluating SUVchange [23] but are omitted here for simplicity of presentation.

The goal of these studies is to optimize use of resources for a prospective multicenter observational study as an intermediate step for validation of SUVchange as a predictive biomarker for breast cancer neoadjuvant therapy response. Therefore, the focus of simulations is on study power for combinations of design strategies (limited versus broad recruitment of sites; unselected versus enrichment design). To represent variability associated with inexperienced sites performing PET for quantitative analysis, we assume two classes of sites: experienced (primary) sites that are likely to produce SUVs with lower measurement error, and secondary sites expected to yield SUVs with higher measurement error due to past performance, limited experience, and/or higher staff turnover. We estimate the loss of power expected when accrual is accelerated through increasing reliance on secondary sites. In the ideal scenario, all sites could achieve the performance of primary sites. However, this expectation is unrealistic; generalizability and practical considerations require the inclusion of secondary sites as well as sites with greater experience and demonstrated measurement precision. We therefore consider the impact on power when a study with a fixed sample size varies in the proportion from primary and secondary sites.

In the *unselected design*, all enrolled patients are followed for early response by FDG PET SUV and pathologic complete response. In the *enrichment design*, patients with SUV < 3 at baseline are not be assessed for metabolic or histopathologic response, but are instead replaced by patients with higher baseline FDG uptake. For a prospective observational study with patients receiving standard therapy, patient scans are the greatest contributor to study costs. To keep total costs consistent, for every two patients with SUV < 3, an additional patient is recruited, and assessed for response if baseline SUV ≥ 3. In this way, we can compare the study power for the unselected design and the enrichment design, for the same total study cost.

Results for both design strategies rely heavily on the measurement error scenarios for individual SUV values. Simulation results will only be valuable if the input variables are based on relevant empirical data. The simulation parameters and their rationale are described below.

Simulation parameters

Key parameters for the simulation are described in Table 1. We assumed that the true distribution of FDG SUV in patients with LABC was similar to that observed in our local imaging study [12]. SUV data for the simulation were generated by sampling from the pairs of SUV values (SUV at baseline and mid-therapy), with replacement [24]. Response data were simulated by random sampling from a binomial distribution with probability estimated by logistic regression models fitted to our local LABC study. Separate logistic regression models were fitted for baseline SUV ≥ 3 (N=50) and baseline SUV < 3 (N=25):

$$\begin{array}{c}\text{baseline SUV} \geq 3\text{: } logit(pCR) = -\ 3.13 - 0.039 * SUV\ change \\ \text{baseline SUV} < 3\text{: } logit(pCR) = -\ 2.29 - 0.015 * SUV\ change.\end{array} \qquad \text{Equation(1)}$$

For example, simulated pCR for a patient with true baseline SUV ≥ 3 and 10% decline would have a 0.06 probability of pCR, in contrast to a 0.49 probability of pCR with 80% decline. For patients with true baseline SUV < 3, estimated pCR probabilities were 0.11 and 0.25 for 10% and 80% decline, respectively.

The sample size of N=100 was chosen as an incremental increase in size over the single center study, and a reasonable moderate size for a multicenter study. As described in the Introduction, rapid multicenter replication of single-center associations is the goal, rather than specific effect sizes or biomarker cutpoint refinement. For an enrichment design, more than 100 patients would have a first scan, but fewer than 100 would have a second scan and be used for analysis. “Percent secondary” refers to the percentage of patients who would be recruited from sites with less experience in monitoring breast cancer chemotherapy using FDG PET.

While simulated response data were generated from the reference SUV data, power under scenarios for measurement error and study design were based on “observed” SUV data generated by adding error to the reference values:

$$\begin{pmatrix} \mathrm{SUV}o_{ij0} \\ \mathrm{SUV}o_{ij1} \end{pmatrix} = \begin{pmatrix} \mathrm{SUV}r_{ij0} \\ \mathrm{SUV}r_{ij1} \end{pmatrix} + \begin{pmatrix} e_{ij0} \\ e_{ij1} \end{pmatrix}.$$

In this equation, SUVo_{ij0} is the observed baseline SUV for the *j*th patient at center *i*, SUVo_{ij1} is the follow-up measurement for the same patient, SUVr_{ij0} and SUVr_{ij0} are the reference SUV values for the same patient (the values sampled with replacement from the N=75 LABC dataset), and e_{ij0} and e_{ij1} are measurement error.

We next describe the measurement error scenarios examined (Table 1). “Percent error” refers to SUV measurement error as it has been described in much of the test-retest repeatability data to date. While the widely accepted measurement error standard deviation from test-retest studies is 10% [25], the average is higher for multicenter studies [26], even for multicenter studies of pseudo-patient phantoms, in which biologic variability does not occur [27, 28]. Based on these and other repeatability and reproducibility studies we chose representative values of 20% error for primary sites and 40% error for secondary sites. We also examined systematic bias, in which observed SUV values are not only more variable than the reference data, but have an expected value (mean) different from that of the reference data. Positive and negative 20% bias were considered for primary sites, and positive and negative 40% bias were considered for secondary sites.

Although PET measurement error is usually described as a percentage, the seminal work of Weber [25] provided evidence to support describing errors in individual PET SUV measurements in terms of absolute values of SUV units. Therefore, we also examined a range of simulations with absolute error and bias. However, since we made the common assumption of a normal (Gaussian) distribution for the measurement error, we used as a starting point the approximately normal distribution of log-transformed SUVs [29]. Values for log(SUV) of 0.25 and 0.5, approximately 1/4 and 1/2 of the pooled standard deviation of baseline and follow-up reference values, were used as values for absolute bias and standard deviation of error for primary and secondary sites respectively. Due to the log-transformation of SUV for absolute error/bias, we expected the distributions of simulated data with percentage error/bias and absolute error/bias to be similar. For simplicity of presentation, we assumed the bias and/or error distribution was the same for all primary sites and for all secondary sites.

With *i* taking the value 0 for primary sites and 1 for secondary sites, the error terms for the “percent bias” scenario have the distribution

$$\begin{pmatrix} e_{ij0} \\ e_{ij1} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathrm{B}_i \times \mathrm{SUV}r_{ij0} \\ \mathrm{B}_i \times \mathrm{SUV}r_{ij1} \end{pmatrix}, \begin{pmatrix} (\mathrm{V}_i \times \mathrm{SUV}r_{ij0})^2 & \rho \times \mathrm{V}_i \times \mathrm{SUV}r_{ij0} \times \mathrm{V}_i \times \mathrm{SUV}r_{ij0} \\ \rho \times \mathrm{V}_i \times \mathrm{SUV}r_{ij0} \times \mathrm{V}_i \times \mathrm{SUV}r_{ij0} & (\mathrm{V}_i \times \mathrm{SUV}r_{ij1})^2 \end{pmatrix}\right),$$

with B, V, and ρ as described in Table 1. The calculations for absolute error and bias are similar, but are applied to log-transformed reference values.

The error and bias models can result in observed SUV data that would likely not be analyzed as observed in the hypothetical trial under consideration. For example, a mass with an SUV of 0.8 at baseline would be unlikely to be identified as a malignancy. Therefore, ceiling and floor values were established for simulated SUV values, with a minimum of 1.3 for baseline and normal breast average of 0.5 [19] for follow-up, and a maximum of 20 for both.

Power is the probability of rejecting the null hypothesis of no relationship in a logistic regression model predicting pathologic complete response (pCR) by percent change in SUV, with a two-sided Type I error rate of $\alpha = 0.05$.

Results

Results for simulations with 10,000 iterations each are presented in Figures 2 and 3, and Table 2. Figure 2 shows power for recruitment strategies in which 25%-75% of patients are from secondary sites, under different scenarios for bias and error in measuring FDG PET SUV. Table 2, column 1 is the power based on "reference" SUV values and responses. Without additional measurement error, the power for finding an association between SUV change and pCR was 0.92 for the unselected design (Figure 2, horizontal line).

Impact of recruitment strategy

Power for the unselected design was substantially reduced (lower than the reference line in Figure 2) under all bias and/or measurement error scenarios examined. When most additional error was of modest size (for example, when 75% of simulated patients were from primary sites with lower additional measurement error), the power was about 10-30 percentage points lower than for the reference sample. The lowest power was for the scenario with negative bias as a percentage of the reference SUV. This was not due to a weaker true relationship between SUVchange and pCR, since for all scenarios the response values were simulated from reference data, rather than simulated "observed" data with additional measurement error. With negative bias it is likely that percent change was attenuated by simulated "observed" data hitting the floor values of 1.3 for baseline and 0.5 for follow-up SUV. Accelerating study enrollment by recruiting 75% of patients rather than 25% from sites expected to have greater measurement error (moving along the X axis in Figure 2) led to an additional decrease in power of 10-30 percentage points.

For the unselected design, scenarios with absolute measurement error and bias resulted in the same power as for absolute measurement error alone (Table 2, columns 5-7). Absolute bias largely cancelled when calculating percentage change in SUV, as would occur if a scanner were miscalibrated but did not drift or undergo further calibration between scans. This presumed cancellation of bias for serial scans may not apply if a patient is not imaged on the same PET/CT scanner each time. Bias as a percentage of baseline SUV and follow-up SUV did not generally cancel when calculating percentage change.

Impact of unselected vs. enrichment design

Approximately 1/3 of the baseline scans (33/100) sampled from the reference data had SUV1 (baseline SUV) less than 3. Not performing follow-up scans for those patients would free funds for recruitment of approximately 16 additional patients, of whom about 11 would have SUV1 ≥ 3, for a final average sample size of N=77. The greater effect size for detecting an association between SUVchange and pCR for patient scans with SUV1 ≥ 3 was offset by the smaller sample size, resulting in lower power (0.87) for the enrichment design using reference data (Table 2, column 1).

Power for the enrichment design (dashed lines in Figure 3) ranged from substantially lower than to slightly higher than for the unselected design (solid lines in Figure 3) for the error and bias scenarios examined. Figure 3(a) shows scenarios with error and/or bias as a percentage of SUV values (columns 2-4 of Table 2). Figure 3(b) shows scenarios with absolute error and/or bias (columns 5-7 of Table 2). For error as a percentage of SUV and no systematic bias, power was similar for unselected and enrichment designs (Table 2, column 2). Under absolute error (column 5) the percentage differential between power of unselected and enriched designs was similar to that for reference data. Power for enrichment designs compared to unselected designs was low under negative bias, especially for absolute bias (column 7). Both scenarios for positive bias (columns 3, 6) had power similar to that of the unselected design. These scenarios did not have higher power (relative to other enrichment design scenarios) because more of the data were generated under the larger SUV1 ≥ 3 model effect size, since the additional bias and/or measurement error does not affect the underlying model for generating response values in the simulation. (Equation 1 always used reference data without additional error.) However, the positive bias did lead to fewer excluded patients under the enriched design: the average sample size was close to 90 for all positive bias scenarios. This was likely the reason that power could be approximately the same as for the unselected design for these scenarios, but never substantially higher. The C-statistic was examined as a preliminary measure of classification accuracy, in anticipation of cutpoint selection for future marker-directed therapy (Figure 1(b)). The average C-statistic for reference data was 0.71 for both unselected and enrichment designs, and was 0.63-0.69 for all measurement error scenarios examined.

Discussion

Power calculations from simulation studies may be used to inform the efficient design of a multicenter study to replicate single-institution findings. This scenario is a common-sense step toward clinical validation of predictive biomarkers when the biomarker cannot be assessed retrospectively in archived data from large clinical trials. As a case study, we examined serial FDG PET to predict early response to breast cancer neoadjuvant chemotherapy, using both unselected designs and enrichment designs assessing only patients with high initial FDG uptake. The number of total scans was held fixed since the cost of scans is a limiting factor for studies involving quantitative imaging. Reference power (from source data with no added error) was 0.92 for N=100 to detect an association between percentage change in SUV and response. With moderate (20%) simulated measurement error for 3/4, 1/2, and 1/4 of measurements and 40% for the remainder, power was 0.70, 0.61, and 0.53 respectively. Reduction of study power was similar for other manifestations of measurement error (bias as a percentage of true value, absolute error, and absolute bias). Any increased power by enriching the sample for patients with initial SUV ≥ 3 was offset by the smaller analysis sample size due to the cost of initial scans for patients who were not assessed for response. The simulations demonstrated that different assumptions about measurement error (drawn from observations from test-retest studies in both patients and pseudo-patient phantoms) had a substantial impact on study power.

In addition to addressing design considerations for a specific scenario, this case study is intended as a template for future simulation studies to design clinical trials involving quantitative imaging. It demonstrates the value of simulation studies in situations where different study designs are under consideration, and where preliminary data are available from complementary sources (breast cancer response to therapy; measurement error parameters for FDG PET). The magnitude of effect of PET measurement error on study power supports the need for cross-calibration of PET scanners, and for thorough understanding of PET measurement error. For example, it is unclear whether PET measurement error manifests in SUV units, log(SUV) units, or as a percentage of true SUV. Results from a reproducibility study with 26 patients suggest that mean-variance relationships between measurement error and SUV magnitude could be avoided by modeling error as a percentage for SUVmax (based on the pixel with maximum uptake – the measure used for clinical reports and for the reference data study) or by absolute SUV units for SUVmean (the average within a region of interest of fixed size, favored for research due to superior reproducibility) [30].

Additional future work may address how to incorporate site-specific calibration results into study analysis. If calibration errors are detected during site initiation they are corrected, but similar errors may be likely to occur at that site in the future. While incorporation of audit results into measurement of study effects [31] does not apply directly in a setting in which error comes from primary data acquisition rather than abstraction errors, errors in variables methods will be applicable for later-stage studies for biomarker refinement.

The primary limitation of this approach to study design is that the simulation parameters could be incorrect, or not generalizable. For example, pCR was simulated without consideration for uncertainty in measuring logistic regression model parameters (Equation 1). Rather than a true modest effect without power to find significance, the observed association between percent change in SUV and pCR when SUV1 < 3 could have been random fluctuation from a truth of no relationship. When the power for the enrichment design is calculated under the assumption of no association between SUVchange and pCR when SUV1 < 3, we find scenarios that would favor the enrichment design (Supplementary Table), including the reference data without additional measurement error. Additionally, there have been changes in chemotherapy regimens and in patient selection (and tumor subtype) for neoadjuvant chemotherapy since the reference study accrued patients (1995-2007) [32].

In summary, the characteristics of measurement error and bias (percentage versus absolute), and the magnitude of error and bias, can have great impact on power and sample size. As study reliance on secondary sites increased to 75%, power was up to 24 percentage points lower than for studies with 25% of subjects from secondary sites, and up to 56 percentage points lower than the reference power with no additional measurement error. Enrichment designs could result in modest increases in power, but for most scenarios had power slightly lower than for an unselected design. Ongoing quantitative characterizations of PET imaging data and sources of measurement error will allow refinement of methods to inform planning of multicenter clinical trials involving quantitative imaging.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Larry Clarke and Robert Nordstrom from the NCI Cancer Imaging Program and members of the Quantitative Imaging Network for many insightful and helpful discussions. In addition, the authors thank the members of the ACRIN PET core laboratory and AAPM Task Group 145 and the RSNA Quantitative Imaging Biomarkers Alliance for support and helpful feedback. This work was supported by National Cancer Institute grants U01-CA148131, P50-CA138293, P30-CA047904, and NCI Contract 24XS036-004.

References

1. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J Clin Oncol. 2009; 27:4027–4034. [PubMed: 19597023]
2. Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: lessons from real trials. Clinical trials. 2010; 7:567–573. [PubMed: 20392785]
3. Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Personalized Medicine. 2010; 7:33–47. [PubMed: 20383292]
4. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res. 2004; 10:6759–6763. [PubMed: 15501951]
5. Lai TL, Lavori PW, Shih MC, Sikic BI. Clinical trial designs for testing biomarker-based personalized therapies. Clinical trials. 2012; 9:141–154. [PubMed: 22397801]
6. Hoering A, Leblanc M, Crowley JJ. Randomized phase III clinical trial designs for targeted agents. Clin Cancer Res. 2008; 14:4358–4367. [PubMed: 18628448]
7. Allegra CJ, Jessup JM, Somerfield MR, et al. American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. J Clin Oncol. 2009; 27:2091–2096. [PubMed: 19188670]
8. Doot RK, Kurland BF, Kinahan PE, Mankoff DA. Design Considerations for using PET as a Response Measure in Single Site and Multicenter Clinical Trials. Acad Radiol. 2012; 19:184–190. [PubMed: 22104290]
9. Wahl RL, Zasadny K, Helvie M, et al. Metabolic monitoring of breast cancer chemohormonotherapy using positron emission tomography: initial evaluation. J Clin Oncol. 1993; 11:2101–2111. [PubMed: 8229124]
10. McDermott GM, Welch A, Staff RT, et al. Monitoring primary breast cancer throughout chemotherapy using FDG-PET. Breast Cancer Res Treat. 2007; 102:75–84. [PubMed: 16897427]
11. Schwarz-Dose J, Untch M, Tiling R, et al. Monitoring primary systemic therapy of large and locally advanced breast cancer by using sequential positron emission tomography imaging with [18F]fluorodeoxyglucose. J Clin Oncol. 2009; 27:535–541. [PubMed: 19075273]

12. Dunnwald LK, Doot RK, Specht JM, et al. PET tumor metabolism in locally advanced breast cancer patients undergoing neoadjuvant chemotherapy: value of static versus kinetic measures of fluorodeoxyglucose uptake. Clin Cancer Res. 2011; 17:2400–2409. [PubMed: 21364034]

13. Tateishi U, Miyake M, Nagaoka T, et al. Neoadjuvant chemotherapy in breast cancer: prediction of pathologic response with PET/CT and dynamic contrast-enhanced MR imaging--prospective assessment. Radiology. 2012; 263:53–63. [PubMed: 22438441]

14. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol. 2005; 23:2020–2027. [PubMed: 15774793]

15. Kasamon YL, Jones RJ, Wahl RL. Integrating PET and PET/CT into the risk-adapted therapy of lymphoma. J Nucl Med. 2007; 48(Suppl 1):19S–27S. [PubMed: 17204717]

16. Young H, Baum R, Cremerius U, et al. Measurement of clinical and subclinical tumour response using [18F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. European Journal of Cancer. 1999; 35:1773–1782. [PubMed: 10673991]

17. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. J Nucl Med. 2009; 50(Suppl 1):122S–150S. [PubMed: 19403881]

18. de Langen AJ, Vincent A, Velasquez LM, et al. Repeatability of 18F-FDG Uptake Measurements in Tumors: A Metaanalysis. J Nucl Med. 2012

19. Zasadny KR, Wahl RL. Standardized uptake values of normal tissues at PET with 2-[fluorine-18]-fluoro-2-deoxy-D-glucose: variations with body weight and a method for correction. Radiology. 1993; 189:847–850. [PubMed: 8234714]

20. Doot RK, Dunnwald LK, Schubert EK, et al. Dynamic and static approaches to quantifying 18F-FDG uptake for measuring cancer response to therapy, including the effect of granulocyte CSF. J Nucl Med. 2007; 48:920–925. [PubMed: 17504870]

21. Castell F, Cook GJ. Quantitative techniques in 18FDG PET scanning in oncology. Br J Cancer. 2008; 98:1597–1601. [PubMed: 18475291]

22. Draft guidance for industry: Pathologic complete response in neoadjuvant treatment of high-risk early-stage breast cancer: Use as an endpoint to support accelerated approval [http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM305501.pdf]

23. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. PLoS medicine. 2012; 9:e1001216. [PubMed: 22675273]

24. Cheng, R.; Currie, C. In: Rossetti, MD.; Hill, RR.; Johansson, B.; Dunkin, A.; Ingalls, RG., editors. Resampling methods of analysis in simulation studies; In 2009 Winter Simulation Conference; New York. Institute of Electrical and Electronics Engineers; p. 2009p. 45-59.

25. Weber WA, Ziegler SI, Thodtmann R, Hanauske AR, Schwaiger M. Reproducibility of metabolic measurements in malignant tumors using FDG PET. J Nucl Med. 1999; 40:1771–1777. [PubMed: 10565769]

26. Velasquez LM, Boellaard R, Kollia G, et al. Repeatability of 18F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. J Nucl Med. 2009; 50:1646–1654. [PubMed: 19759105]

27. Westerterp M, Pruim J, Oyen W, et al. Quantification of FDG PET studies using standardised uptake values in multi-centre trials: effects of image reconstruction, resolution and ROI definition parameters. European journal of nuclear medicine and molecular imaging. 2007; 34:392–404. [PubMed: 17033848]

28. Fahey FH, Kinahan PE, Doot RK, et al. Variability in PET quantitation within a multicenter consortium. Med Phys. 2010; 37:3660–3666. [PubMed: 20831073]

29. Thie JA, Hubner KF, Smith GT. The diagnostic utility of the lognormal behavior of PET standardized uptake values in tumors. J Nucl Med. 2000; 41:1664–1672. [PubMed: 11037996]

30. Nahmias C, Wahl LM. Reproducibility of standardized uptake value measurements determined by 18F-FDG PET in malignant tumors. J Nucl Med. 2008; 49:1804–1808. [PubMed: 18927325]

31. Shepherd BE, Shaw PA, Dodd LE. Using audit information to adjust parameter estimates for data errors in clinical trials. Clinical trials. 2012; 9:721–729. [PubMed: 22848072]

32. Schott AF, Hayes DF. Defining the benefits of neoadjuvant chemotherapy for breast cancer. J Clin Oncol. 2012; 30:1747–1749. [PubMed: 22508810]

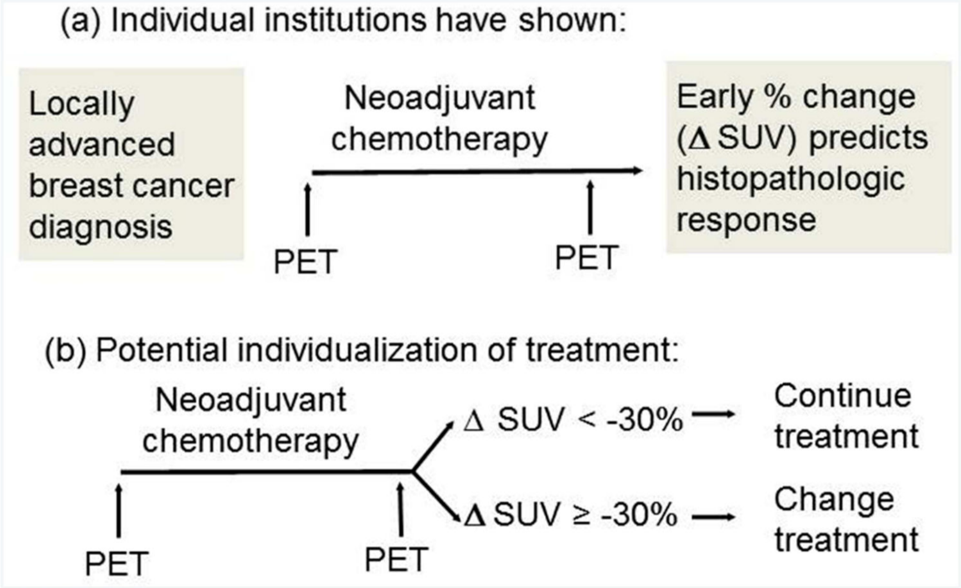
(a) Individual institutions have shown:
Locally advanced breast cancer diagnosis
Neoadjuvant chemotherapy
PET
PET
Early % change (Δ SUV) predicts histopathologic response
(b) Potential individualization of treatment:
Neoadjuvant chemotherapy
PET
PET
Δ SUV < -30%
Continue treatment
Δ SUV ≥ -30%
Change treatment

Figure 1.

Schemas for using quantitative FDG PET imaging as a predictive biomarker. (a) Design for single-institution observational studies, and for proposed multicenter replication. (b) Potential clinical application of marker-driven therapy.

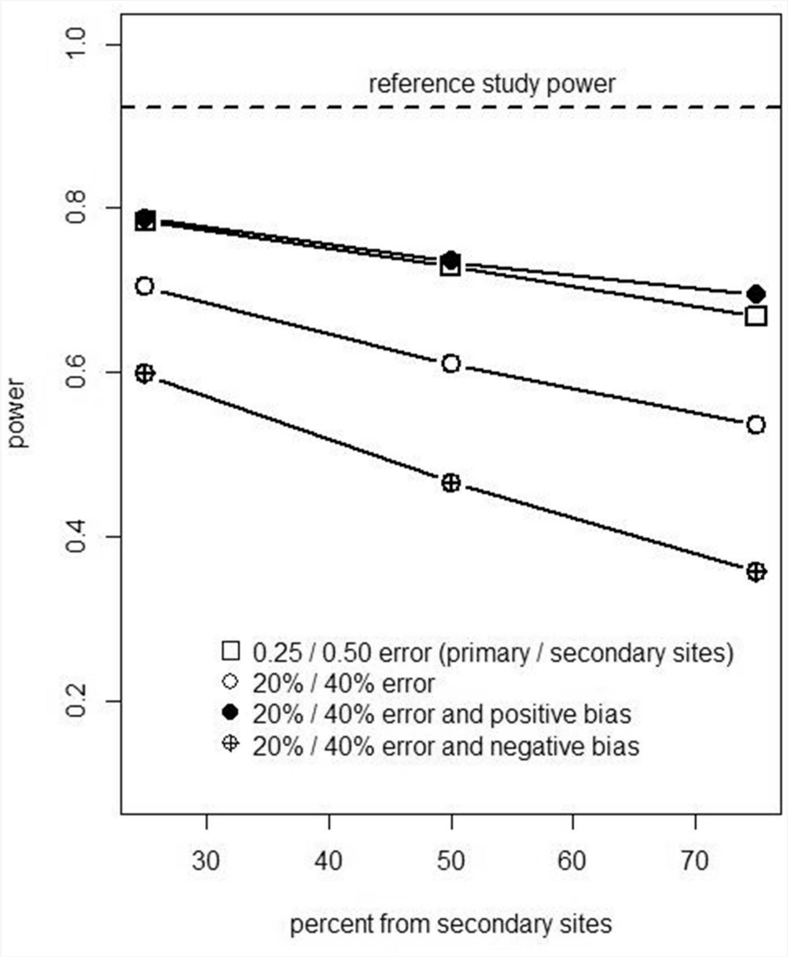

Figure 2.

Graphical presentation of Table 2 error conditions 1-5: impact of recruitment strategy (balance of primary and secondary sites) on study power, N=100 patients (200 scans).

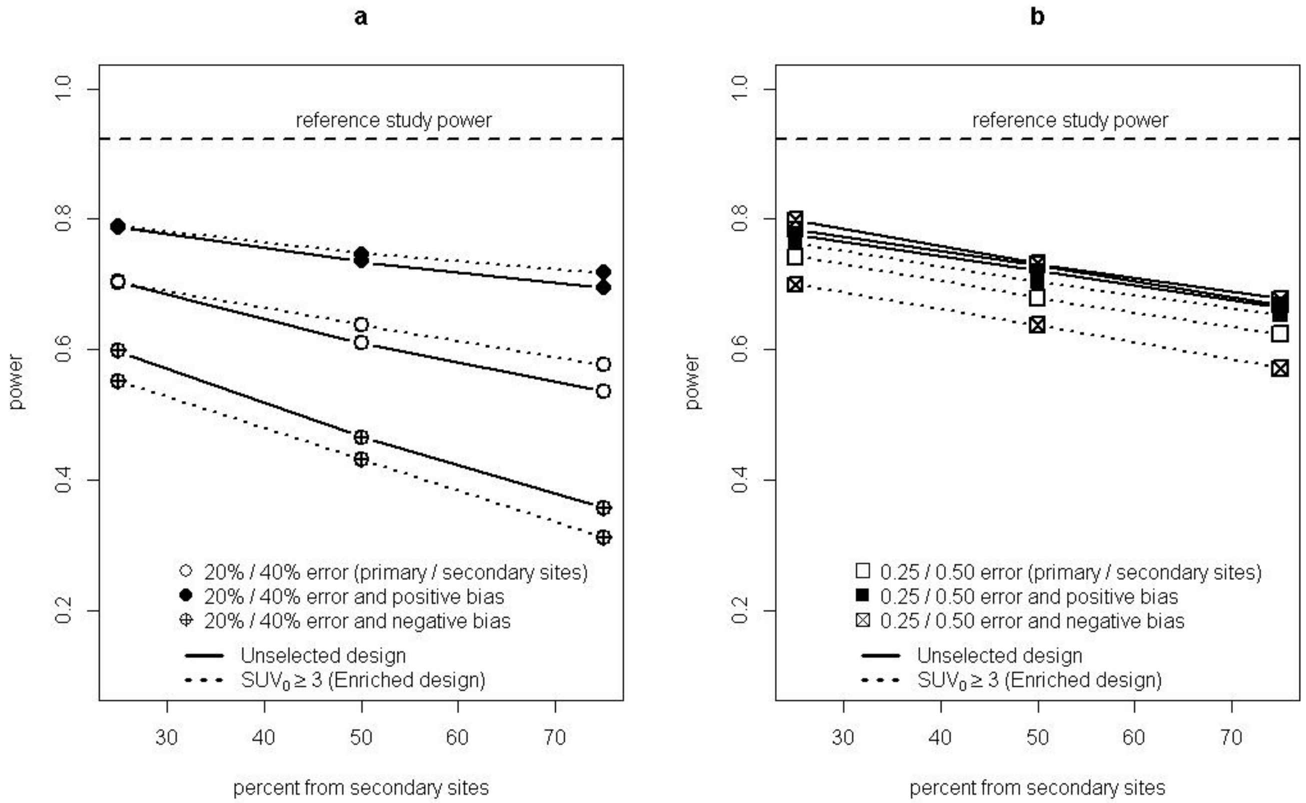

Figure 3.

Graphical presentation of relative power for unselected design (solid lines) and enrichment design (dashed lines), for a study with N=200 scans. (a) error and/or bias as a percentage of reference values. (b) error and/or bias in absolute log(SUV) units.

Table 1

Key parameters in the simulation of power calculations for a multicenter trial estimating the association between percent change in FDG PET SUV and pathologic response to neoadjuvant chemotherapy for patients with locally advanced breast cancer.

Simulation input	Description	Values explored Primary sites	Values: explored: Secondary sites
Reference SUV data	Baseline and mid-therapy SUV values were sampled together from the University of Washington locally advanced breast cancer (LABC) dataset (N=75), with replacement, to form simulation datasets.		
Response data	1 = pathologic complete response 0 = other than pathologic complete response Response (0/1) was generated from a binomial distribution with the response probability modeled by logistic regression for percentage change in true SUV.	See equation 1	
Sample size	Sample size for hypothetical study (all centers combined, without exclusions based on baseline SUV)	N=100	
Percent secondary	Percent of SUV measurements taken from sites expected to have higher measurement error	25% 50% 75%	
Percent error (V%)	Random noise (with mean 0 and standard deviation V% of the reference SUV value) added to reference SUV values	20%	40%
Percent bias (B%)	Random noise (with mean B% of the true SUV value, and standard deviation V% of the true SUV value) added to the reference SUV values	20% -20%	40% -40%
Absolute error (V)	Random noise (with mean 0 and standard deviation V) added to the reference log(SUV) values	0.25	0.50
Absolute bias (B)	Set of observed SUV values are generated by adding random noise (with mean B and standard deviation V) to the reference log(SUV) values	0.25 -0.25	0.50 -0.50
Correlation (ρ)	Correlation of noise added to baseline and mid-therapy reference SUV values for the same person to create observed values. Random noise is generated by a bivariate normal distribution.	0.1	0.1

Table 2

Power calculations for a multicenter trial estimating the association between percent change in FDG PET SUV and pathologic response to neoadjuvant chemotherapy for patients with locally advanced breast cancer, calculated by simulation of 10,000 datasets with characteristics described in Table 1.

Error conditions for each scenario:	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	Ref. power	**20%/40%**	**20%/40%**	**−20%/−40%**	**0.25/0.50**	**0.25/0.50**	**−0.25/−0.50**
Simulation scenarios:		**pct error**	**pct bias**	**pct bias**	**abs error**	**abs bias**	**abs bias**
75% from primary sites							
- all data analyzed (no selection by SUV0)	0.92	0.71	0.78	0.60	0.78	0.78	0.79
- enriched sample (observed SUV0 ≥ 3)	0.87	0.70	0.79	0.54	0.73	0.76	0.70
25% from primary sites							
- all data analyzed (no selection by SUV0)	0.92	0.54	0.69	0.36	0.67	0.66	0.68
- enriched sample (observed SUV0 ≥ 3)	0.87	0.57	0.71	0.32	0.62	0.65	0.57

SUV0 is SUV of first scan

N=100 for unselected design (200 scans), N < 100 for enrichment design (200 scans)