Published in final edited form as:

Acad Radiol. 2017 August ; 24(8): 1036–1049. doi:10.1016/j.acra.2017.03.002.

Beyond Correlations, Sensitivities, and Specificities: A Roadmap for Demonstrating Utility of Advanced Imaging in Oncology Treatment and Clinical Trial Design

Erich P. Huang, PhD*,

Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, NIH. Bethesda, MD

Frank I. Lin, MD, and

Cancer Imaging Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, NIH. Bethesda, MD

Lalitha K. Shankar, MD, PhD

Cancer Imaging Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, NIH. Bethesda, MD

Abstract

Despite the widespread belief that advanced imaging should be very helpful in guiding oncology treatment decision and improving efficiency and success rates in treatment clinical trials, its acceptance has been slow. Part of this is likely attributable to gaps in study design and statistical methodology for these imaging studies. Also, results supporting the performance of the imaging in these roles have largely been insufficient to justify their use within the design of a clinical trial or in treatment decision-making. Statistically significant correlations between the imaging results and clinical outcomes are often incorrectly taken as evidence of adequate performance. Assessments of whether the imaging can outperform standard techniques or meaningfully supplement them are also frequently neglected. This manuscript provides guidance on study designs and statistical analyses for evaluating the performance of advanced imaging in the various roles in treatment decision guidance and clinical trial conduct. Relevant methodology from the imaging literature is reviewed; gaps in the literature are addressed using related concepts from the more extensive genomic and in vitro biomarker literature.

1 Introduction

Advanced imaging, namely novel imaging methods or standard of care imaging used in novel clinical contexts, has proven valuable in guiding oncology treatment, clinical trial design, and drug development. Endpoints of many Phase II trials such as overall response

*Corresponding author: erich.huang@nih.gov, 9609 Medical Center Drive, MSC 9735, Bethesda, MD 20892-9735, Phone: (240) 276-6151, Fax: (240) 276-7888.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

rate (ORR) via Response Evaluation Criteria in Solid Tumors (RECIST) [1] were based on anatomic imaging. Changes in tumor metabolism in gastric cancer patients following the initial cycles of chemotherapy as measured by positron emission tomography with [18F]-fluorodeoxyglucose (FDG-PET) can serve as an early measure of response and thus help guide treatment adjustments in these patients after the first few cycles of chemotherapy [2–4]. Possible roles of imaging are summarized in Table 1; definitions were borrowed from literature on clinical trial design and in vitro and genomic biomarkers in order to maintain a common vocabulary between these research communities.

Although the imaging literature currently contains many promising results, the incorporation of advanced imaging into patient care and clinical trial design has not been ubiquitous [5]. In addition to the high cost of the investigation and development of imaging agents [6] and relatively high regulatory barriers for using imaging in a clinical study (e.g. Investigational New Drug [IND] application requirements [7]), most of the results so far, although important, are insufficient evidence of the utility of advanced imaging in guiding disease management or as an integral part of the design of a treatment clinical trial. Many clinical imaging studies indicate statistically significant correlations and high sensitivities and specificities, which by themselves neither necessarily translate into adequate performance in a particular role nor sufficiently justify the extra effort and resources to adopt the imaging.

This manuscript serves as a roadmap for research to advance an imaging procedure to the point where it can justifiably be used in guiding disease management or as an integral part of the design of an oncology treatment clinical trial. In general, this entails evaluating the performance of the imaging through a sequence of progressively larger and more definitive clinical imaging studies.

Gatsonis and Hillman and Gatsonis define a framework of such a sequence that draws upon parallels with Phase I, Phase II, and Phase III clinical trials of oncology treatments [8, 9]. The first studies, which Gatsonis and Hillman and Gatsonis call “Phase I”, involve *discovery* and include those focusing on standardization of the image acquisition and processing protocol and evaluation of metrological aspects such as test-retest repeatability. Next come *introductory studies* evaluating the association between imaging measurements and clinical outcomes (e.g. association between baseline FDG standardized uptake value [SUV] and overall survival [OS] in gastric cancer patients undergoing chemotherapy) or of the ability of the imaging to facilitate detection of clinically relevant characteristics such as metastases (“Phase II” according to Gatsonis and Hillman and Gatsonis). Then come larger *mature studies* directly and definitively evaluating the performance of the imaging in its intended role, typically in a multi-institutional setting (“Phase III”); for example, such a study of FDG-PET in assessment of early response to chemotherapy in gastric cancer patients may involve a randomized study comparing the survival of early non-responders (i.e. those not showing appreciable decreases in FDG SUV after the initial cycles of chemotherapy) that switch treatments to that of early non-responders that do [10]. So far, these mature studies have not received much attention in the imaging literature. Not only have comparatively few of them been performed to date, but also, the imaging literature contains substantial gaps in methodology regarding how to design and execute such studies. However, related concepts

have been addressed more extensively in the genomic and in vitro biomarker literature and can be adapted to imaging to bridge these gaps.

The sequences of clinical studies needed to evaluate imaging in each of its possible roles are presented, with relevant designs and statistical analyses. The emphasis will be more on guiding the reader through this process rather than the statistical details of the analysis methodology or designs of such studies. Presenting novel study designs or statistical methodology is not the purpose of this manuscript either. Study designs and analysis techniques from the imaging literature are reviewed whenever such methodology is available. Otherwise, analogous concepts from genomic and in vitro biomarker literature are presented with proper adaptations to imaging.

Note that the sequence of studies depends heavily on the intended use; those appropriate for showing the utility of imaging in one role are generally not for evaluating its utility in another. Although the same imaging procedure can have utility in multiple scenarios (e.g. one used for early response assessment may also be useful as the basis of a trial endpoint), proper evaluation should occur separately for each potential role. Additionally, results on the fitness of the use of the imaging in one patient population should not be generalized to another; evaluations should occur separately for each target population.

It is assumed that prior to the conduct of any of the studies in this manuscript, acquisition and processing protocols have already been sufficiently standardized and metrological performance has been shown to be adequate. This manuscript will not focus on these topics. A review of study designs and statistical methodology for evaluating metrological aspects can be found in the Quantitative Imaging Biomarkers Alliance Metrology series [11–15].

2 Diagnosis and Staging

Initial assessments often involve estimation of the proportion of cases in which a human observer correctly identifies the primary tumor through visualization of the imaging data. For example, in Hannah et al, nuclear medicine physicists reviewed the pre-surgical FDG-PET scans of patients with clinically diagnosed head and neck squamous cell carcinoma (HNSCC) and indicated where the primary tumor was located [16]. Sensitivities, specificities, and positive and negative predictive values (PPV and NPV respectively) of human observers in visually detecting nodal involvement and metastases are also often assessed. In the ongoing American College of Radiology Imaging Network (ACRIN) 6685 study, radiologists reviewed FDG-PET scans for 292 HNSCC patients prior to surgery to identify nodal involvement in order to estimate the NPV of visual nodal involvement detection [17]. Target sensitivities, specificities, or positive and negative predictive values will vary between modalities and disease types, but should reflect a trade-off between ideal expectations and realistic performance of the imaging; for instance, in the ACRIN 6685 study, investigators target an NPV of 90%, a very high value but not an unrealistic one given previous results.

Subsequent evaluations may include demonstrating superior sensitivity, specificity, PPV, or NPV of the imaging in detecting the primary tumor, nodal involvement, or metastases

compared to standard methods. Other imaging procedures such as anatomic imaging with computed tomography (CT) and magnetic resonance imaging (MRI), both of which are currently used for numerous disease types, may be considered standard methods in some cases. For example, Adams et al and Laubenbacher et al compare the sensitivity and specificity of lymph node metastasis detection in HNSCC based on pre-surgical FDG-PET to those based on CT and MRI [18, 19].

However, high sensitivities, specificities, or positive and negative predictive values of tumor, metastasis, or nodal involvement detection based on the imaging are insufficient evidence of its utility in guiding disease management, as are superior values of these metrics relative to those associated with standard methods. Subsequent studies should focus on assessing the impact of the imaging in terms of disease management decision and consequently patient outcome. Examining the impact of imaging on disease management decisions may involve assessing the proportion of patients for whom the imaging changes the disease stage, or treatment decision compared to a standard method. For example, Ha et al performed a retrospective study in which 36 patients with previously untreated HNSCC underwent staging with CT or MRI and with FDG-PET/CT; they computed the proportion of early-stage and of late-stage patients whose treatment plans were modified given FDG-PET results [20]. Alternatively, in some studies, investigators assess the level to which the imaging agrees with that of a standard method. Abramnyuk et al describe inferences on the weighted κ statistic between the stages according to the two techniques for this purpose [21]. Low values of κ , traditionally accepted as 0.4 or less [22], are evidence that the stages according to the two techniques do indeed differ.

Assessing the therapeutic impact of the imaging, namely the improvement in patient outcome resulting from incorporating the imaging into diagnosis and staging, is also strongly recommended. Lu and Gatsonis present the Paired Design to assess the therapeutic impact of the imaging. All patients undergo both the new imaging-based procedure and a standard procedure. Patients for whom both procedures agree simply undergo treatment according to their stages. Patients for whom the procedures disagree are randomized to one of two arms: one in which the treatment is dictated by the results of the new imaging-based procedure and one in which it is dictated by the results of the standard procedure. The clinical outcomes (e.g. response rate, OS or PFS, or OS or PFS past a prescribed landmark time point) are compared between these groups of patients [23]. A schematic of this design is given in Figure 1.

Disease stage does indeed have prognostic value (see §3). However, the evaluation of imaging in a prognostic context is different from its evaluation for diagnosis and staging; thus, imaging-based prognostic markers will be discussed separately.

3 Imaging-Based Prognostic Markers

In oncology, prognosis indicates the expected outcome on standard therapy. For example, the area of residual tumor following surgical resection based on MRI is prognostic in medulloblastoma patients. Those with less than 1.5 cm² of residual tumor and no metastases are considered average-risk and are expected to have excellent outcomes on

chemoradiotherapy; the others are considered high-risk, and their outcomes on chemoradiotherapy are less promising [24, 25].

Such prognostic markers would be useful in identifying lower-risk patients whose eventual outcomes on the standard treatment are good enough to warrant forgoing potentially more intensive and toxic alternative treatments that may only provide marginal benefit [26]. Prognostic markers can thus be useful for selecting patients for a clinical trial assessing the non-inferiority of reduced therapy. For example, the ACNS0331 study aims to assess non-inferiority of outcomes resulting from reduction of the intensity of radiation therapy in average-risk medulloblastoma patients [27]. Similarly, prognostic markers may also be useful in identifying higher-risk patients whose outcomes are poor enough to warrant alternative treatments, as well as identifying those that may be good candidates for trials evaluating more intensive therapy regimens.

3.1 Assessing Associations with Clinical Outcomes

Preliminary assessments often involve estimation of the association between imaging measurements and clinical outcomes (e.g. progression-free survival [PFS] or OS) among patients undergoing standard therapy for their stage. These patients may come from the control arm of a randomized trial; for example, the ongoing ACRIN 6697 study is assessing the association of baseline tumor hypoxia as measured using PET with [18F]-fluoromisonidazole (FMISO-PET) [28, 29] with OS in 50 non-small cell lung cancer (NSCLC) patients randomized to the standard radiation therapy arm [30]. Data may also be gathered retrospectively from completed clinical trials in order to bypass the lengthy follow-up times required for a sufficient number of progressions or deaths [31].

Statistical methods for assessing associations between the imaging and outcomes include Cox regression for time-to-event outcomes and logistic regression or receiver operating characteristic (ROC) curve analysis for binary outcomes (e.g. response). If multiple imaging measurements are being studied, these associations are often assessed individually, with corrections for multiple testing (e.g. Benjamini-Hochberg procedure [32]). Associations between the imaging variables and the clinical outcome after adjusting for standard prognostic variables may also be assessed through multivariate Cox or logistic regression. For example, Okamoto et al demonstrate a statistically significant association between FDG maximum SUV and 6-month recurrence following resection in pancreatic cancer patients through a multivariate logistic regression model including FDG maximum SUV, age, performance status, and carbohydrate antigen (CA19-9) [33].

However, these methods are insufficient to provide evidence that use of the imaging to guide disease management decision has a meaningful impact on patient outcome. Subsequent studies (described in §3.3) are needed to directly show that either a very high proportion of lower-risk patients actually do have favorable outcomes on standard therapy, or that treatment decisions using information from the imaging result in improved patient outcomes over standard of care treatment given standard prognostic variables.

3.2 Defining Risk Groups

In practice, the imaging, with or without standard prognostic variables, would be used to classify a patient into one of two or more risk groups, namely subsets of patients with similar expected outcomes on standard therapy. The definitions of these risk groups given the imaging and standard clinical variables need to be fully determined as treatment decisions and clinical trial execution is dictated by a patient's risk group assignment. For example, higher-risk patients may be those with FMISO maximum SUV exceeding a prescribed cutoff whereas the others are lower-risk. Alternatively, risk group definitions can be based on risk scores involving multiple imaging and standard prognostic variables. For example, Aerts et al propose a score equal to a weighted sum of four quantitative imaging variables derived from pre-treatment CT scans related to tumor density, tumor shape, and within-tumor heterogeneity [34]. Risk group categorization would be based on whether this score exceeds a prescribed cutoff.

For risk group categorization based on a single imaging variable, the cutoff is sometimes set equal to the median value or some other quantile of the imaging measurement among patients undergoing standard therapy. However, this may not be ideal as it ignores clinical outcome. One approach may be to identify the cutoff leading to the greatest separation in clinical outcome (e.g. largest difference in OS among patients whose imaging measurements exceed the cutoff versus those whose measurements do not). For example, a cutoff for FMISO maximum SUV can be identified through the following procedure for multiple possible thresholds: each patient is categorized as higher-risk or lower-risk based on whether their SUV exceeds the threshold, and a log-rank test of the OS of the risk groups is performed. The threshold associated with the lowest log-rank test p-value is selected as the cutoff [31].

Meanwhile, if multiple variables are to be used, the specific imaging and standard prognostic variables to be included in the computation of the risk score and an explicit mathematical formula combining the values of these variables need to be determined in addition to cutoffs in scores differentiating the risk groups. A review of the numerous existing statistical techniques that can be used for this step (e.g. forward stepwise selection for variable selection, Cox regression for determining the exact mathematical formula of the score) can be found in Hastie, Tibshirani, and Friedman [35].

Assessment of whether these different risk groups actually have significantly differing outcomes should also occur. The data used for this should be completely separate from that used for derivation of the risk score or the risk group definitions as using the same data for both usually leads to overly optimistic estimates of the magnitude of the differences in outcome among the risk groups [36, 37]. Patients in the new data set would be categorized into risk groups, without any alterations to cutoffs, weights, included variables, and mathematical forms. Outcomes are then compared between risk groups. Molinaro et al and Hastie, Tibshirani, and Friedman review resampling and cross-validation techniques that are also useful here [35, 38].

3.3 Demonstrating Usefulness of the Imaging in Guiding Treatment Decision

Demonstrating that the imaging is useful in guiding treatment decision can entail showing that a very high proportion of lower-risk patients actually do have favorable outcomes on standard therapy. Low rates of recurrence, progression, or death by a landmark time point among patients labeled as lower-risk according to the imaging may be assessed prospectively in a study in which these patients uniformly receive standard therapy [39]. Again, a target rate is context-dependent and should reflect a trade-off between ideal expectations and realistic performance. As an example, Packer et al showed that the 5-year survival rate of average-risk medulloblastoma patients undergoing radiation therapy followed by adjuvant chemotherapy was 86% [25].

An alternative is to demonstrate that treatment decisions based on the imaging lead to improved patient outcomes. The Paired Design described in §2 is useful for this assessment; although this design has rarely been used for imaging so far, it has been used in analogous trials of genomic markers such as the Microarray in Node Negative Disease May Avoid Chemotherapy Trial (MINDACT) [40]. Another option is a Modified Marker Strategy Design depicted in Figure 2. For the Modified Marker Strategy Design, treatment decisions based on the risk groups must be specified prior to the study; for example, lower-risk patients according to FMISO-PET (FMISO maximum SUV less than a prescribed threshold) undergo standard radiation therapy whereas higher-risk patients undergo intensified radiation therapy. All patients undergo a baseline scan, and, based on the imaging plus standard prognostic variables, are categorized into a risk group. Patients for whom the treatment decision given the imaging is the same as standard of care go off study. All other patients are randomized to receive treatment based on imaging plus standard clinical variables or standard of care treatment [26]. Investigators then test for improved clinical outcome in the experimental arm through a log-rank test.

4 Imaging-Based Assays of Predictive Biomarkers

A predictive biomarker is a baseline characteristic that indicates response or lack thereof to a particular class of therapies [41]. For example, estrogen receptor (ER) expression is a predictive biomarker indicating whether a breast cancer patient will benefit from endocrine therapy relative to standard chemotherapy alone [42–45]. Predictive biomarkers are thus not only useful for treatment selection, but also to enrich clinical trials by restricting entry to only biomarker-positive patients (i.e. those with the biomarker present or higher levels of biomarker expression) expected to benefit from the treatment under investigation. This results in a significant reduction of the sample size while maintaining power if only a minority of all patients is expected to be biomarker-positive. The sample size would be reduced by approximately a factor of 16 is expected when 25% of the patient population are biomarker-positive and by approximately a factor of 100 is expected when only 10% are [46, 47].

Imaging is enticing as a predictive biomarker assay due to its reduced invasiveness relative to in vitro assays and its ability to simultaneously assay multiple lesions, allowing easier assessment of biomarker heterogeneity. Investigations of some imaging procedures such as PET with [¹⁸F]-fluoroestradiol (FES-PET) for ER expression and perfusion CT for blood

flow and volume in advanced pancreatic neuroendocrine tumors are currently underway [48].

4.1 Concordance with In Vitro Assay Measurements and Associations with Clinical Outcomes

Preliminary studies can involve assessments of the concordance between imaging and more standard in vitro assay measurements of the underlying predictive biomarker, such as inferences on the Kendall tau rank correlation coefficient between FES SUV and quantitative immunohistochemistry measurements of ER expression [49]. They may also involve assessment of the association between the imaging and clinical outcomes among patients uniformly treated with an investigational therapy linked to the biomarker (e.g. endocrine therapy for ER). In these studies, all patients undergo baseline imaging prior to initiation of treatment and are followed up for outcome. Associations between baseline imaging measurements and outcomes are assessed through Cox regression or log-rank tests for time-to-event outcomes such as PFS or Mann-Whitney U tests or Fisher's exact tests for binary outcomes such as response (i.e. tests of whether measurements differ between responders versus non-responders) [50].

In order to provide evidence that the predictive biomarker is actually useful in guiding treatment decision, subsequent studies should assess qualitative interactions between treatment and biomarker-positivity, namely that outcomes on the investigational treatment are superior to those on standard therapy among biomarker-positive patients, but among biomarker-negative patients, outcomes on the investigational treatment are similar or inferior [51]. A qualitative interaction is depicted in Figure 3 and discussed further in §4.3. But first, a definition of biomarker-positivity according to the imaging must be prospectively specified. §4.2 describes how to accomplish this.

4.2 Defining Biomarker-Positivity Based on the Imaging

Definitions of biomarker-positivity may be based on whether a quantitative imaging measurement exceeds a pre-specified cutoff (e.g. breast cancer patients with a baseline FES SUV of 2 or higher [50]) or whether a particular characteristic is present on the imaging. Cutoffs may be selected from preliminary studies described above; one possible approach is to select a cutoff such that the difference in clinical outcomes among biomarker-positive versus biomarker-negative patients is maximized, similar to what Dehdashti et al do [50]. Such a cutoff could also be adjusted based on whether incorrectly labeling a biomarker-negative patient as biomarker-positive is preferable, which could happen if the investigational treatment has low toxicity and presents little additional inconvenience, or vice versa.

4.3 Relative Efficacy of Treatments as a Function of Biomarker-Positivity

Qualitative interactions between treatment and biomarker-positivity may be assessed through designs such the Marker by Treatment Interaction Design depicted in Figure 4, which have been used for studies of in vitro and genomic assays of predictive markers [52]. Patients enrolled in a prospective randomized controlled trial of the investigational treatment versus standard therapy undergo baseline imaging prior to initiation of therapy, and then are

randomized to a treatment arm independently of the results of the imaging. A variant of this design, depicted in Figure 5, involves randomization being stratified among biomarker-positive and biomarker-negative patients according to the imaging. This variant avoids a reduction in sample size for the evaluation of the imaging by guaranteeing that all patients undergo the imaging. However, this variant also requires that the imaging procedure be ready by the time randomization is to occur, whereas the former one simply requires it to be ready before closure to accrual [26].

Gail and Simon review statistical techniques for testing for a qualitative interaction between treatment and biomarker-positivity [51]. A simple statistical interaction (i.e. the difference in efficacy between the two treatments is unequal among biomarker-positive and among biomarker-negative patients) is insufficient. In a scenario similar to that depicted in Figure 6, in which the investigational treatment is superior in both biomarker-positive and biomarker-negative patients but the difference in efficacy is significantly larger among the former, the interaction between treatment and biomarker positivity is still statistically significant, but the imaging is not useful for guiding treatment selection; based on this data, all patients should receive the investigational treatment.

4.4 Practical Considerations

Studies using the Marker by Treatment Interaction Designs may be difficult to execute in practice. Randomized controlled treatment trials in which such a design could be incorporated are often too small; the required sample sizes are often much larger than those required to detect treatment effects [53]. Plus, randomized controlled treatment trials of the therapy associated with the predictive biomarker would have already been completed in many cases as research on imaging-based predictive biomarker assays often lag behind drug development [5]. Simon, Paik, and Hayes describe retrospective analysis of completed randomized studies investigating the efficacy of the associated class of treatments to evaluate in vitro predictive biomarker assays, which could also circumvent the sample size problem by combining data from multiple trials [52, 54]. However, adapting this approach to imaging requires that the imaging procedure was sufficiently widely available at the time most of these trials were performed, which is often not the case. Imaging data would need to be available for a sufficiently large portion of patients from these completed trials in order to ensure adequate power and to involve a reasonably representative subset of the target patient population.

Also, widespread availability of in vitro assays of the predictive biomarker and ubiquitous use of the class of treatments associated with the predictive biomarker can make a randomized study to test for qualitative interactions difficult to justify. For example, immunohistochemistry is already widely used to assay ER expression in breast cancer; most researchers and clinicians would not be willing to disregard this information, nor would they be likely to withhold treatment with endocrine therapy, which is currently used to treat many breast cancer cases. Furthermore, existence of a widely used in vitro assay necessitates demonstrating that the imaging has added value in terms of whether basing treatment decision on the imaging leads to improved clinical outcome or, given the reduced invasiveness of the imaging, whether the resulting clinical outcomes are non-inferior.

A modified marker-strategy design similar to that described in §3.3 may also be useful in evaluating an imaging-based predictive biomarker assay in light of these considerations. First, appropriate treatment decisions given the imaging (e.g. those with FES SUV exceeding 2 receive endocrine therapy whereas the others receive standard therapy) and given the standard assay (e.g. those with Allred scores of 7 or 8 receive endocrine therapy whereas the others receive standard therapy [55]) need to be identified. Prior to treatment, all enrolled patients would undergo the imaging and the standard assay. Patients for whom treatment decision given the imaging coincides with that given the standard assay go off study. All other patients are randomized to the control arm, in which therapy is dictated by the standard assay, or to the experimental arm, in which therapy is dictated by the imaging. The clinical outcomes in the two arms would be compared through log-rank tests or two-sample t-tests.

5 Imaging-Based Pharmacokinetic and Pharmacodynamic Markers

Often, evaluation of imaging-based pharmacokinetic and pharmacodynamic markers consists of assessment of the concordance of the imaging measurement with an independent measure of the underlying characteristic of interest. For example, Bhattacharyya et al perform inferences on the Kendall tau rank correlation coefficient between [⁸⁹Zr]-panitumumab uptake and epidermal growth factor receptor expression according to immunohistochemistry [56]. Such assessments usually take place in pre-clinical studies. Evaluations may also include assessment of quantitative changes in imaging measurements from baseline after initiation of treatment and whether these changes substantially differ from zero. As a secondary objective in the Sarcoma Alliance for Research through Collaboration (SARC) 022 trial, investigators performed a Wilcoxon signed-rank test of a decrease in FDG SUV from baseline following initiation of treatment with the inhibitor of the insulin-like growth factor 1 receptor linsitinib in patients with gastrointestinal stromal tumors [57]. These assessments typically occur within a Phase I or Phase II treatment trial in which patients receiving a therapeutic dose of the drug undergo baseline imaging and follow-up imaging at a pre-specified time point during treatment.

Pharmacokinetic and pharmacodynamic markers may take on other roles later. Subsequent studies should be geared toward these roles. For example, pharmacokinetic markers such as [⁸⁹Zr]-panitumumab uptake may evolve into predictive biomarker assays; studies described in §4 should then be performed. Pharmacodynamic markers may evolve into clinical trial endpoints, in which case studies described in §7.1 are recommended [58].

6 Response Assessment

Response assessments through imaging are useful in several contexts. Interim response assessment, namely examination of changes in the imaging after the first few cycles of therapy, can guide subsequent treatment; those showing response at this point may proceed with the remainder of the therapy regimen whereas others may switch to an alternative regimen. For example, gastric cancer patients that do not show early response to pre-operative chemotherapy (less than 35% decrease in FDG SUV from baseline after one cycle) may switch to a more intensive treatment regimen involving salvage chemotherapy in

addition to surgery and post-surgical salvage chemotherapy [10]. These response assessments may also serve as the basis of treatment clinical endpoints. For example, as mentioned in §1, endpoints such as ORR according to anatomic imaging have been used for numerous Phase II trials and complete metabolic response according to FDG-PET have been used in Phase II trials of radiochemotherapy in cervical cancer [59, 60]. Assessment of progression via RECIST for determining PFS has also served as the basis of some Phase III trials [5]. Endpoints based on response assessments in this manner are often ascertainable earlier than more definitive ones such as OS; particularly for disease types with lengthy OS and PFS, use of these alternative endpoints can thus help reduce costs and expedite the drug development process [61–63].

6.1 Assessing Associations Between Changes in the Imaging and a More Definitive Endpoint

Initial evaluations often involve assessments of associations between a more definitive endpoint (e.g. OS) and changes in the imaging from baseline. Associations between changes in the imaging (e.g. percent change in FDG SUV) and time-to-event outcomes can be assessed through Cox regression or log-rank tests; those between changes in the imaging and binary outcomes such as histopathological response can be assessed through Mann-Whitney U tests or Fisher's exact tests. For example, Ott et al compare outcomes in patients showing a decrease in FDG SUV of 35% or more to those not showing such a decrease in a cohort of 65 patients with adenocarcinomas of the esophagogastric junction undergoing treatment with cisplatin plus leucovorin and fluorouracil and then surgery [64].

6.2 Defining Response

Prior to the studies described in §6.3, §6.4, and §6.5, response categories must be defined. They may be based on existing sets of response criteria such as RECIST, Response Assessment in Neuro-Oncology (RANO) criteria [65], or PET Response Criteria in Solid Tumors (PERCIST) [66]. However, some existing criteria may not have a basis in empirical data or have gone through the evaluations described in §6.1.

Response may be also based on whether changes in the imaging exceed a specific threshold (e.g. decrease in FDG SUV by more than 35% or absence of metabolic activity in a lesion or reduction in intensity to one similar to that of the blood pool according to FDG-PET [67]). Data from the initial studies described above can be used to define response. A cutoff in the quantitative changes in imaging measurements differentiating response versus non-response could be selected in order to optimize sensitivity and specificity in differentiating ultimate responders versus ultimate non-responders; Weber et al use this technique to define interim response as a decrease in FDG SUV of more than 35% [4].

Definitions of response categories may also be derived through modifications of existing criteria to improve upon the latter or to accommodate advancements in imaging technology and cancer treatments. Aspects that were previously highly subjective or variable may be specified in greater detail; for example, RECIST 1.0 involved a standardization of numerous aspects of World Health Organization (WHO) criteria whereas RECIST 1.1 involved more detailed guidance on new lesions and progression of non-target lesions compared to RECIST

1.0 [1]. Additional imaging or clinical variables may be incorporated into the existing response criteria; for example, in RECIST 1.1, lymph nodes were added to response categories from RECIST 1.0 [1] and in RANO criteria, corticosteroid use and clinical deterioration were added to the Macdonald criteria definition of progression and changes in non-enhancing lesions were added to Macdonald criteria response categories [68]. Criteria may be simplified (e.g. using unidimensional measurements in RECIST 1.0 instead of bidimensional measurements in WHO criteria or basing the response assessment on five target lesions in RECIST 1.1 instead of ten in RECIST 1.0 [1]).

Concordance between response assessments according to the modified criteria and those according to the original ones should be evaluated; for example, Bogaerts et al examine the percent agreement between response assessments according to RECIST 1.0 and according to RECIST 1.1 [69]. The association between a more definitive endpoint (e.g. OS) and response assessments according to the modified criteria should also be shown to be at least as strong as that between the endpoint and assessments according to the original criteria.

6.3 Assessing Outcomes After Switching Treatments Based on Interim Response Assessments

The true test of the utility of the imaging in interim response assessment is whether treatment modification based on changes in the imaging lead to improved clinical outcome. One possible study design is the one used in the A021302 trial; 324 gastric cancer patients underwent baseline FDG-PET imaging and then one cycle of treatment with epirubicin, oxaliplatin or cisplatin, and capecitabine or fluorouracil before follow-up FDG-PET imaging. Those who experienced an interim response (35% or more decrease in FDG SUV from baseline) went off study whereas those that did not were randomized either to undergo surgery plus post-surgery chemotherapy and radiotherapy or to receive two cycles of salvage chemotherapy with docetaxel and irinotecan followed by surgery and then three additional cycles [3, 10]. Investigators then test for improved outcome among those switching treatments through a log-rank test of the OS or PFS between the two arms.

However, this study design can only answer the question whether switching treatments leads to improved outcome among those not experiencing interim response; switching treatments may also lead to improved outcome among those experiencing interim response, in which case, the imaging will not be useful in guiding treatment decision. An alternative study design is one similar to the Marker by Treatment Interaction Design described in §4.3 in which early responders remain on study and are also randomized to either remain on their current therapy regimen or to switch. Statistical techniques from Gail and Simon described in §4.3 can be used to test for a qualitative interaction between early response and switching treatments, namely that while early non-responders that switch treatments experience improved outcome, early responders that switch treatments experience similar or worse outcomes than those that do not [51]. Although this study design requires significantly more patients, it allows for the comparison of the improvement in outcome from switching treatments among both early responders and early non-responders.

6.4 Phase II Trial Endpoints

When imaging-based response assessments are used as an endpoint for a Phase II clinical trial, they are primarily intended to guide decisions as to whether to move forward to a Phase III trial of the treatment [70]. The true test of such an endpoint is whether it moves forward truly efficacious treatments to a subsequent Phase III study, namely whether a large proportion of Phase II trials that are positive based on the endpoint are also associated with positive Phase III trials (i.e. a high PPV of decisions to move forward to Phase III based on the imaging) [71]. Demonstration of a sufficiently high PPV requires multiple Phase II trials of relevant agents across a variety of disease types and the associated Phase III trials. For example, Ratain performed such an analysis of Phase II trial positivity according to response rate in predicting Phase III trial positivity for treatments for various disease types, using data from Goffin et al [72, 73]. These analyses can also be done using Phase II/III trials, or retrospectively using data from completed Phase II and Phase III trials.

6.5 Phase III Trial Endpoints

When imaging-based response assessments are used as the basis of an endpoint for a Phase III treatment trial, then such an endpoint should be shown to be a surrogate for a more definitive one such as OS, as Phase III trials are intended to gather evidence for drug approval and provide recommendations regarding changes in treatment or practice among the medical community [74]. In other words, effects of the treatment on such an endpoint need to capture a sufficiently large proportion of the effects of the treatment on the more direct endpoint [75, 76]. Between-arm differences in such an endpoint (e.g. control versus treatment arm PFS hazard ratios) need to be shown to have strong concordance with between-arm differences in a more direct endpoint (e.g. control versus treatment arm OS hazard ratios) [75, 77]. Assessing associations between the two endpoints is insufficient in this case as statistically significant correlations do not necessarily mean that the new endpoint captures between-arm differences in the more definitive one [70, 78].

Here, meta-analyses of randomized trials where both endpoints are measured are required; acquisition of data may occur retrospectively, where researchers collect available imaging, treatment, and clinical outcome data from completed randomized trials [61, 63, 75]. Methodology for assessing such a surrogate endpoint is reviewed in Sargent et al [75]. Korn et al and Buyse et al describe R^2 -type statistics that can be used to assess the proportion of the variability in between-arm differences in the more direct endpoint that is explained by between-arm differences in the surrogate endpoint. This quantity can take on values between zero and one, with values closer to one providing strong evidence that treatment effects on the surrogate endpoint do indeed capture those on the definitive endpoint [79, 80]. Initially, Prentice proposed that the value of this quantity should be one, but that has been considered too strict to be practical [80]. Instead, Freedman et al and Korn et al propose that a value of this quantity exceeding 0.5 is evidence of the viability of the imaging-based endpoint as a surrogate [77, 81].

7 Discussion

Prior to the incorporation of advanced imaging into patient disease management and clinical trial design, it should undergo a series of clinical studies evaluating its performance in its intended role. Proper design and execution of such studies are crucial in producing convincing evidence of the utility of the imaging in its intended role, especially given the major expenses incurred by the fact many imaging modalities under investigation are not standard of care. The intended use of the imaging and its target population should be carefully specified prior to the initiation of such studies, as should a clear hypothesis that the study is intended to address. The design and statistical analysis should be appropriate for the hypothesis, the intended role, and the target population. The use of terminology should be accurate and consistent.

But also, these studies should also aim to justify the extra effort and resources needed for the imaging. The study should be designed and the analysis planned in a way that demonstrates not only adequate performance of the imaging in the intended role, but also improved patient outcome and clinical trial efficiency resulting from incorporation of the imaging. For instance, imaging measurements, together with standard prognostic variables, should be shown to predict clinical outcome in patients undergoing standard treatment more accurately than the latter can alone, and treatment decisions based on the imaging plus the standard prognostic variables should be shown to lead to improved clinical outcome relative to decisions based on the latter only.

This manuscript provides guidance regarding the study design, statistical analyses, and best practices for conducting studies to rigorously evaluate the performance of imaging in these roles. These types of studies are also fertile ground for development of novel study designs and statistical analysis methodology, and research into these problems are another area of future work.

Acknowledgments

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

LIST OF ABBREVIATIONS

ACRIN	American College of Radiology Imaging Network
CT	Computed tomography
ER	Estrogen receptor
FDG	[18F]-fluorodeoxyglucose
FDG-PET	Positron emission tomography with [18F]-fluorodeoxyglucose
FES	[18F]-fluoroestradiol

FES-PET	Positron emission tomography with [18F]-fluoroestradiol
FMISO	[18F]-fluoromisonidazole
FMISO-PET	Positron emission tomography with [18F]-fluoromisonidazole
HNSCC	Head and neck squamous cell carcinoma
MRI	Magnetic resonance imaging
NPV	Negative predictive value
ORR	Overall response rate
OS	Overall survival
PET	Positron emission tomography
PFS	Progression-free survival
PPV	Positive predictive value
RECIST	Response Evaluation Criteria in Solid Tumors
ROC	Receiver operating characteristic
SUV	Standardized uptake value

References

1. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*. 2009; 45(2):228–247. [PubMed: 19097774]
2. Ott K, Fink U, Becker K, et al. Prediction of response to preoperative chemotherapy in gastric carcinoma by metabolic imaging: Results of a prospective trial. *Journal of Clinical Oncology*. 2003; 21(24):4604–4610. [PubMed: 14673049]
3. Lordick F, Ott K, Krause B-J, et al. PET to assess early metabolic response and to guide treatment of adenocarcinoma of the oesophagogastric junction: The MUNICON Phase II trial. *The Lancet Oncology*. 2001; 8(9):797–805.
4. Weber WA, Ott K, Becker K, et al. Prediction of response to preoperative chemotherapy in adenocarcinomas of the esophagogastric junction by metabolic imaging. *Journal of Clinical Oncology*. 2001; 19(12):3058–3605. [PubMed: 11408502]
5. Lin FI, Huang EP, Shankar LK. Beyond correlations, sensitivities, and specificities: Case examples of the evaluation of advanced imaging in oncology clinical trials. *Academic Radiology*. 2017 advance access.
6. Nuun AD. The cost of developing imaging agents for routine clinical use. *Investigative Radiology*. 2006; 41(3):206–212. [PubMed: 16481902]
7. Guidance for clinical investigators, sponsors, and IRBs: Investigational new drug applications (INDs) – Determining whether human research studies can be conducted without an IND. Available at <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM229175.pdf>. Accessed January 3, 2017
8. Gatsonis CA, Hillman BJ. When is the right time to conduct a clinical trial of a diagnostic imaging technology? *Radiology*. 2008; 248(1):12–17. [PubMed: 18566166]
9. Gatsonis C. Design of evaluations of imaging technologies: Development of a paradigm. *Academic Radiology*. 2000; 7(9):681–683. [PubMed: 10987328]

10. Alliance for Clinical Trials in Oncology. FDG-PET directed treatment in improving response in patients with locally advanced stomach or gastroesophageal junction cancer. Available at <https://clinicaltrials.gov/ct2/show/NCT02485834>. Accessed October 1, 2015
11. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Statistical Methods in Medical Research*. 2015; 24(1):9–26. [PubMed: 24919826]
12. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment. *Statistical Methods in Medical Research*. 2015; 24(1):27–67. [PubMed: 24919831]
13. Obuchowski NA, Reeves AP, Huang EP, et al. Quantitative imaging biomarkers: A review of statistical methods for computer algorithm comparisons. *Statistical Methods in Medical Research*. 2015; 24(1):68–106. [PubMed: 24919829]
14. Obuchowski NA, Barnhart HX, Buckler AJ, et al. Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example. *Statistical Methods in Medical Research*. 2015; 24(1):107–140. [PubMed: 24919828]
15. Huang EP, Wang X-F, Roy Choudhury K, et al. Meta-analysis of the technical performance of an imaging procedure: Guidelines and statistical methodology. *Statistical Methods in Medical Research*. 2015; 24(1):141–174. [PubMed: 24872353]
16. Hannah A, Scott AM, Tochon-Danguy H, et al. Evaluation of 18F-fluorodeoxyglucose positron emission tomography and computed tomography with histopathologic correlation in the initial staging of head and neck cancer. *Annals of Surgery*. 2002; 236(2):208–217. [PubMed: 12170026]
17. American College of Radiology Imaging Network. ACRIN Protocol 6685: A multicenter trial of FDG-PET/CT staging of head and neck cancer and its impact on the N0 neck surgical treatment in head and neck cancer patients. Available at http://www.acrin.org/6685_protocol.aspx. Accessed September 11, 2015
18. Adams S, Baum RP, Stuckensen T, et al. Prospective comparison of 18F-FDG PET with conventional imaging modalities (CT, MRI, US) in lymph node staging of head and neck cancer. *European Journal of Nuclear Medicine*. 1998; 25(9):1255–1260. [PubMed: 9724374]
19. Laubenbacher C, Saumweber D, Wagner-Manslau C, et al. Comparison of fluorine-18-fluorodeoxyglucose PET, MRI, and endoscopy for staging head and neck squamous-cell carcinomas. *Journal of Nuclear Medicine*. 1995; 36(10):1747–1757. [PubMed: 7562038]
20. Ha PK, Hdeib A, Goldenberg D, et al. The role of positron emission tomography and computed tomography fusion in the management of early-stage and advanced-stage primary head and neck squamous cell carcinoma. *Archives of Otolaryngology – Head and Neck Surgery*. 2006; 132(1):12–16.
21. Abramiyuk A, Appold S, Zophel K, et al. Quantitative modifications of TNM staging, clinical staging, and therapeutic intent by FDG-PET/CT in patients with non-small cell lung cancer scheduled for radiotherapy – A retrospective study. *Lung Cancer*. 2012; 78(2):148–152. [PubMed: 22922126]
22. Landis JR, Koch GG. The measurement of observer agreement from categorical data. *Biometrics*. 1977; 33(1):159–174. [PubMed: 843571]
23. Lu B, Gatsonis C. Efficiency of study designs in diagnostic randomized clinical trials. *Statistics in Medicine*. 2013; 32(9):1451–1466. [PubMed: 23071073]
24. Zeltzer PM, Boyett JM, Finlay JL, et al. Metastasis stage, adjuvant treatment, and residual tumor are prognostic factors for medulloblastoma in children: Conclusions from the Children’s Cancer Group 921 randomized Phase III study. *Journal of Clinical Oncology*. 1999; 17(3):832–845. [PubMed: 10071274]
25. Packer RJ, Gajjar A, Vezina G, et al. Phase III study of craniospinal radiation therapy followed by adjuvant chemotherapy for newly diagnosed average-risk medulloblastoma. *Journal of Clinical Oncology*. 2006; 24(25):4202–4208. [PubMed: 16943538]
26. Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized Medicine*. 2010; 7(1):33–47. [PubMed: 20383292]
27. Children’s Oncology Group. Protocol ACNS0331: A study evaluating limited target volume boost irradiation and reduced dose craniospinal radiotherapy (18.00 Gy) and chemotherapy in children

with newly diagnosed standard risk medulloblastoma: A Phase III double randomized trial. Available at http://www.skion.nl/workspace/uploads/ACNS0331DOC-versie-28092012_1.pdf. Accessed September 14, 2015

28. Koh W-J, Bergman KS, Rasey JS, et al. Evaluation of oxygenation status during fractionated radiotherapy in human non-small lung cancers using [F-18] fluoromisonidazole positron emission tomography. *International Journal of Radiation Oncology, Biology, and Physics*. 1995; 33(2):391–398.
29. Rajendran JG, Mankoff DA, O'Sullivan F, et al. Hypoxia and glucose metabolism in malignant tumors: Evaluation by [18F]-fluoromisonidazole and [18F]-fluorodeoxyglucose positron emission tomography imaging. *Clinical Cancer Research*. 2004; 10(7):2245–2252. [PubMed: 15073099]
30. American College of Radiology Imaging Network. ACRIN Protocol 6697: Randomized Phase 2 trial of individualized adaptive radiotherapy using during-treatment FDG-PET/CT and modern technology in locally advanced non-small cell lung cancer (NSCLC). Available at http://www.acrin.org/6697_protocol.aspx. Accessed September 16, 2015
31. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer*. 1994; 69(6):979–985. [PubMed: 8198989]
32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society-Series B (Methodological)*. 1995; 57(1):289–300.
33. Okamoto K, Koyama I, Miyazawa M, et al. Preoperative 18[F]-fluorodeoxyglucose positron emission tomography/computed tomography predicts early recurrence after pancreatic cancer resection. *International Journal of Clinical Oncology*. 2011; 16(1):39–44. [PubMed: 20862596]
34. Aerts HJWL, Rios Velazques E, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*. 2014; 5
35. Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition)*. New York: Springer Science and Business Media; 2009.
36. Hilsenbeck SG, Clark GM, McGuire WL. Why do so many prognostic factors fail to pan out? *Breast Cancer Research and Treatment*. 1992; 22(3):197–206. [PubMed: 1391986]
37. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: Ready for clinical use? *Journal of the National Cancer Institute*. 2010; 102(7):464–474. [PubMed: 20233996]
38. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: A comparison of resampling methods. *Bioinformatics*. 2005; 21(15):3301–3307. [PubMed: 15905277]
39. Subramanian J, Simon R. What should physicians look for in evaluating prognostic gene-expression signatures? *Nature Reviews: Clinical Oncology*. 2010; 7(6):327–334.
40. Cardoso F, Piccart-Gebhart M, Van't Veer L, et al. The MINDACT trial: The first prospective clinical validation of a genomic tool. *Molecular Oncology*. 2007; 1(3):246–251. [PubMed: 19383299]
41. Clark GM. Prognostic factors versus predictive factors: Examples from a clinical trial of erlotinib. *Molecular Oncology*. 2008; 1(4):406–412. [PubMed: 19383314]
42. McGuire AH, Dehdashti F, Siegel BA, et al. Positron tomographic assessment of 16 alpha-[18F]-fluoro-17 beta-estradiol uptake in metastatic breast carcinoma. *Journal of Nuclear Medicine*. 1991; 32(8):1526–1531. [PubMed: 1869973]
43. Mintun MA, Welch MJ, Siegel BA, et al. Breast cancer: PET imaging of estrogen receptors. *Radiology*. 1988; 169(1):45–48. [PubMed: 3262228]
44. Mortimer JE, Dehdashti F, Siegel BA, et al. Positron emission tomography with 2-[18F]fluoro-2-deoxy-D-glucose and 16 alpha-[18F]fluoro-17beta-estradiol in breast cancer: Correlation with estrogen receptor status and response to systemic therapy. *Clinical Cancer Research*. 1996; 2(6): 933–939. [PubMed: 9816253]
45. Dehdashti F, Flanagan FL, Mortimer JE, et al. Positron emission tomographic assessment of 'metabolic flare' to predict response of metastatic breast cancer to antiestrogen therapy. *European Journal of Nuclear Medicine*. 1999; 26(1):51–56. [PubMed: 9933662]
46. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*. 2004; 10(20):6759–6763. [PubMed: 15501951]

47. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine*. 2005; 24(3):329–339. [PubMed: 15551403]
48. National Cancer Institute. Ziv-aflibercept in treating and computed tomography perfusion imaging in predicting response in patients with pancreatic neuroendocrine tumors that are metastatic or cannot be removed by surgery. Available at <http://www.cancer.gov/about-cancer/treatment/clinical-trials/search/view?cdrid=759683&version=HealthProfessional&protocolsearchid=6072212>. Accessed September 28, 2015
49. Peterson LM, Mankoff DA, Lawton T, et al. Quantitative imaging of estrogen receptor expression in breast cancer with PET and [18F]-fluoroestradiol. *Journal of Nuclear Medicine*. 2008; 49(3): 367–374. [PubMed: 18287268]
50. Dehdashti F, Mortimer JE, Trinkaus K, et al. PET-based estradiol challenge as a predictive biomarker of response to endocrine therapy in women with estrogen-receptor-positive breast cancer. *Breast Cancer Research and Treatment*. 2009; 113(3):509–517. [PubMed: 18327670]
51. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*. 1985; 41(2):361–372. [PubMed: 4027319]
52. Sargent DJ, Conley BA, Allegra C, et al. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*. 2005; 23(9):2020–2027. [PubMed: 15774793]
53. Polley M-YC, Freidlin B, Korn EL, et al. Statistical and practical considerations for clinical evaluation of predictive biomarkers. *Journal of the National Cancer Institute*. 2013; 105(22):1677–1683. [PubMed: 24136891]
54. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *Journal of the National Cancer Institute*. 2009; 101(21):1446–1452. [PubMed: 19815849]
55. Collins LC, Botero ML, Schnitt SJ. Bimodal frequency distribution of estrogen receptor immunohistochemical staining results in breast cancer. *American Journal of Clinical Pathology*. 2005; 123(1):16–20. [PubMed: 15762275]
56. Bhattacharyya S, Kurdziel K, Wei L, et al. Zirconium-89 labeled panitumumab: A potential immuno-PET probe for HER-1 expressing carcinomas. *Nuclear Medicine and Biology*. 2013; 40(4):451–457. [PubMed: 23454247]
57. National Cancer Institute. Linsitinib in treating patients with gastrointestinal stromal tumors. Available at <https://clinicaltrials.gov/ct2/show/study/NCT01560260>. Accessed October 1, 2015
58. Amur S, LaVange L, Zineh I, et al. Biomarker qualification: Toward a multiple stakeholder framework for biomarker development regulatory acceptance, and utilization. *Clinical Pharmacology and Therapeutics*. 2015; 98(1):34–46. [PubMed: 25868461]
59. Kunos CA, Radivoyevitch T, Waggoner S, et al. Radiochemotherapy plus 3 aminopyridine-2-carboxaldehyde thiosemicarbazone (3-AP, NSC #663249) in advanced-stage cervical and vaginal cancers. *Gynecologic Oncology*. 2013; 130(1):75–80.
60. National Cancer Institute. Cisplatin and radiation therapy with or without triapine in treating patients with previously untreated stage IB-IVA cervical cancer or stage II-IVA vaginal cancer. Available at <https://clinicaltrials.gov/ct2/show/NCT01835171>. Accessed October 1, 2015
61. Lesko LJ, Atkinson AJ. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Pharmacology and Toxicology*. 2001; 41:347–366.
62. Verweij J, Therasse P, Eisenhauer E, et al. Cancer clinical trial outcomes: any progress in tumour-size assessment? *European Journal of Cancer*. 2009; 45(2):225–227. [PubMed: 19068275]
63. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*. 2001; 69(3):89–95. [PubMed: 11240971]
64. Ott K, Weber WA, Lordick F, et al. Metabolic imaging predicts response, survival, and recurrence in adenocarcinomas of the esophagogastric junction. *Journal of Clinical Oncology*. 2006; 24(29): 4692–4698. [PubMed: 16966684]

65. Wen PY, Macdonald DR, Reardon DA, et al. Updated response assessment criteria for high-grade gliomas: Response Assessment in Neuro-Oncology Working Group. *Journal of Clinical Oncology*. 2010; 28(11):1963–1972. [PubMed: 20231676]
66. Wahl RL, Jacene H, Kasamon Y, et al. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. *Journal of Nuclear Medicine*. 2009; 50(Supplement 1): 122S–150S. [PubMed: 19403881]
67. MacManus MP, Seymour JF, Hicks RJ. Overview of early response assessment in lymphoma with FDG-PET. *Cancer Imaging*. 2007; 7:10–18. [PubMed: 17766210]
68. Chinot OL, Macdonald DR, Abrey LE, et al. Response assessment criteria for glioblastoma: Practical adaptation and implementation in clinical trials of antiangiogenic therapy. *Current Neurology and Neuroscience Reports*. 2013; 13(5):347. [PubMed: 23529375]
69. Bogaerts J, Ford R, Sargent D, et al. Individual patient data analysis to assess modifications to the RECIST criteria. *European Journal of Cancer*. 2009; 45(2):248–260. [PubMed: 19095437]
70. Ratain MJ, Sargent DJ. Optimising the design of Phase II oncology trials: The importance of randomisation. *European Journal of Cancer*. 2009; 45(2):275–280. [PubMed: 19059773]
71. Korn EL, Sachs MC, McShane LM. Statistical controversies in clinical research: Assessing pathologic complete response as a trial-level surrogate end point in early-stage breast cancer. *Annals of Oncology*. 2015 advance access.
72. Ratain MJ. Phase II oncology trials: Let's be positive. *Clinical Cancer Research*. 2005; 11(16): 5661–5662. [PubMed: 16115898]
73. Goffin J, Baral S, Tu D, et al. Objective responses in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval. *Clinical Cancer Research*. 2005; 11(16):5928–5934. [PubMed: 16115935]
74. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER); Food and Drug Administration. Cancer trial endpoints for the approval of non-small cell lung cancer drugs and biologics: Guidance for industry. Available at <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM259421.pdf>. 2007; accessed February 8, 2016
75. Sargent DJ, Rubinstein L, Schwartz L, et al. Validation of novel imaging methodologies for use as cancer clinical trial end-points. *European Journal of Cancer*. 2009; 45(2):290–299. [PubMed: 19091547]
76. Shi Q, Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *International Journal of Clinical Oncology*. 2009; 14(2):102–111. [PubMed: 19390940]
77. Korn EL, Albert PS, McShane LM. Assessing surrogates as trial endpoints using mixed models. *Statistics in Medicine*. 2005; 24(2):163–182. [PubMed: 15515150]
78. Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. *Statistics in Medicine*. 2012; 31(25):2973–2984. [PubMed: 22711298]
79. Buyse M, Molenberghs G, Burzykowski T, et al. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. 2000; 1(1):49–67. [PubMed: 12933525]
80. Prentice RL. Surrogate endpoints in clinical trials: Definitions and operational criteria. *Statistics in Medicine*. 1989; 8(4):431–440. [PubMed: 2727467]
81. Freedman LS, Graubard BI, Schatzkin S. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*. 1992; 11(2):167–178. [PubMed: 1579756]

Figure 1.

The Paired Design to compare clinical outcomes resulting from basing treatment decisions on a new imaging-based staging procedure to those resulting from basing treatment decisions on a standard procedure. Paired Designs can also be used for comparing clinical outcomes resulting from using a new imaging-based prognostic marker to guide treatment to those resulting from using standard prognostic variables.

Figure 2. The Modified Marker Strategy Design to assess the improvement in clinical outcome resulting from using an imaging-based prognostic marker to guide treatment decision.

Figure 3.

Illustration of a predictive marker. Biomarker-positive patients experience a noticeable improvement in survival on the investigational treatment, whereas no such improvement is observed among biomarker-negative patients.

Figure 4. Marker by Treatment Interaction Design without stratification to test the value of an imaging-based predictive marker.

Figure 5. Marker by Treatment Interaction Design with stratification to test the value of an imaging-based predictive marker.

Figure 6.

Illustration of a case where a predictive biomarker is not useful for guiding treatment selection. Although the improvement in survival on the investigational treatment is more pronounced in biomarker-positive patients, biomarker-negative patients also experience improved outcome on the investigational treatment as well. Thus, the investigational treatment is superior regardless of biomarker positivity.

Table 1

The roles of advanced imaging discussed in this manuscript, with definitions, examples, and the evidence needed to demonstrate adequate performance of the imaging in each context.

Role	Purpose	Evidence of adequate performance	Example
Diagnosis and Staging	Indicates the location of the primary tumor and the extent of the disease in terms of size of the primary lesion, nodal involvement, and metastases.	Superior sensitivity and specificity (or positive and negative predictive values) in detecting nodal involvement and metastases. A significant number of patients for whom the stage according to the imaging differs from that according to standard procedures.	Anatomic imaging through CT
Prognostic marker	A baseline measurement that forecasts the expected outcome of a patient undergoing standard therapy.	High rates of lengthy progression-free, recurrence-free survival, or overall survival in patients identified as low risk. Superior outcomes among patients whose treatment decisions were dictated by the imaging.	Size of residual tumor according to MRI in medulloblastoma
Predictive biomarker assay	A measurement of a baseline characteristic that can differentiate between patients whose outcomes on a particular class of therapies is expected to be superior to outcomes on standard therapy versus patients not expected to experience such a benefit.	Superior outcome on investigational treatment relative to comparator treatment in biomarker-positive patients (those expected to benefit) and similar or worse outcome on investigational treatment in biomarker-negative patients.	FES SUV as a measurement of ER expression in breast cancer
Pharmacokinetic marker	Changes from baseline indicating the trajectory and distribution of a drug throughout the body.	Adequate metrological performance.	[89Zr]-panitumumab SUV
Pharmacodynamic marker	Changes from baseline indicating the downstream biological effects of the drug.	Adequate metrological performance.	FDG uptake following administration of therapies targeting insulin-like growth factor 1
Interim response assessment	A change from baseline observed after the initial cycles of treatment that forecasts eventual outcome.	Superior outcomes resulting from switching treatment among those not showing interim response. Similar or inferior outcomes resulting from switching treatment among those showing interim response.	Decrease in FDG SUV of 35% or more in gastric cancer patients undergoing chemotherapy after the first cycle
Basis of a Phase II trial endpoint	A pre- to post-treatment change that indicates a drug's anti-tumor activity used to determine whether to proceed to the subsequent Phase III investigation.	A high proportion of positive Phase III studies associated with Phase II studies in which trial positivity was declared using an endpoint based on the imaging.	Complete metabolic response according to FDG-PET in cervical cancer
Basis of a Phase III trial endpoint	A pre- to post-treatment change that serves as a surrogate for a definitive clinical endpoint.	A high concordance of between-arm differences in the imaging-based endpoint with between-arm differences in the definitive endpoint.	PFS based on anatomic imaging

Table 2

Statistical terminology found in the manuscript, their purposes in the context of clinical trials, and examples.

Term	Purpose	Examples
Log-rank test	To test whether the distributions of a time-to-event outcome between two groups are equal.	Comparison of the OS of early non-responders who switch treatments to that of early non-responders who do not.
Logistic regression	To model the probability of an event (e.g. pathologic complete response) occurring as a function of one or more explanatory variables.	Assessment of the association between FDG SUV and response (with or without adjustment for standard prognostic variables such as age and performance status).
Cox regression	To model the rate at which an event (e.g. progression or death) occurs as a function of one or more explanatory variables	Assessment of the association between FDG SUV and OS (with or without adjustment for standard prognostic variables such as age and performance status).
Fisher's exact test	To test the association between two categorical variables.	Test of differences in response rates in breast cancer patients to endocrine therapy among biomarker-positive (high FES SUV patients) versus biomarker-negative patients.
Mann-Whitney U test	To test whether the distributions of a quantitative variable between two groups are equal.	Comparison of FES SUV in breast cancer patients who ultimately respond to endocrine therapy to that in patients who do not.
Test for qualitative interaction	To test whether one treatment is more efficacious than another in one subset of patients but not in another subset.	Test of whether an investigational treatment is more efficacious than standard therapy among biomarker-positive patients but in biomarker-negative patients, the investigational treatment is either as efficacious or less efficacious.
Kendall tau rank correlation coefficient	A measure of the association between two quantitative or ordered categorical variables.	Assessment of the association between FES SUV and ER expression according to immunohistochemistry.
Wilcoxon signed-rank test	To test whether repeat measurements on a particular patient differ.	Assessment of changes in FDG SUV from baseline to post-treatment.