EUROPEAN JOURNAL OF CANCER 45 (2009) 290–299

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Validation of novel imaging methodologies for use as cancer clinical trial end-points

D.J. Sargent[a,*], *L. Rubinstein*[b], *L. Schwartz*[c], *J.E. Dancey*[d], *C. Gatsonis*[e], *L.E. Dodd*[b], *L.K. Shankar*[b]

[a]*Mayo Clinic, 200 First Street, SW, Rochester, MN 55905, United States*
[b]*National Cancer Institute, Bethesda, MD, United States*
[c]*Memorial Sloan Kettering, New York, NY, United States*
[d]*National Cancer Institute of Canada Clinical Trials Group, Kingston, Ontario, United States*
[e]*Brown University, Providence, RI, United States*

ARTICLE INFO

Article history:
Received 17 October 2008
Accepted 29 October 2008
Available online 16 December 2008

Keywords:
Surrogate endpoint
Phase II clinical trial
Meta-analysis
Imaging endpoint
RECIST

ABSTRACT

The success or failure of a clinical trial, of any phase, depends critically on the choice of an appropriate primary end-point. In the setting of phases II and III cancer clinical trials, imaging end-points have historically, and continue presently to play a major role in determining therapeutic efficacy. The primary goal of this paper is to discuss the validation of imaging-based markers as end-points for phase II clinical trials of cancer therapy. Specifically, we outline the issues that must be considered, and the criteria that would need to be satisfied, for an imaging end-point to supplement or potentially replace RECIST- defined tumour status as a phase II clinical trial end-point. The key criteria proposed to judge the utility of a new end-point primarily relate to its ability to accurately and reproducibly predict the eventual phase III end-point for treatment effect, which is usually assessed by a difference between two arms on progression free or overall survival, both at the patient and more importantly at the trial level. As will be demonstrated, the level of evidence required to formally and fully validate a new imaging marker as an appropriate end-point for phase II trials is substantial. In many cases, this level of evidence will only become available by conducting a series of coordinated prospectively designed multicentre clinical trials culminating in a formal meta-analysis. We also include a discussion of situations where flexibility may be required, relative to the ideal rigorous evaluation, to accommodate inevitable real-world feasibility constraints.

1. Introduction

The success or failure of a clinical trial, of any phase, depends critically on the choice of an appropriate primary end-point. The end-point must be sensitive to the effect of the treatment under study, be able to be unambiguously and reliably measured, and optimally be highly clinical relevant. In the setting of phases II and III cancer clinical trials, imaging end-points have historically, and continue to presently play a major role in determining therapeutic efficacy. The utility of imaging-based end-points in the context of cancer is based on several factors. The non-invasive or minimally invasive nature of imaging allows the integration of tumour biology information at each site of disease, and for most imaging methodologies,

* *Corresponding author:* Tel.: +1 507 284 5380; fax: +1 507 266 2477.
E-mail address: Sargent.daniel@mayo.edu (D.J. Sargent).
0959-8049/$ - see front matter
doi:10.1016/j.ejca.2008.10.030

over multiple sites of metastases throughout the body. Non-invasive imaging assays also lend themselves to serial evaluation during surveillance and evaluation for recurrent disease. Increasingly, with advances in molecular and functional imaging, we will gain the ability to assess not only the morphology of the primary tumour and its metastases, but also the metabolic, hypoxic, proliferative and receptor status of the lesions, suggesting that the role of imaging may further increase.[1–6]

The primary goal of this paper is to discuss the validation of imaging-based markers as end-points for phase II clinical trials of cancer therapy. We will touch upon end-points for phase III studies but only insofar as this is necessary for us to pursue our primary goal. Phase II clinical trials in cancer are designed to provide evidence of biological drug activity. Phase II trials have traditionally used imaging-related end-points, such as tumour shrinkage or delayed tumour growth, as anti-tumour activity signals. The utility of tumour response as a phase II end-point is supported by biology (tumours rarely shrink by themselves) and history; drugs that induce tumour responses in early clinical trials are more likely to subsequently lead to positive phase III trials and drug registration.[7,8] However, tumour response by itself does not constitute a necessary or sufficient demonstration of clinically meaningful drug efficacy, as (1) tumour response may not result in an improvement in survival or quality of life and (2) patients may benefit from therapy without obtaining a tumour response.[9,10] Therefore, increasingly tumour growth (progression) despite drug administration is viewed as evidence of drug inactivity, and progression-free survival (PFS), either overall or at a fixed time point, is increasingly being used as a phase II clinical trial end-point.

Since 2000, a standard for evaluating imaging-related end-points in solid tumour cancer clinical trials has been that defined by the RECIST project.[11] These criteria specify the manner by which data from standardised imaging modalities, such as CT and MRI, are used to define clinical trial end-points. In this volume, the RECIST criteria are updated to address multiple issues that have arisen since the initial RECIST publication in 2000.[12] Given the importance of imaging-related end-points in cancer clinical trials, and the rapid pace at which new imaging modalities are becoming available, in this paper we focus on methodological issues that must be considered for a new imaging end-point to be appropriately validated as a primary end-point for phase II clinical trials. Specifically, we outline the issues that must be considered, and the criteria that would need to be satisfied, for an imaging end-point to supplement or potentially replace RECIST- defined tumour status as a phase II clinical trial end-point. For example, a SUV decrease from a FDG-PET scan, if appropriately validated, might be accepted as an alternative to a RECIST-based partial response as assessed by a CT scan, or provide an additional mechanism to upgrade a patient from a partial responder to a complete responder.[13,14] Alternatively, volumetric imaging, if validated as a more accurate predictor of subsequent therapeutic benefit (demonstrated in phase III trials), could potentially replace uni-dimensional imaging as currently specified by RECIST.

To evaluate the quality of a study and to compare results across studies, established standards for reporting relevant elements of study design and analysis are vital. Guidelines for evaluating and reporting results from the studies of tissue-based biomarkers have recently been developed. The Reporting Tumour Marker Prognostic Study (REMARK) guidelines, for example, describe a list of basic elements that should be documented in any report of a tissue-based tumour marker study.[15] These guidelines include the reporting of study design, pre-specified hypotheses, patient characteristics and the statistical analysis plans. Similarly, the Tumour Marker Utility Grading System (TMUGS) established a standardised technique to allow evaluating the utility of a known marker based on the existing evidence.[16] The principles outlined in these tissue-based biomarker guidelines in general are equally appropriate in the context of imaging-based biomarkers. Here, we focus on study design issues to validate an imaging biomarker; adherence to these standards, and reporting the studies per REMARK guidelines, will allow the generation of TMUGS level '+++' or '++' evidence for imaging modalities.

Several authors have proposed that the alternative uses of criteria based on the measurements obtained via morphologic imaging may be preferred to tumour response as a predictor of improvement in the clinically relevant end-points in phase III clinical trials. For example, it has been proposed that progression-free survival, assessed by the current RECIST status, provides greater predictive accuracy than tumour response for phase III end-points,[17,18] or that a continuous measure of tumour size change may be preferred to the categorical definitions of RECIST.[19] In this paper we focus on the technological advances designed to supplement or potentially replace tumour assessments based on RECIST, as opposed to an end-point that uses RECIST-based anatomical imaging data in an alternative manner. In addition, we acknowledge that anatomically based tumour assessments in phase II clinical trials, using either response rate or PFS based on the current RECIST criteria, as well as for previous criteria such as WHO, are well documented to provide imperfect prediction of subsequent therapeutic benefit in phase III clinical trials.[20–22] However, at the present time, the RECIST criteria remain a clearly defined and recognised standard, and for a new approach to be advocated, it must provide clear advantage to the recognised standard. We therefore will consider RECIST as the default competitor for the new imaging approaches.

2. Background and current state assessment

Tumour response as an end-point in therapeutic trials was first codified by the World Health Organization (WHO) based upon the initial publications that focus on the *reproducibility* of the metric for assessing tumour response and progression. The specific response categorisations, as well as the cutpoint values for response categorisation (50% for the World Health Organization's bi-dimensional metric, corresponding to a partial response of 30% by uni-dimensional measurements), have remained basically unchanged in the evolution of response assessment. The modalities used to assess tumour size, however, have evolved substantially. In addition, novel non-response inducing agents (cytostatic) are being developed, in addition to new cytotoxic agents.[4,23–25] Taken together, these factors have led to a recognition that

end-points based on RECIST have limitations in certain primary tumour types, and with certain therapeutic agents. Indeed, there are clear limitations to the universal use of RECIST in all tumour types for all agents. However, many of the shortcomings that have been noted in the literature represent either lack of proper clinical interpretation of radiologic images, or intrinsic limitations of any scoring system where categorical response criteria are binned into categories whilst the data actually represent continuous change. These limitations will be relevant to any imaging modality or other biomarker technique. Here, we present selected examples to illustrate these concepts.

In gastrointestinal stromal tumours, divergent strategies to RECIST have been proposed which, in single institutional trials, improve correlation with survival outcomes.[23] These criteria have been based on modifying the RECIST cut point for progression, as well as utilising change in tumour density to assess disease status. The criteria involving changes in tumour density on post-contrast CT scans are particularly appealing as they introduce a functional element into anatomic criteria. When properly applied, these approaches may be invaluable, principally by providing additional clinically relevant data from the existing scanning technology. However, at this time the reproducibility of such criteria amongst multiple institutions requires further validation in independent data sets. Considering the variability in image acquisition techniques amongst centres, and the resultant wide variability in density or perfusion measurements post-contrast, proposed criteria such as these must be carefully vetted not only for correlation with outcome but for reproducibility of the measurement metric. Further, correlation of within-patient changes in a biomarker and patient outcome is inadequate to conclude that a therapy that alters the biomarker will also alter the ultimate patient outcome – a correlate does not a surrogate make.[26]

Other criteria, such as those proposed in mesothelioma, have carefully been created in an attempt to address the reproducibility issue. For example, there is considerable heterogeneity of response seen within tumours such as mesothelioma, evident on multiple CT slices, with innumerable potential linear diameters. Approaches to provide reproducible measures are clearly a necessary part of the strategy to documenting response in this patient population[27] (for specific details see Appendix I). However, these responses must not only be reproducible, but correlate with true clinical outcome through validation in large multi-institutional data sets.

3. Novel imaging modalities of key interest

Three imaging modalities or metrics currently posed to play a role in disease assessment are PET (including but not limited to the FDG tracer), dynamic contrast-enhanced MRI (DCE-MRI) and three-dimensional tumour measurement. In this section, we discuss key performance characteristics as they relate to the potential widespread use and acceptance of these three imaging modalities.

When an imaging assay is used serially to assess changes in tumour characteristics, a change analysis is being performed. In the setting of developing a technique for wide usage in clinical trials, as with any other assay, the performance characteristics of the imaging assays in the multicentre setting must be established. At this time, the methods used to obtain FDG-PET scans and to assess FDG metabolism and uptake are clearly varied[28,29]; this is also true of studies evaluating DCE-MRI.[30,31] The accuracy, variance and reproducibility of the imaging technology must be determined to assure a quantitative or semi-quantitative index, which is biologically meaningful.

To provide guidance, and standardise the acquisition analysis and interpretation of FDG-PET in clinical trials, the Cancer Imaging Program of the NCI convened a workshop in 2005, resulting in consensus guidelines that are currently being used in NCI trials as well as several studies developed and performed by the pharmaceutical industry.[29] The guidelines include recommendations on patient preparation, image acquisition, image reconstruction, quantitative and semi-quantitative analysis of FDG-PET images, quality assurance issues, reproducibility and other parameters of importance to be used in FDG-PET studies before and after a therapeutic intervention. The NCI Cancer Imaging Program has also engaged the MRI community in a similar process to develop consensus guidelines for the performance of dynamic contrast-enhanced MRI (DCE-MRI) as well as magnetic resonance spectroscopy (MRS) (http://imaging.cancer.gov/).

In the specific setting of FDG-PET, multiple studies have evaluated the role of FDG-PET in assessing response to treatment in non-small cell lung cancer (NSCLC), oesophageal cancer, head and neck cancer, breast cancer and many other tumours.[1,6,14,32,33] To date, these studies have been primarily performed in single institutions with small numbers of patients. Similarly, although promising, the data for 3D measures of treatment response are scant, involve relatively small numbers of patients and are not readily comparable. Therefore, at the present time we feel there are inadequate data to support the inclusion of functional imaging (PET and DCE-MRI) or expanded morphologic imaging (3D measurement) RECIST version 1.1 criteria as presented in this volume.[12] In a later section, we provide specific details of an ongoing trial seeking to provide components of the necessary information to determine whether FDG-PET and/or 3D tumour measurements can ultimately serve as valid trial end-points.

4. Specific criteria prior to the launch of validation studies for imaging-based end-points

Prior to initiating definitive studies to validate imaging-based end-points, several criteria must be met (Table 1). The technology must be at a relatively stable stage, and have the potential for broad availability across centres, which will perform the therapeutic intervention in the clinical trial. All aspects of image acquisition including the frequency of scanning, modality, timing of image acquisition relative to injection of contrast agents or radiolabelled tracers, and pulse sequence parameters or other imaging parameters, must be specified. A standardised protocol for interpreting images, qualitative or quantitative, must be established, taking into

Table 1 – Criteria necessary prior to definitive evaluation studies.

Technology stable
Broad availability
Image acquisition parameters specified – scanning frequency, timing relative to contrast, pulse sequence parameters, etc.
Standardised interpretation protocol
Documented reproducibility
Normal ranges defined

account the specific modality parameters and reproducibility of the measurement metric.

Standardisation of technique will also help to limit variability across readers, although such variability is unlikely to disappear even in modalities that produce quantitative test results. For example, SUV measurements in PET studies are subject to variability related to the determination of a Region of Interest (ROI) by the test interpreter. The assessment of variability across readers is particular to imaging and remains an important consideration in imaging marker evaluation studies. Accordingly, studies to evaluate imaging reproducibility and to document a normal range of values for replicate acquisitions and interpretations should be conducted. Such studies should include an evaluation of the rating system to establish the categories of response or progression, optimally based on biologically relevant cut values. Further, the appropriate patient population should be well defined, along with an understanding of any limitations of the technique in certain diseases or disease sub-types.

In addition to these technical issues, in most if not all cases, it is assumed that a sound biological rationale exists for the use of an imaging technique as an end-point. For example, the use of radiographic tumour response in the phase II studies of cytotoxic agents assumes that tumour shrinkage is an outcome reflecting drug activity. Biologic confounders must also be accounted for, for example, if assessing tumour metabolism via FDG-PET, one must consider treatment specific issues such as non-specific uptake in inflammation post radiotherapy, which can impact the optimal time to assess post-treatment response. This is usually minimised by waiting several weeks post radiation therapy to obtain the post-treatment FDG-PET scan, allowing the inflammation time to subside. Whilst critical, for the validation of future phase II end-points, we stress that biological plausibility alone is inadequate to allow any end-point to be validated as without a demonstration of correlation with a true patient benefit (phase III) outcome. We return to this point in the discussion.

5. Criteria to validate new end-point

We will outline, in general, critical issues, constraints and goals associated with the validation of a new imaging end-point, to provide guidance and a conceptual framework for the validation of individual imaging end-points. It is not our purpose to precisely prescribe how to validate any specific new phase II trial imaging end-point as this will depend on the specific characteristics and purpose of the particular end-point, whether its use is restricted to a certain patient subgroup, the current state of development of the end-point and the technology to measure it.

The primary purpose of an imaging end-point in the phase II setting is to serve as an early but accurate indicator of a promising treatment effect. As such, the key criteria for judging the utility of a new end-point will be its ability to predict accurately the phase III end-point for treatment effect, which is usually assessed by a difference between two arms on PFS or overall survival (OS). More precisely, the measure of treatment effect on the phase II end-point must correlate sufficiently well with the measure of treatment effect on the phase III primary end-point that the former can be considered reasonably predictive of the latter.

An initial question to be addressed is whether the new end-point is destined to be '+++' or only '++', according to the TMUGS criteria[16] – in other words, will the end-point be useable, by itself, as the primary criterion for moving to a phase III study, or will it be useable as one of several such criteria. In this paper, we focus on validating early end-points that are anticipated to be '+++'. The second question relates to the current utility of RECIST in the disease setting under exploration. In a disease setting where RECIST (or existing alternatives) predicts phase III outcomes poorly, improved prediction of outcome over the current standard would be of clear utility, even if the imaging modality does not meet criteria for full end-point validation.

It is not sufficient that the end-point being considered for a phase II trial be a prognostic indicator of clinical outcome, although it will usually be the case that early end-points are prognostic of clinical outcome even in the absence of a treatment effect. Within the context of a clinical trial, the early end-point must capture at least a component of treatment benefit, a concept that specifies that a change due to treatment in the early end-point predicts a change in the ultimate clinical end-point. Theoretical principles to define treatment benefit were outlined by Prentice,[34] although capturing the full treatment benefit (as measured by the phase III end-point) has been recognised as too strict to be useful in practice.[35,36] A more practical and demonstrable criterion requires that the early end-point captures a substantial proportion of the treatment benefit, for example, more than 50%.[20,35,36] This approach has been used to establish the utility of end-points such as tumour response and progression-free survival (PFS) by demonstrating that they are sufficiently predictive of OS, even if they do not satisfy the Prentice criterion.[18,20,21,37–42]

Establishing the utility of the end-point can be separated into an early development and a later validation stage (Table 2). Even in the early development stage, optimally work should be performed in the context of randomised studies, which most reliably allow the measurement of treatment benefit.[35] Practically, much early development work will by necessity occur in the context of prospective cohort studies, which should at minimum have patients with uniform treatment. In the early development stage of a new imaging end-point, utility determination will likely be restricted to demonstrating that in single studies the end-point captures much of the treatment benefit *at the individual patient level*. Such a demonstration suggests, but does not prove, that the end-point may also capture much of the treatment benefit at the trial level. Freedman et al.[35] describe one approach to estimating the

Table 2 – Early and late phases of end-point validation.

Attribute	Early phase validation	Late phase validation
Goal	Individual patient level outcome prediction	Trial level outcome prediction
Setting	Single randomised trials or uniformly treated patients from non-randomised trials	Meta-analysis of randomised clinical trials
Methods	Correlation analyses between end-points within patients	Correlation analyses between trial level effects on both end-points

proportion of treatment effect explored by modelling the treatment effect on the ultimate end-point (Appendix II).

Success at this early validation phase, by demonstrating a high correlation at the patient level between the early end-point and the ultimate clinical end-point within a trial, randomised or not, is not sufficient to validate an end-point. Such a correlation may be a result of prognostic factors that influence both end-points, rather than a result of similar treatment effect on the two end-points. Despite this caveat, a reasonably high patient level correlation (for example, >50%) would suggest the possible utility of the early end-point and the value of subsequently assessing, by means of a larger analysis, the predictive ability of the early end-point for the ultimate phase III end-point for treatment effect at the trial level.

In the later stages of validation, as argued by Korn et al.,[36] the true test of the validity of an end-point is whether it captures treatment benefit *at the trial level*. In other words, there must be a strong association between the measure of treatment effect as assessed by the early end-point and the measure of treatment effect as assessed by the end-point to be used in a phase III trial, which is most likely the estimated treatment hazard ratio associated with PFS or OS. In virtually all cases, such an assessment must be performed in the context of a meta-analysis of phase III trials, where both end-points are measured. Such a meta-analysis may be performed using trials already conducted, if imaging data are available. However, the methodologic aspects of meta-analysis itself must be defined prospectively in order to be statistically convincing. Such analyses have been performed for the relationship between tumour response and OS in advanced colorectal cancer[18,20] and in metastatic breast cancer.[21] In each case, the proportion of variation in the treatment effect on OS explained by the log OR of tumour response is less than 50%. In metastatic breast cancer, tumour response was seen to capture a much greater proportion of the treatment benefit reflected by PFS (92%). Such meta-analyses are substantial undertakings; the breast study included 11 trials, whilst the colorectal studies included 18–28 trials.

Even with a substantial number of trials included in a planned meta-analysis, obtaining adequate power to demonstrate that a substantial proportion of the treatment benefit, at the trial level, is captured by the early imaging end-point is challenging (see Appendix III). In the end, it will be necessary to compromise and accept that one cannot always prospectively assure the desired power to achieve the desired lower confidence bound. We stress that whatever form the meta-analysis is to take, it must be pre-specified formally in a protocol. An ad hoc approach will increase the probability for bias in the estimation of correlation between the two measures of treatment benefit (that associated with the early end-point versus that associated with the primary phase III end-point).

The recommendations above are based, in large part, on guidelines to validate a phase III surrogate end-point. Although the basic principles behind validation of a phase II end-point remain similar, in specific contexts the standards may appropriately be adapted for a phase II end-point. For example, a meta-analysis of fewer trials may be all that is possible, and/or an imaging end-point may be considered acceptable for use in the phase II trials with a lower correlation between the treatment effect of interest and that estimated by the imaging end-point (for example, capture of 50% of the treatment effect may be adequate). We further note that there may be scenarios to allow refinements to RECIST based on the technical or other advances, in which the above standards of validation are not required. For example, an existing concern regarding RECIST is the reproducibility of tumour measurements across readers. If a more reproducible anatomic method were available (e.g. a computer-assisted diagnostic (CAD) algorithm) that consistently provided the same result as an expert reader across sites, this would be an improvement upon standard RECIST and would likely be acceptable without a meta-analytic validation.

6. Specific example of an ongoing imaging validation trial

There are currently ongoing national trials within the United States designed to provide data at the early validation phase (Table 2) for FDG-PET as a biomarker for response in lymphoma and non-small cell lung cancer. These multicentre trials seek to validate the results of single and other multicentre trials that provided promising evidence of the utility of these biomarkers to make patient-level biomarker outcome prediction, thus reflecting the early phase of biomarker validation. These trials have been designed for the purpose of biomarker validation by optimising the image acquisition parameters within the real world limitations of a multicentre trial, and by providing both local and expert assessments of the imaging results. Validating imaging methods as potential biomarkers for tumour response to treatment requires the demonstration of a high degree of test–retest reproducibility for the imaging method. Therefore, test–retest reproducibility will also be an important element of these trials.

As a specific example, ACRIN protocol 6678 (FDG-PET/CT as a Predictive Marker of Tumour Response and Patient Outcome: Prospective Validation in Non-small Cell Lung Cancer) will explore three types of evaluation of imaging biomarkers that relate to their potential role as clinical trial end-points. Specifically, the study includes (a) comparison of time-to-event distributions for biomarker ‘responders’ and ‘non-responders’, (b) assessment of the predictive accuracy of the biomarkers and (c) assessment of the test–retest reliability

of the imaging measurement. The primary aim of the study is to assess whether a metabolic response, defined as a $\geqslant$25% decrease in peak tumour SUV post-cycle 1 of chemotherapy, provides early prediction of treatment outcome as determined by 1-year patient survival. A secondary aim of the study is to compare the predictive value of FDG-PET/CT for 1-year overall survival after one and two cycles of chemotherapy. Further secondary end-points assess the test–retest reproducibility of standardised uptake values (SUVs) measured by PET/CT systems.

In addition to the evaluation of FDG-PET-based markers, ACRIN protocol 6678 also includes an exploration of tumour volumetry. This study will permit an early assessment of whether volumetric analysis is feasible and reproducible in the multicentre trial setting, and whether volumetric change analysis early in the course of therapy has the potential to predict a phase III end-point (long-term survival), as an independent or complementary variable to FDG-PET. Additional detail on ACRIN 6678 is available in Appendix IV, and the full protocol is available at http://www.acrin.org/TabID/162/Default.aspx.

7. Discussion

It is clear from the preceding discussion that the level of evidence required to formally and fully validate a new imaging marker as an appropriate end-point for the phase II trials is substantial. In many cases, this level of evidence will only become available by conducting a series of coordinated prospectively designed clinical trials, such as the example described above. As the financial and time burdens involved in the prospective clinical trials are considerable, it is necessary to consider whether selected elements of the validation of a new technology may be performed retrospectively, that is, on data from patients who have already been enrolled, treated and assessed on a previous clinical trial (or even who were not on a clinical trial). Clearly, each component of a validation analysis must prospectively specify the hypothesis, the analytic techniques, the patient population and the precise imaging algorithms to be used. If these elements are clearly specified prospectively in a protocol, it is possible to derive evidence from a situation in which patients may have already been enrolled in a randomised clinical trial, as long as the imaging results are available from the vast majority of patients without selection bias. Such a retrospective evaluation may be most appropriate at the early validation phase of an end-point's development, where the focus is on the individual patient-level treatment benefit prediction. As standardisation is attained in measurements based on the new imaging modalities, if the data are stored in a queriable database, such analyses may become possible. An implication of this recommendation is that ongoing and future phase III trials should incorporate the appropriate collection of imaging end-points whenever feasible.

The considerable enthusiasm surrounding the use of new imaging modalities must be tempered by a number of examples that suggest that end-points reflecting a biological effect of an agent may not result in improvements in a clinically meaningful end-point in a phase III trial. For example, a clear and measurable change in vascular permeability and blood flow as assessed by dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) failed to predict for improved survival, when the VEGFR targeted agent vatalanib (PTK787/ZK22584) was tested in the phase III trials in colorectal carcinoma patients.[43,44] Another example may be the use of FDG-PET to determine response in patients with GIST. FDG-PET response has been shown to be an early and sensitive evaluation of the effectiveness of imatinib in this disease,[23] and as such FDG-PET is useful to evaluate imatinib's activity in individual patients, and to screen for activity in the phase II trials for drugs with a similar mechanism of action to imatinib. However, it is not clear whether similar FDG-PET effects would occur with other agents that differ in mechanism of action from imatinib. Furthermore, occurrence of FDG-PET changes similar to those seen with imatinib in GIST patients may not reflect changes seen in other settings that will correlate with clinical benefit.

This raises the critical and difficult issue in the validation of new imaging techniques of the degree to which the validation of an early imaging end-point may be universal versus being disease site and agent specific. Clearly generalisability is never guaranteed; however, independent validation for each imaging modality for each disease site/agent class is clearly intractable. As a general guideline, if one performs a rigorous evaluation, according to the principles outlined here, in one particular setting, then the level of evidence required for that imaging modality for disease sites that have historically performed consistently, and agents with similar mechanisms of action, may be lessened. Consistency of the novel imaging results with results obtained using other biomarkers (RECIST, PFS, etc.) which may be observed later in the trial strengthens the evidence for a new biomarker. In the end, the study design, study end-points and the level of validation required must be informed by the careful examination of both the biology of the imaging marker and the mechanism of action of the therapeutic intervention. A critical difference exists between upstream markers that may be pathway or target specific versus downstream markers (metabolism, apoptosis, proliferation) that are intended to measure biologic activity in the tumour; these guidelines focus on the downstream markers. Ultimately, researchers must balance cost and time efficiencies against potential bias. To achieve this, an iterative strategy may be adopted through which until substantial evidence of lack of prediction exists, researchers may proceed as though previous results of predictive ability established in similar settings continue to apply in a new setting. However, as new knowledge becomes available, studies may need to become more clearly disease or agent-class focused. This strategy clearly only applies in the phase II setting; phase III end-points must have been appropriately validated to allow practice changing decisions.

Ultimately, whether the biological measurements allowed by advanced imaging are meaningful predictors of drug efficacy and patient benefit depends on multiple factors, including the importance of the biologic effect being assessed on tumour growth and survival, and the magnitude, duration and frequency of occurrence of the biologic effect in a given patient population. The relevance of these factors must be understood for each modality and in each clinical situation. We remain optimistic that through the careful design of

prospective trials, coupled with protocol-specified analyses of existing, standards-based datasets, promising imaging modalities may be properly validated for inclusion into future RECIST versions.

Appendix I

The modified RECIST criteria employed in mesothelioma response assessment[27] consist of measuring tumour perpendicular to the chest wall or mediastinum in two positions at three separate levels on transverse cuts of CT scan. The sum of these six measurements defines a pleural uni-dimensional measurement. The transverse cuts are recommended to be at least 1 cm apart and related to anatomical landmarks in the thorax to enhance reproducibility on follow-up scans. At the follow-up scans, the pleural thickness is measured at the same position and at the same level.

Appendix II

The Freedman approach[35] involves estimating the treatment effect on the true end-point, defined as τ, and then assessing the proportion of treatment effect explained by the early end-point by $1-\hat{\tau}_a/\hat{\tau}$, where the ratio is that of estimated treatment effect, adjusted for the early end-point, divided by the unadjusted estimated treatment effect. Thus, for an early end-point that captures no treatment benefit $\hat{\tau}_a = \hat{\tau}$, the proportion of treatment effect explained is 0%. At the opposite extreme, for an early end-point that captures all the treatment benefit $\hat{\tau}_a = 0$, satisfying the Prentice criterion, the proportion of treatment effect explained is 100%. However, as noted by Freedman, this approach has statistical power limitations that will generally preclude conclusively demonstrating that a substantial proportion of the treatment benefit at the individual patient level is explained by the early end-point. In addition, it has been recognised that the proportion explained is not indeed a true proportion, as it may exceed 100%, and that whilst it may be estimated within a single trial, that data from multiple trials are required to provide a robust estimate of the predictive end-point.[37]

Appendix III

In the setting of conducting a meta-analysis of randomised clinical trials, even if 81% of the variation in the primary phase III end-point is explained by the early end-point, it will require 28 trials to achieve 90% power to demonstrate that the proportion of variation is at least greater than 50%, with 95% confidence. (Relaxing the requirement to 90% confidence does little – the requirement is reduced to 23 trials.) If only 64% of the variation in the primary phase III end-point is explained by the early end-point ($r = .8$), a more realistic assumption, 28 trials yield 90% power to achieve a lower 95% confidence limit of at least 23% on the proportion of variation explained – which is not very satisfactory. One possible approach to this dilemma is to separate individual trials into homogeneous strata defined by appropriate prognostic variables, and correlate the two measures of treatment effect over the much greater number of separate strata. This may be particularly useful if the treatment effects vary over the strata within trials. As long as care is taken that the strata are not so sparse that the estimates of treatment effect become statistically unstable, the increase in precision of the correlation estimate should overcome the decrease in precision of the two sets of treatment benefit estimates.

Appendix IV. Summary of ACRIN 6678: FDG-PET/CT as a predictive marker of tumour response and patient outcome: prospective validation in non-small cell lung cancer

Objectives

This study has four objectives:

1. To test whether a metabolic response, defined as a $\geqslant$25% decrease in peak tumour SUV post-cycle 1 of chemotherapy, provides early prediction of treatment outcome (tumour response and patient survival).
2. To determine the test–retest reproducibility of quantitative assessment of tumour FDG uptake by SUVs.
3. To study the time course of treatment-induced changes in tumour FDG uptake.
4. To evaluate in an exploratory analysis, changes in tumour volume during chemotherapy by multislice CT.

The two specific hypotheses underlying this trial are (i) a metabolic response, defined as a $\geqslant$25% decrease in peak tumour SUV post-cycle 1 of chemotherapy, provides early prediction of treatment outcome (tumour response and patient survival) and (ii) tumour glucose utilisation can be measured by FDG-PET/CT with high reproducibility.

Primary end-point

The primary end-point of this study is the prediction of 1-year overall survival by monitoring changes in tumour metabolic activity during the first chemotherapy cycle, where metabolic response is classified as $\geqslant$25% decrease in SUV of the primary tumour relative to baseline (pre-chemotherapy).

Secondary end-points

1. Assessment of the association between a metabolic response after one cycle of chemotherapy and subsequent best tumour response according to standard anatomic response evaluation criteria (RECIST).
2. Assessment of the association between a metabolic response after the first chemotherapy cycle and progression free survival.
3. Comparison of the predictive value of FDG-PET/CT for 1-year overall survival after one and two cycles of chemotherapy.
4. Assessment of the test–retest reproducibility of standardised uptake values (SUVs) measured by PET/CT systems.

Exploratory analyses

In addition to the specific end-points described above, the trial provides data for hypothesis-forming analyses. Specifically, the following questions will be addressed:

1. Could ROC analysis be used to estimate an optimal threshold for the SUV differences in defining a metabolic response?
2. Can changes in tumour volume be assessed by multi-detector CT early during the course of chemotherapy?
3. Are tumour volumetric changes correlated with patient outcomes?
4. Can one develop parameters that combine metabolic and volumetric data, and do these parameters allow a better prediction of patient outcome than metabolic changes alone?
5. How does the prognostic value of a metabolic response in PET compare with the prognostic value of tumour response according to standard tumour response assessment according to RECIST?
6. What is the correlation between metabolic changes in the primary tumour and in the metastatic lesions?
7. How should changes in FDG uptake of multiple metastatic lesions be quantified?

Participant population

Inclusion criteria

1. Histologically or cytologically proven NSCLC.
2. Tumour stage IIIB (with malignant pleural effusion) or stage IV.
3. Tumour staging minimum requirements:
   - CT scan of the chest and upper abdomen (to include liver and adrenal glands) within 4 weeks prior to registration;
   - History/physical examination within 6 weeks prior to registration;
   - CT scan of the brain if there is headache, mental/physical impairment, or other signs or symptoms suggesting brain metastases.
4. Measurable primary tumour or other measurable intrathoracic lesion according to response evaluation criteria in solid tumours (RECISTs).
5. Performance status of 0–2 on the Eastern Cooperative Oncology Group (ECOG) scale.
6. Scheduled to be treated with a platinum-based dual agent chemotherapy regimen administered at 3 week intervals.
7. 18 years of age or older.
8. Women of childbearing potential must not be pregnant and all participants must use medically appropriate contraception if sexually active.
9. Ability to give study-specific informed consent.
10. Ability to tolerate PET imaging required by protocol, to be performed at an ACRIN-qualified facility.
11. Laboratory testing (within 4 weeks of registration) including at a minimum CBC, glucose, BUN, creatinine, PT, PTT and liver function tests (to include at minimum alkaline phosphatase) to demonstrate that there are no contraindications for chemotherapy.

Exclusion criteria

1. Small cell carcinoma histology.
2. Pure bronchioloalveolar carcinoma histology.
3. Prior thoracic radiotherapy, lung surgery or chemotherapy within 3 months prior to inclusion in the study.
4. Poorly controlled diabetes (defined as fasting glucose level >200 mg/dl) despite medications.
5. Prior malignancy other than basal cell/squamous cell carcinoma of the skin, carcinoma in situ or other cancer, from which the participant has been disease free for less than 3 years.
6. Pregnancy or participants of reproductive potential who are sexually active and not willing/able to use medically appropriate contraception.
7. Planned to undergo chemoradiotherapy.
8. Clinical or radiographic signs of post-obstructive pneumonia.
9. Symptomatic brain metastases.
10. Maximum diameter of the chest lesion <2 cm.

Summary of study design

The trial will examine the association between changes in tumour FDG uptake during chemotherapy and patient survival. Furthermore, it will determine the test–retest reproducibility of quantitative measurements of tumour FDG uptake. The trial will also evaluate the time course of changes in tumour glucose metabolism during chemotherapy and measure changes in tumour FDG uptake after one and two cycles of chemotherapy, because the optimal time point to predict patient outcome by FDG-PET is currently unknown. Since it is not practical for participants to undergo a total of four[4] PET/CT scans (two prior to therapy and two during therapy), study participants will be randomised into two groups. Group A will undergo two PET scans prior to therapy and one PET scan after the first chemotherapy cycle. Group B will undergo one PET scan prior to therapy and two PET scans during therapy (after the first and the second chemotherapy cycles). For both groups A and B, follow-up CT imaging after every other chemotherapy cycle will be used to determine best clinical response according to RECIST criteria. The participant's treating oncologist will be contacted every three months for 1 year or until death, whichever occurs first, to obtain observational data to determine the primary end-point of one-year overall survival.

Accrual goal

Total of 228 participants will be enrolled into the study at a minimum of 8 institutions. Of the 228 eligible participants, 57 participants will be assigned to group A and the remaining 171 participants will be assigned to group B.

REFERENCES

1. Kelloff GJ, Krohn KA, Larson SM, et al. The progress and promise of molecular imaging probes in oncologic drug development. *Clin Cancer Res* 2005;**11**(22):7967–85.
2. Kelloff GJ, Hoffman JM, Johnson B, et al. Progress and promise of FDG-PET imaging for cancer patient management and

oncologic drug development. *Clin Cancer Res* 2005;**11**(8):2785–808.
3. Richter WS. Imaging biomarkers as surrogate endpoints for drug development. *Eur J Nucl Med Mol Imaging* 2006;**33** (Suppl. 1):6–10.
4. Miller JC, Pien H, Sorensen AG. Imaging biomarker applications in oncology drug development. *Drug Inform J* 2007;**41**:561–72.
5. Jaffe CC. Measures of response: RECIST, WHO, and new alternatives. *J Clin Oncol* 2006;**24**(20):3245–51.
6. Shankar L, Menkens A, Sullivan D. Molecular imaging in cancer. In: Kaufman Howard L et al., editors. *Molecular targeting in oncology (cancer drug discovery and development)*. Humana Press; 2008. p. 675–92.
7. El-Maraghi RH, Eisenhauer EA. Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. *J Clin Oncol* 2008;**26**(8):1346–54 [Epub 2008 February 19, Review].
8. Chan JK, Ueda SM, Sugiyama VE, et al. Analysis of phase II studies on targeted agents and subsequent phase III trials: what are the predictors for success? *J Clin Oncol* 2008;**26**(9):1511–8 [Epub 2008 February 19].
9. Llovet JM, Ricci S, Mazzaferro V, et al. Sorafenib in advanced hepatocellular carcinoma. *New Engl J Med* 2008;**359**:378–90.
10. Grothey A, Hedrick EE, Mass RD, et al. Response-independent survival benefit in metastatic colorectal cancer: a comparative analysis of N9741 and AVF2107. *J Clin Oncol* 2008;**26**(2):183–9.
11. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* 2000;**92**:205–16.
12. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumors: Revised RECIST guideline (version 1.1). *Eur J Cancer* 2009;**45**(2):228–47.
13. Weber WA. Positron emission tomography as an imaging biomarker. *J Clin Oncol* 2006;**24**(20):3282–92.
14. Van den Abbeele AD, Badawi RD. Use of positron emission tomography in oncology and its potential role to assess response to imatinib mesylate therapy in gastrointestinal stromal tumors (GISTs). *Eur J Cancer* 2002;**38**(Suppl 5):S60–5.
15. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor markers prognostic studies (REMARK). *J Natl Cancer Inst* 2005;**97**(16):1180–4.
16. Hayes DF, Bast RC, Desch CE, et al. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. *J Natl Cancer Inst* 1996;**88**(20):1456–66.
17. Ballman KV, Buckner JC, Brown PD, et al. The relationship between six-month progression-free survival and 12-month overall survival end points for phase II trials in patients with glioblastoma multiforme. *Neuro Oncol* 2007;**9**(1):29–38.
18. Buyse M, Burzykowski T, Carroll K, et al. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *J Clin Oncol* 2007;**25**:5218–24.
19. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non small-cell lung cancer. *J Natl Cancer Inst* 2007;**99**(19):1455–61.
20. Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P. Relation between tumor response to first line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. *Lancet* 2000;**356**:373–8.
21. Burzykowski T, Buyse M, Piccart-Gebhart MJ, et al. Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. *J Clin Oncol* 2008;**26**:1987–92.
22. Goffin J, Baral S, Tu D, Nomikos D, Seymour L. Objective responses in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval. *Clin Cancer res* 2005;**11**(16):5928–34.
23. Benjamin RS, Choi H, Macapinlac HA, et al. We should resist using RECIST, at least in GIST. *J Clin Oncol* 2007;**25**(13):1760–4.
24. Ratain MJ, Eckhardt SG. Phase II studies of modern drugs directed against new targets: if you are fazed, too, then resist RECIST. *J Clin Oncol* 2004;**22**(22):4442–5.
25. Ratain MJ. Phase II oncology trials: let's be positive. *Clin Cancer Res* 2005;**11**:5661–2.
26. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996;**125**(7):605–13.
27. Byrne MJ, Nowak AK. Modified RECIST criteria for assessment of response in malignant pleural mesothelioma. *Ann Oncol* 2004;**15**(2):257–60.
28. Young H, Baur R, Cremerius U, et al. Measurement of clinical and subclinical tumour response using [18F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. *Eur J Cancer* 1999;**35**:1773–82.
29. Shankar LK, Hoffman JM, Bacharach S, et al. Consensus recommendations for the use of 18F-FDG PET as an indicator of therapeutic response in patients in National Cancer Institute Trials. *J Nucl Med* 2006;**47**:1059–66.
30. O'Connor JP, Jackson A, Parker GJ, Jayson GC. DCE-MRI biomarkers in the clinical evaluation of antiangiogenic and vascular disrupting agents. *Brit J Cancer* 2007;**96**:189–95.
31. Rosen M, Schnall M. Dynamic contrast-enhanced magnetic resonance imaging for assessing tumor vascularity and vascular effects of targeted therapies in renal cell carcinoma. *Clin Cancer Res* 2007;**13**(2 Suppl.).
32. Weber WA. Use of PET for monitoring cancer therapy and for predicting outcome. *J Nucl Med* 2005;**46**:983–95.
33. Langen A, Hoekstra OS, Dingemans A, et al. Response monitoring with positron emission tomography (PET) in patients with advanced non-small cell lung cancer (NSCLC) treated with bevacizumab and erlotinib: a phase II study. *Eur J Cancer* 2007;**5**(Suppl.).
34. Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Stat Med* 1989;**8**:431–40.
35. Freedman LS, Graubard BI, Schatzkin S. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* 1992;**11**:167–78.
36. Korn EL, Albert PS, McShane LM. Assessing surrogates as trial endpoints using mixed models. *Stat Med* 2005;**24**:163–82.
37. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998;**54**:1014–29.
38. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Appl Stat* 2001;**50**:405–22.
39. Burzykowski T, Molenberghs G, Buyse M. The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. *J Royal Stat Soc A* 2004;**167**:103–24.
40. Bruzzi P, Del Mastro L, Sormani MP, et al. Objective response to chemotherapy as a potential surrogate end point of survival in metastatic breast cancer patients. *J Clin Oncol* 2005;**23**:5117–25.
41. Sargent DJ, Wieand HS, Haller DG, et al. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. *J Clin Oncol* 2005;**23**: 8664–70.
42. Sargent DJ, Patiyil S, Yothers G, et al. End points for colon cancer adjuvant trials: observations and recommendations

based on individual patient data from 20,898 patients on 18 randomized trials from the ACCENT group. *J Clin Oncol* 2007;**25**:4569–74.

43. Morgan B, Thomas AL, Drevs J, et al. Dynamic contrast-enhanced magnetic resonance imaging as a biomarker for the pharmacological response of PTK787/ZK 222584, an inhibitor of the vascular endothelial growth factor receptor tyrosine kinases, in patients with advanced colorectal cancer and liver metastases: results from two phase I studies. *J Clin Oncol* 2003;**21**(21):3955–64 [Epub 2003 September 29].
44. Hecht JR, Trarbach T, Jaeger E, et al. A randomized, double-blind, placebo-controlled, phase III study in patients (Pts) with metastatic adenocarcinoma of the colon or rectum receiving first-line chemotherapy with oxaliplatin/5-fluorouracil/leucovorin and PTK787/ZK 222584 or placebo (CONFIRM-1). *J Clin Oncol* 2005;**23**(16S, Part I or II, June 1 Supplement):3.