Evaluation of the Efficacy of Targeted Imaging Agents

Michael M. Graham¹ and Wolfgang A. Weber²

1. Department of Radiology, University of Iowa Hospitals and Clinics, Iowa City, IA, USA.
2. Department of Radiology, Memorial Sloan-Kettering Cancer Center, New York, NY, USA.

First and corresponding author:

Michael M. Graham, PhD, MD
Division of Nuclear Medicine
Department of Radiology, Room 3863 JPP
University of Iowa Hospitals and Clinics
200 Hawkins Drive
Iowa City, IA 52242

Tel: (319) 356-3380; Fax (319) 356-2220
michael-graham@uiowa.edu

Word count: 5778 (exclusive of title page and references)

Running title: Efficacy of Targeted Imaging Agents
ABSTRACT

This paper is an adaptation of the hierarchical model of the efficacy of diagnostic imaging systems by Fryback and Thornbury (1). The original scheme was designed to evaluate new medical imaging systems, but is less successful when applied to evaluation of new radiopharmaceuticals. The new proposed scheme is specifically directed toward the evaluation of targeted imaging agents. The six levels of the new scheme are: In-vitro characterization, In-vivo animal studies, Initial human studies, Impact on clinical care (change in management), Impact on patient outcome, and Societal efficacy. These levels, particularly the first four, implicitly define the sequence of studies that are necessary to move a new agent from the radiochemistry synthesis laboratory to the clinic. It is suggested that completion of Level Four (Impact on clinical care) should be sufficient for initial approval and reimbursement. It is hoped that these suggestions will help streamline the process and assist in bringing new, targeted radiopharmaceuticals to approval in the next few years.

Key words: Efficacy, Evaluation, Radiopharmaceutical, FDA Approval Process
INTRODUCTION

In 1991 Fryback and Thornbury (1) published an important paper that defined a hierarchical model of efficacy of diagnostic imaging. This six level scheme has become a widely accepted guideline for the evaluation of new diagnostic tests. The levels they defined are: 1. Technical efficacy; 2. Diagnostic accuracy; 3. Diagnostic thinking; 4. Therapeutic efficacy; 5 Patient outcomes; and 6. Societal efficacy.

Fryback and Thornbury described a systematic approach to establish the diagnostic accuracy of an imaging test and to determine its impact on therapeutic efficacy and eventually its benefit for society. About 10 years after the publication of this influential paper the concept of molecular imaging was introduced and has become a major focus of imaging research during the last 15 years. While several definitions of molecular imaging probes have been proposed, an essential goal of molecular imaging is repetitive and quantitative assessment of the expression or function of molecular targets. Detecting the presence or absence of a certain disease (diagnostic accuracy) remains an important goal of molecular imaging, but molecular imaging also addresses several other clinical problems, such as assessing prognosis, predicting and monitoring response to molecularly targeted interventions, and assessment of distribution and binding occupancy of receptors. The important concepts introduced by Fryback and Thornbury, are only partially applicable for the evaluation of molecular imaging. Conversely, validation of molecular imaging requires additional steps that are not described by Fryback and Thornbury. Several modifications of the Fryback and Thornbury scheme have been published, but none have addressed the evaluation of new radiopharmaceuticals (2). A new hierarchical scheme needs to be defined to help developers as well as regulators understand how to evaluate the efficacy of these new agents.
The agents that require this new scheme are ones that target a specific receptor or metabolic pathway in tissue where the imaging results can make a major difference in how a patient will be treated. Examples of such agents include new Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) agents designed to bind to cell surface receptors of malignant tumors or ligands binding to amyloid deposits in the cerebral cortex. Often the goal is to identify tumors with high levels of receptor expression that can then be treated with a similarly targeted therapeutic agent. This concept is often referred to as “Theranostics”. Another goal is to monitor pharmacodynamic effects of targeted drugs in order to predict and to determine if there is a response to therapy.

Because the targeted imaging approach goes beyond the type of diagnostic imaging envisioned by Fryback and Thornbury, the hierarchical scheme needs to be significantly modified in order to allow for efficient and appropriately designed clinical trials of molecular imaging. In making this modification, we have retained the underlying philosophy, starting with fundamental assessment of the technical details of the test, then moving to practical approaches to evaluate the test in clinical practice, and finally looking at the impact of the test on patient outcome and on society.

The hierarchy suggested in this paper (see Table 1) follows the temporal sequence that is necessary to develop a new targeted agent: initial evaluation in the laboratory, studies in animals, human studies that are needed to obtain regulatory approval, and then use in clinical management. The suggested approach is not intended to be completely comprehensive, but makes recommendations for the most important steps at each level and, in the discussion below also identifies important steps or testing that are inappropriate or not feasible for evaluating new targeted agents.
It takes into account that significant preclinical evaluation is needed for molecularly targeted imaging probes to ensure that the imaging probe specifically visualizes and quantifies its target. The new hierarchy also reduces emphasis on diagnostic accuracy studies, because the quantitative nature of molecular imaging goes beyond the binary classification of presence or absence of disease that is fundamental to the concept of diagnostic accuracy.

Another reason for de-emphasizing the concept of diagnostic accuracy is that in oncologic imaging there is frequently no unbiased reference standard to determine the presence or absence of disease. As a consequence, diagnostic accuracy cannot be determined in an unbiased way for many important applications of molecular imaging.

The main goal in writing this paper is to provide a clear pathway for efficient development, evaluation, and application of new targeted imaging agents. It is not a simple process and there is a definite need to develop better strategies and understanding of the processes, which should lead to accelerated clinical application of targeted imaging agents in clinical applications.

**Level 1. In Vitro characterization**

In the Fryback and Thornbury hierarchy this level is “Technical efficacy”, and is concerned with image quality (resolution, modulation transfer function, etc.), a type of assessment, which is inappropriate for imaging agents. Instead, this level is concerned with necessary preliminary in-vitro detailed characterization that is essential for a new agent to demonstrate that it is likely to successfully bind to the target of interest. Initial studies are usually done in cell-free systems, such a columns with bound receptors. More complex studies require cell suspensions or tissue culture. Chemical and metabolic stability can be studied by incubation in buffer, cell medium, and blood or plasma, after which the degradation products,
including free radionuclide, can be detected, identified and quantified. A recent paper by Wynendaele et al. contains a detailed description of many additional aspects of vitro characterization of radiopharmaceuticals (3).

**Level 2. In Vivo animal studies**

In the Fryback and Thornbury hierarchy this level is “Diagnostic accuracy”. The major emphasis was determination of test accuracy in terms of sensitivity and specificity, including use of Receiver Operating Characteristic (ROC) curves.

Sensitivity and specificity of imaging tests has been extensively studied in the literature. Calculation of sensitivity and specificity is appropriate if there is an established reference standard (“gold standard”) that is compared with the results of the imaging test. An example is the evaluation of solitary pulmonary nodules by FDG PET using histology as the reference standard. However, for whole body tumor staging - one of the most common applications of imaging tests – histology generally cannot be used to evaluate the accuracy of all sites that are considered as positive or negative by a novel imaging modality. Therefore, a common approach has been to use a consensus interpretation of all available imaging studies (the new imaging modality and the conventional imaging tests combined) as a reference standard. However, histologic evaluation can only be performed for sites that are positive on at least one imaging test and perhaps a limited number of sites that are negative on imaging. Consequently, there is always a verification bias for sensitivity and specificity, which causes a systematic overestimation of both sensitivity and specificity. Therefore, the concept of sensitivity and specificity is problematic. For example, it is a common observation that the sensitivity of novel imaging systematically decreases over time. The sensitivity of $^{111}$In-octreotide imaging for
neuroendocrine tumors was described to be close to 100% when the technology was introduced in the 1990s (4). However, in a more recent study comparing $^{111}$In-octreotide SPECT imaging with $^{68}$Ga-DOTATOC PET/CT, its sensitivity was only about 50%. This striking difference is obviously due to the fact that, in the more recent paper $^{111}$In-octreotide SPECT was compared with a more sensitive imaging modality and that therefore the denominator for calculation of sensitivity markedly increased (5).

An additional problem of using the terms sensitivity and specificity for cancer staging is that both are dependent on the number of sites analyzed. This can be illustrated by many studies evaluating PET/CT for lymph node staging. Imaging findings can be correlated with the presence or absence of lymph node metastases on a whole patient basis, on the location of lesions in relatively large lymph node regions (for example, left and right side of the pelvis), smaller lymph node regions (for example left internal iliac nodes), or on individual potentially involved nodes. When the size of the analyzed sites decreases, their total number generally increases. As a consequence, the specificity of an imaging test will be higher if smaller lymph node regions are analyzed, because the number of false positive findings will be divided by a larger total number of true negative sites.

Even more fundamental, there can be cases that can be considered true positive and false negative at the same time. For example, consider a $^{11}$C-choline PET/CT scan in a patient with biochemical recurrence of prostate cancer that shows a choline avid left internal iliac node. Let’s also assume that the patient then undergoes salvage lymphadenectomy that finds a lymph node metastases in the left internal and the left common iliac region. On a patient basis the scan is true positive for lymph node metastasis, it is also true positive if lymph nodes are classified as “left
and right iliac nodes”, but the study is false negative and true positive if the iliac lymph node regions are subdivided into external and internal iliac nodes.

In summary, determination of sensitivity requires knowledge of all “true positive” sites of disease. For whole body cancer staging, positive sites can only be identified by imaging tests, making estimates of sensitivity and specificity inherently biased. With improvements in instrumentation, sensitivity and specificity consequently change over time.

Calculation of specificity requires knowledge of all true negative sites. Because the number of true negative sites is most of the body, specificity becomes critically dependent on how many regions the body is divided into. It is possible to determine specificity on an individual patient basis, but like sensitivity, it is dependent on the capabilities of instrumentation used as reference and cannot be accurately measured.

Instead of sensitivity and specificity, in the proposed hierarchy the emphasis at this level is on accurate characterization of the behavior of the new agent in animals prior to human administration. A potential pitfall with these studies is that the agent may behave quite differently in animals and in humans. Accordingly, it may be best to move rapidly to early human studies, rather than investigating multiple animal species.

The essential testing done at this level is defined in the sub-headings of level 2: in vivo stability, target specificity, pharmacokinetics, radiochemistry optimization, radiation dosimetry, and toxicity.

In vivo stability is essential. There are numerous circulating enzymes in blood that may rapidly degrade the new agent. Slow degradation may be acceptable, but has to be defined and the behavior of the metabolites should be characterized.
Target specificity is also essential. Note that this is very different than test performance specificity. Target specificity of a novel targeted agent is difficult to assess in-vitro and only when it is injected and imaged in an animal will it become clear if there is a high target-to-background ratio. Additional studies are also often needed to show lack of uptake in tissue that lacks the target receptor. Target specificity is usually evaluated with transplanted tumors in immuno-incompetent mice. Ideally one tumor has receptors and shows good uptake, while another type of similar tumor lacking receptors shows no uptake. This type of study firmly demonstrates agent specificity and makes it unlikely that increased capillary permeability as the primary mechanism of uptake. The other major approach for demonstration of specificity is by blocking uptake using relatively high levels of known receptor targeting ligand. This is often a stable non-radioactive version of the study agent. Note that in such studies the high dose may have physiologic effects that can perturb the delivery of the probe molecule by effects other than simply blocking the receptors.

Pharmacokinetics is quite important. How rapidly the agent is taken up into the target and how rapidly it is excreted will define the timing of imaging and will also define the limits for the acceptable half-life of the radioactive label. If the agent takes days to localize it is not feasible to label it with a short-lived isotope. Pharmacokinetics should also include identification of metabolites, particularly those that retain the radioisotope label. Because labeled metabolites can represent a significant fraction of the detected radioactivity, their appearance and time course should be determined.

Both radiation dosimetry and toxicity studies are essential before an agent can be injected into humans. Because most of these agents are injected at microgram levels and below, it may be practically impossible to reach truly toxic levels even in small animal such as mice.
Therefore it is acceptable to show no acute toxicity in one species (usually mice) with 100 times
the dose (mg/kg) likely to be used in humans and when the anticipated human dose is less than
100 μg (6, 7).

**Level 3. Initial human studies**

In the Fryback and Thornbury hierarchy this level is “Diagnostic thinking efficacy”. This
is relatively vague and difficult to quantitate, since it addresses how the test results change the
thinking of the referring clinicians. In the proposed hierarchy this level is concerned with
demonstrating safety of the agent in humans and preliminary demonstration of efficacy. These
Phase 1 studies are typically closely monitored and may be conducted in patients or normal
volunteers. These studies are designed to determine the metabolism and pharmacokinetics of the
drug in humans, any side effects, and, if possible, to gain early evidence on effectiveness. (8)

An important part of these studies usually consists of measurement of vital signs, ECG and
a limited panel of blood chemistries before the study and then at one or more time points
afterwards. It is also important to ask the subjects if they experienced any symptoms
immediately after the injection or later. It should not be necessary to continue subject follow-up
for longer than about 5 half-lives of the injected agent. Note that the follow-up period should be
based on the biologic half-life of the non-radioactive agent, not the physical half-life of the
radioisotope label. The selection of the blood tests should reflect reasonable postulated toxicity
based on the structure of the compound and on the animal toxicity study.

Human dosimetry can be performed using quantitative pharmacokinetic PET data, and is
necessary before proceeding to Phase 2 clinical efficacy studies. Collection of data and
calculation of radiation dose to various organs and to the whole body can be accomplished, once
quantitative PET data has been obtained at several time points post injection. Similar studies can be done with SPECT, although the methodology is more challenging and accuracy may be lower.

Target specificity for targeted agents can be defined as the ratio of uptake in target tissue divided by uptake in normal tissue. Most of the new agents have relatively high target to background ratios. This ratio is particularly important if there is a companion theranostic agent, when the ligand is labeled with a beta or alpha emitter and the intention is delivery of high radiation dose to the target while avoiding unacceptable radiation dose to normal tissue.

Pharmacokinetic studies involve collection of blood time-activity data, derived from sampling, and of target and normal tissue, typically derived from sequential quantitative imaging. These data are used in the calculation of radiation dosimetry, to determine optimal imaging time after injection, and the appropriate injected dose for the radiopharmaceutical.

Reproducibility determination of the quantitative behavior of a new agent is essential before attempting to use it for assessing response to therapy. Once reproducibility is determined it is possible to define the degree of change in uptake that is significant and often shows improvement or progression of disease. Typically such studies are done by repeating the imaging of the same subject within a few days, with no therapy in the interval.

Although initial biodistribution studies of radiopharmaceuticals previously used in humans can be performed under Radioactive Drug Research Committee (RDRC) oversight, an FDA Investigational New Drug Application (IND) will need to be filed and approved before clinical research can be conducted. First-in-man studies require an IND or an exploratory IND (eIND). An essential part of the IND is a section on Chemistry, Manufacturing, and Controls (CMC). The CMC documentation should be developed at this level. The IND application needs to
include any available information on safety and dosimetry, a CMC section, and a study protocol that describes the proposed trial design in detail.

**Level 4. Impact on clinical care (change in management)**

In the Fryback and Thornbury hierarchy this level is “Therapeutic efficacy”. It was essentially defined as the fraction of tests that resulted in a change of management. Similarly, in the new scheme at this level the emphasis is on change of management in several specific settings.

The studies at this level are Phase 2 and Phase 3 clinical trials (8). The FDA expects diagnostic PET drugs to be produced under the rules associated with Good Manufacturing Practice (GMP) for Phase 3 clinical trials and for subsequent manufacturing for marketing of the new drug.

Diagnosis.

The test may be useful in patients with symptoms, laboratory findings and imaging results that suggest the presence of the target disease. The measure of diagnostic efficacy is yield, i.e. the fraction of patients studied with the new agent who are found to have the target disease. This then should result in initiation of appropriate therapy.

Staging.

In patients with known malignant disease the new agent may be useful for staging, i.e. accurately defining the extent and location of disease. Because of the problems in determining sensitivity and specificity in this setting (as discussed above), we propose that existing and new
imaging tests be compared for discrepant findings that lead to a change in management. For example, if the new imaging test detects a bone metastasis that will change treatment from curative to palliative, a biopsy of this metastasis should be performed and used as the reference for comparison of the two tests. Findings that are concordantly positive or have no impact on patient management do not need to undergo further evaluation. The discrepantly positive and negative relevant findings of the new and existing imaging test can be tested for statistical significance by McNemar’s test. The result of McNemar’s test depends only on the discrepant cases. Therefore, no reference standard is needed for sites that are concordantly positive or negative. McNemars’s test for lesions that were positive on histology but discrepant on the two imaging modalities will show if the new imaging modality is significantly more sensitive than the existing modality. Conversely, McNemar’s test for all lesions that are negative on histology will show if the new modality is more specific than the existing imaging modality. Thus, improvements in sensitivity compared to existing imaging technologies can be determined even if the absolute sensitivity and specificity is unknown for the reasons discussed earlier.

Demonstration of accurate identification of abnormal tissue is often sufficient to show that imaging with a new agent will result in change of management. A recent paper written by FDA personnel (9) stated: “The FDA imaging product guidance recognizes how the clinical usefulness of some imaging information may be obvious in certain clinical settings, such as the staging of cancer or the detection of clinically important pathology”. In addition, the paper states: “In such situations, imaging drug developers are not expected to perform clinical studies that demonstrate again the clinical benefit of the imaging information”.

Response to therapy.
A significant advantage of PET imaging is that uptake can be determined quantitatively. This capability lends itself to assessing response to therapy by quantitative comparison of uptake prior to therapy to uptake at some time afterward. The optimum timing of the follow-up studies and the criteria for response have to be determined in appropriate clinical trials. The major rationale for such studies is that if the studies show lack of response to therapy another treatment regimen can be implemented. Early identification of lack of response may also benefit patients, by limiting the duration patients are exposed to ineffective but potentially toxic drugs. The reference standard in this setting is often change in tumor size, as seen with anatomic imaging using Response Evaluation Criteria In Solid Tumors (RECIST) criteria (10). In addition, survival of patients classified as responders or non-responders can be compared.

Monitoring tumor response to therapy is related to assessment of pharmacodynamic effects of targeted drugs. For example, blocking of androgen receptors by anti-androgens can be imaged with fluorine-18 labeled dihydrotestosterone. The reference standard in this setting can be biopsies showing down-regulation of target dependent signaling pathways that correlate with a decrease of the uptake of the imaging agent.

Evaluation for targeted therapy.

New diagnostic targeted agents are often developed in conjunction with a companion therapeutic agent, which differs only in the radioisotope label, e.g. Ga-68 DOTATATE and Lu-177 DOTATATE, for diagnosis and treatment of neuroendocrine tumors. Diagnostic imaging with the appropriate targeted agent is essential before administration of the therapeutic companion to assure high uptake into the targets and acceptable uptake into normal tissue. In this setting the accurate quantitation of radiotracer uptake and calculation of radiation dose with
a clinically feasible imaging protocol is the key outcome parameter, not the sensitivity and specificity for detection of metastases.

**Level 5. Impact on patient outcome**

In the Fryback and Thornbury hierarchy this level is “Patient outcome efficacy”. This includes the fraction of patients improved because of the test (compared without the test), morbidity avoided by having the test, change in quality-adjusted life years (QALY), and the cost per QALY saved (cost effectiveness). In the new proposed hierarchy the goals are similar.

Assessment of implementation of the change in management.

At level 4 the major goal is assessment of the frequency that clinical management is changed in response to the information obtained with the new test. Usually this is done by requiring the treating physician to record the treatment plan prior to seeing the results of the new test and then recording the new plan, once the results are available. This is really looking at the change of intended management. At level 5 a more rigorous criterion is required, confirmation that the changes were actually implemented and were appropriate.

Assessment of correctness of the change in management

This is not easily determined. It either requires a panel of experts to assess the situation and determine appropriateness of the change or follow-up to see how well the patient does following the change. Both approaches have inherent weaknesses. The experts may not have sufficient knowledge or information to accurately determine appropriateness in all settings, and follow-up can not reveal what would have happened if the test had not been done. In addition,
subsequent testing and changes in management may occur which are not related to the original diagnostic test.

It would be ideal to determine outcome (relapse-free survival, overall survival) in a randomized trial with patients who had and did not have the new test. The practical problems with implementation of this approach are lack of clinical equipoise and lack of control of subsequent treatment decisions. Many of these new agents show very high target to background ratios and after inspection of a few examples, it is often intuitively compelling that the new agent is superior to prior approaches. This makes it very hard to recruit subjects for a randomized clinical trial. Even if recruited, many subjects will attempt to have the new test done outside the scope of the trial. This can have the effect of making a rigorous survival trial impossible. The other factor that also affects the feasibility and meaningfulness of a survival trial is lack of control over subsequent treatment decisions. The only setting where subsequent treatment is well controlled is in a therapeutic clinical trial, when it becomes very problematic to introduce an experimental diagnostic agent to make decisions in terms of change in management.

Although it would be ideal to be able to calculate cost per quality-adjusted life year (QALY), similar problems are encountered, in that there is often lack of uniformity in the subsequent treatment of patients in the months and years following imaging with the new agent and survival and quality of life may be only very loosely related to the test results.

An important limitation of using survival as an endpoint is the very large number of patients needed to demonstrate differences in outcome. An alternative treatment is unlikely to improve survival in all patients that do not respond to the first treatment. The fraction of patients that can potentially improve the outcome of the whole patient population becomes therefore small. For example, if 50% of the patients are classified as non-responders by the new test and
an alternative treatment improves survival in 20% of these patients, only 10% of the patients will ultimately benefit from the use of the new test to assess response. Studies with sufficient statistical power to detect an improvement in overall survival in such a setting will generally require randomization of several hundred patients. In addition, the results are likely to be confounded by patients in the control group that are identified as nonresponders by conventional imaging at a later point in time. These patients are likely to receive alternative treatments as well and some of these patients will likely benefit from the alternative therapy.

In practice, the only feasible way to make meaningful conclusions about survival and cost effectiveness is to model the probability of subsequent events following the test, given knowledge of the fraction of patients in which there was a change in management and knowledge (or assumptions) about appropriate subsequent therapy decisions and outcomes (11) (see Figure 1).

**Level 6. Societal efficacy.**

It is certainly desirable and important to determine if a new test is valuable at the societal level, particularly in extending useful life span and in lowering overall health care costs. As discussed above, explicit determination of these measures is essentially impossible, however may be feasible with appropriate modeling.

**DISCUSSION**

The goal of this paper is to present an organized consensus view of a logical and efficient approach for evaluation of the efficacy of new targeted radiopharmaceuticals. It is built on the framework of a prior publication, “The Efficacy of Diagnostic Imaging” by Dennis G. Fryback
and John R. Thornbury, published in 1991. It is implicit that the levels represent the sequence of studies that are necessary to move a new agent from the radiochemistry synthesis laboratory to the clinic. It is suggested that completion of Level 4 (Impact on clinical care) should be sufficient for initial approval and reimbursement. Work on levels 5 (Impact on patient outcome) and 6 (Societal efficacy) should be addressed once a new agent becomes widely available.

Others have considered the issue of optimal strategy for approval of diagnostic imaging agents. A major issue in prior discussion has been the question of the need for for randomized clinical trials (RCTs). This issue was considered at length by Valk in 2000 (12) and more recently by Hicks et al. in 2012 (13) in a critique of a paper on a review of RCTs in PET (14). Both papers (Valk and Hicks) clearly make the point that RCTs are not necessary, feasible, or effective in the assessment of new radiopharmaceuticals.

Vach et al. (15) have also addressed this question and have considered the problem of “Generating evidence for clinical benefit of PET/CT in diagnosing cancer patients”. They consider two different RCT designs, but conclude that practical issues of clinical equipoise, time to conduct a trial, and the need for multiple trials to address all the possible scenarios make the RCT approach frequently impractical. They propose that decision modeling following determination of actual change in management is an efficient way to generate evidence of clinical benefit. This approach depends on making reasonable assumptions about the management changes that were correct, as well as the expected benefit or detriment of change of therapy for both correct and incorrect changes. If a consensus can be reached between the medical researchers and the regulatory authorities on the validity of the assumptions, then it should be feasible to move forward to approval and reimbursement.
Although there is not complete agreement, there is consensus that appropriate observational studies, carefully done, can be sufficient to establish safety and efficacy of a new agent. A Medical Imaging & Technology Alliance (MITA) conference addressed this question, although with a more specific focus on research endpoints appropriate for Medicare coverage of new PET radiopharmaceuticals (16). At the outset of the conference there was general agreement on specific issues presented by Louis Jacques, who was then head of the CMS Coverage and Analysis Group. The key principle presented was: “The potential benefits of diagnostic tests relate to their providing information to optimize treatment plans and, thereby, improve clinical care and health outcomes.” The key take-home point from the conference was “Coverage of new PET radiopharmaceuticals should depend on clinical evidence of effect on intermediate endpoints, such as a beneficial change in clinical management (i.e., change in subsequent therapeutic or diagnostic interventions) that can be linked to improved health outcomes.” These same principles should be applied to approval by the FDA, as well as coverage by CMS. While the link to improved health outcomes could conceivably be done with RCTs and/or long-term observational studies, the only practical way to make the link is with well-designed decision modeling studies.

Change in management study trial design depends on how the new agent is likely to be used in clinical practice. For example, the four major possibilities for an imaging agent used for cancer staging are: 1. The new test may detect previously unknown disease in patients and treatment is undertaken as a result. 2. The new test may detect unsuspected distant disease and a futile operation or other treatment is avoided. 3. Unnecessary treatment is avoided in patients who test negative for disease. 4. Minimal disease may be detected, of no clinical significance, e.g. stable prostate or thyroid cancer and unnecessary treatment is undertaken. In addition, study
design needs to consider if the test is a replacement for or in addition to an existing test, the potential consequences of both positive and negative results, and if intended changes were actually implemented. Although change in management is a potential powerful tool for assessing the efficacy and need for a new agent, its measurement is not trivial and must be approached with care. (17)

In addition to molecular imaging agents intended for guiding clinical management, another class of agents is intended for assessing the pharmacokinetics and pharmacodynamic behavior of novel therapeutic drugs during early development (18). This characterization can have a major effect in the decision making about subsequent development of the new drug. These agents often include a radiolabeled version of the therapeutic drug or may be targeted at a specific metabolic pathway presumed to be blocked or stimulated by the therapeutic drug. Often these agents are never intended for clinical use but are essential in the initial characterization of a new therapeutic drug. These agents need to be characterized carefully at levels 1, 2 and 3 to demonstrate the validity of the behavior of the agent. These agents may be specific for the study drug (e.g. showing the biodistribution and tumor uptake of a specific antibody). However others may be more generic and show the expression of a target for multiple drugs, such as the density of free estrogen or androgen receptors.

**CONCLUSION**

Currently, there are several targeted radiopharmaceuticals being developed by multiple academic and commercial groups throughout the United States and the world. Many of these new agents have significant potential to make a real difference in how medicine is practiced in the future and are likely to be a major part of true “personalized medicine”. However, because
of uncertainty and inconsistency regarding the optimum pathway from discovery to clinical application, the development of these agents is less efficient, more expensive and slower than it should be. It is hoped that the suggestions that are implicit in the levels of efficacy presented in this paper will help streamline the process and assist in bringing many of these new agents to approval in the next few years.

Disclosure

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The authors wish to thank the members of the Society of Nuclear Medicine and Molecular Imaging FDA Task Force for numerous comments and critique during the formulation and writing of this paper. This paper has been formally endorsed by the SNMMI FDA Task Force.
References


8. Federal Code of Regulations 21CFR 312.21
https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=312.21


<table>
<thead>
<tr>
<th>New Proposed Hierarchical Scheme</th>
<th>Fryback and Thornbury Scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. In Vitro characterization</strong></td>
<td><strong>1. Technical efficacy</strong></td>
</tr>
<tr>
<td>Kon, Koff, Kd, Bmax, IC50</td>
<td></td>
</tr>
<tr>
<td>Partition coefficient, Binding potential</td>
<td></td>
</tr>
<tr>
<td>Labeling efficiency, yield</td>
<td></td>
</tr>
<tr>
<td>In-vitro label stability</td>
<td></td>
</tr>
<tr>
<td><strong>2. In Vivo animal studies</strong></td>
<td><strong>2. Diagnostic accuracy</strong></td>
</tr>
<tr>
<td>In-vivo stability</td>
<td></td>
</tr>
<tr>
<td>Target vs non-target tissue specificity</td>
<td></td>
</tr>
<tr>
<td>Pharmacokinetics</td>
<td></td>
</tr>
<tr>
<td>Radiochemistry optimization</td>
<td></td>
</tr>
<tr>
<td>Dosimetry, Toxicity</td>
<td></td>
</tr>
<tr>
<td><strong>3. Initial human studies</strong></td>
<td><strong>3. Diagnostic thinking</strong></td>
</tr>
<tr>
<td>Safety, Dosimetry, Target specificity</td>
<td></td>
</tr>
<tr>
<td>Tracer stability in-vivo</td>
<td></td>
</tr>
<tr>
<td>Pharmacokinetics (including metabolites)</td>
<td></td>
</tr>
<tr>
<td>Reproducibility</td>
<td></td>
</tr>
<tr>
<td>Determination of sensitivity, specificity, PPV, NPV</td>
<td></td>
</tr>
<tr>
<td>Investigational New Drug (IND) application</td>
<td></td>
</tr>
<tr>
<td>CMC (Chemistry, Manufacturing, and Controls) development</td>
<td></td>
</tr>
<tr>
<td><strong>4. Impact on clinical care (change in management)</strong></td>
<td><strong>4. Therapeutic efficacy</strong></td>
</tr>
<tr>
<td>Diagnosis (patients with suspicion of disease)</td>
<td></td>
</tr>
<tr>
<td>Staging (patients with known disease)</td>
<td></td>
</tr>
<tr>
<td>Response to therapy (imaging pre and post therapy)</td>
<td></td>
</tr>
<tr>
<td>Evaluation for targeted therapy</td>
<td></td>
</tr>
<tr>
<td>CGMP implementation</td>
<td></td>
</tr>
<tr>
<td><strong>5. Impact on patient outcome</strong></td>
<td><strong>5. Patient outcomes</strong></td>
</tr>
<tr>
<td>Assessment of implementation of change in management</td>
<td></td>
</tr>
<tr>
<td>Assessment of correctness of change in management</td>
<td></td>
</tr>
<tr>
<td>Survival with and without test (Kaplan-Meier plots)</td>
<td></td>
</tr>
<tr>
<td>Cost per quality-adjusted life year (QALY)</td>
<td></td>
</tr>
<tr>
<td><strong>6. Societal efficacy</strong></td>
<td><strong>6. Societal efficacy</strong></td>
</tr>
<tr>
<td>Cost-benefit analysis</td>
<td></td>
</tr>
<tr>
<td>Risk-benefit analysis</td>
<td></td>
</tr>
<tr>
<td>Post-approval monitoring for side effects</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.
Figure 1. General scheme for modeling the outcomes expected with a new medical test. While it is straightforward to define disease incidence and test sensitivity and specificity, it is more challenging to define the impact of treatment in terms of probability of cure and probability of complications for each test result pathway. It is also challenging to estimate the expected survival years and overall costs for each pathway.