Methods and Limitations of Assessing New Noninvasive Tests

Part II: Outcomes-Based Validation and Reliability Assessment of Noninvasive Testing

Rory Hachamovitch, MD, MSc; Marcelo F. Di Carli, MD

Outcomes-based approaches are the preferred methodology for technology validation. As discussed in part I of this review, the difficulties of performing an unbiased diagnostic evaluation are increasingly appreciated. An outcomes-based approach is advantageous in that it mimics the clinical application of testing. By risk stratifying patients, the results of this approach can be applied directly to clinical practice. Nonetheless, outcomes-based technology validation is not without challenges and limitations. These issues include the need for multivariable modeling for observational data, end point selection, and limitations of estimating posttest risk. Finally, it is increasingly appreciated that the future “gold standard” for outcomes-based assessments will be demonstrating whether imaging can identify which therapeutic approach optimizes patient benefit rather than merely identifying risk.

**Outcomes-Based Validation**

**Study Design**

Requirements of imaging studies include a relevant study population, comparison with an appropriate control group, and follow-up for outcomes. Designing randomized controlled trials (RCTs) that address imaging questions is challenging. The RCT may utilize imaging results as inclusion criteria or the basis for therapeutic assignment. Randomization to strategies with versus without imaging is problematic. The use of imaging per se affects outcomes only if a therapy is triggered; hence, in these studies, therapy and imaging results must be linked. Comparisons of imaging methods/modalities must mandate that test results are acted on rather than “available to the physician” (thus, an emphasis on efficacy rather than effectiveness). Because observational studies are far more common, they will be our focus. Although limited by inherent design flaws (eg, selection biases, potentially spurious observations, missing covariates), patients in observational studies better represent those seen in practice.

**Patient Selection**

Unlike diagnostic-based validation in which only patients referred to a gold standard after testing are included, in prognostic-based approaches all eligible patients are followed up at a preselected time point after testing to determine their status relative to the events of interest. The selection of the cohort for a given study can be challenging as issues of power (eg, increasing risk and event rates in patients with versus those without prior coronary artery disease), patient availability for follow-up, generalizability to other settings, and impact of posttest treatment and biases occur. Study end points should be clinically relevant, easily ascertainable, sensitive to the effects under evaluation, and verifiable. In prognostic studies, 2 end points predominate: cardiac death and all-cause death. The former is profoundly limited and susceptible to misclassification bias (for review, see Lauer et al). Death certificates are often erroneous, and the gold standard (autopsy) is rarely performed. The mechanism of demise and the actual cause of death are often confused. Death in patients with coronary artery disease is usually assumed to be cardiac. Hence, the use of cardiac death as an end point has significant flaws and unacceptably high error rates.

All-cause death is a “harder” end point, relatively unbiased, easily ascertained, and the most valid. However, this end point has limitations. Do studies modeling all-cause death in patients aged <55 years, in whom <20% to 25% of deaths are cardiac, have the same meaning as those with an older cohort? In patients undergoing preoperative evaluation, with more frequent and serious comorbidities, noncardiac mortality will likely be higher, obfuscating all-cause death rates. Therefore, we recommend all-cause death as a primary end point but cardiac death as a secondary end point. Because the social security death index is used to determine all-cause death, a “follow-up” study may not need to contact patients. This is potentially problematic because if the use of early revascularization is not known, its impact on outcomes cannot be determined.
Hard events, such as combined cardiac death and nonfatal myocardial infarction (MI), have the aforementioned limitations and those of MI verification2,5 (eg, nonblinding of ascertaining observers and defining and identifying MI) and thus lack rigor but may be helpful as secondary end points. Interestingly, predictors of these 2 events differ, and therefore a study’s relative proportions of MI and cardiac death will influence observed test performance.2,5 This differential prediction may aid posttest therapeutic decision making, but further investigation is necessary. However, because these 2 events are not independent, modeling of MI alone requires a more advanced Cox proportional hazards methodology.7

Composite end points (eg, combined hard events, late revascularizations, catheterizations, hospitalizations) have numerous limitations. Although their use reduces needed sample size and accrual time, they assume risk homogeneity among the component events. The differential impact of treatment on each end point yields ambiguous and problematic results, including reduced power due to unaffected end points. Physician-influenced outcomes (catheterization, revascularization, hospitalization) are often more amenable to treatment than harder end points, and trials using these outcomes more frequently have positive results.

A study of the prognostic value of computed tomographic angiography (CTA) in 100 patients used a composite end point of cardiac death, MI, unstable angina, and revascularization.8 Revascularizations represented 24 of 32 events that occurred. Although the authors concluded that CTA predicted the composite end point, CTA actually predicted revascularization—an end point it triggered—but predicted other end points questionably. Hence, a study reporting composite end points when one event predominates is potentially misleading. Although broadening the spectrum of events may enhance the value of the study, care must be exercised. The component events should be presented as secondary end points, with event rates for individual as well as combined end points reported.

Test Performance Metrics
Prognostic studies avoid reporting sensitivity, specificity, and predictive values and focus on aggregate and/or annualized event rates for the overall cohort and the subgroups with normal and abnormal test results. Risk stratification, evidenced by increasing event rates as a function of worsening test abnormalities (eg, small versus large myocardial perfusion abnormalities), requires demonstration. Historically, demonstration of “low” event rates after normal studies has been a mantra because it reassures physicians about the safety of managing these patients conservatively.2 Two standards for low risk are used: a hard event rate <1% per year and, more recently, a <1% cardiac death rate per year. The latter lends itself to comparisons of risks and benefits for revascularization as a therapeutic approach. Numerous stress imaging reports have claimed <1% hard event risk in normal studies;2 stress cardiovascular magnetic resonance imaging, computed tomography, positron emission tomography (PET), and perfusion echocardiography will eventually require similar studies. However, these observational reports of unadjusted risk are limited, and no evidence suggests that this approach enhances patient outcomes. In addition, newer paradigms question the validity of fixed thresholds in defining “low risk.”2

Multivariable Modeling and Risk-Adjustment Techniques
Prognostic imaging studies are commonly observational, compromised by multiple factors (eg, bias, confounding) that necessitate multivariable techniques to enhance validity and accuracy. (Multivariable is defined as the simultaneous modeling of multiple variables; multivariate is defined as the simultaneous consideration of multiple outcomes.) Several important concepts will be reviewed, and the reader will be referred to reviews in this area.7

Briefly, multivariable models express the association between an end point (y) and a combination of factor(s) (x’s).7 Modeling may serve 2 purposes: descriptive (determining whether y is associated with 1 or more xs and assessing the effect of x on y after adjustment for other factors) and predictive (predicting the value of y at specific values of x).

Types of Models
The first step in modeling is selecting a model form. Linear regression is used for continuous end points (eg, blood pressure, exercise tolerance), logistic for dichotomous end points (eg, disease presence), and Cox proportional hazards for survival analyses (survival analyses model time to event rather than the event occurrence). In specific situations, it may be advantageous to model survival with parametric survival or logistic models.

Variable Selection
Ideally, candidate variable selection for model entry is based on clinical experience and judgment, the research question posed, and common sense.7 Importantly, the input to and findings of a model should make clinical sense. Known predictors or confounders merit inclusion, with extraneous variables ideally omitted. Variable selection based on univariate analysis (inclusion of independent predictors) is inappropriate because it wrongly rejects potentially important variables in the setting of uncontrolled confounders; avoiding this error depends on careful data analysis.7 Variable selection based on automated algorithms, including stepwise approaches, also has serious limitations, introducing numerous potential errors while including significant noise and failing to include >50% of actual predictors.7

Different studies using similar cohorts and end points may identify different models and predictors. This is due to varying patient selection; the variables that were examined; the manner in which they were defined, collected, and coded; and the manner in which the modeling was performed and tested. It may also be due to inherent limitations of classic regression techniques.9

Model Assumptions, Power, and Overfitting
All models are based on multiple assumptions that must be examined7 (Table). Unfortunately, the reader is at the mercy of the authors as to whether these were examined, and most published studies do not state whether this was done.
Interactions are a means of addressing issues relative to the additivity assumption, and they are also a source of significant clinical insight in the modeling process. For example, identification of a therapeutic benefit by means of multivariable modeling is often based on the presence of an interaction between therapy given and a metric of disease burden. The presence of this interaction identifies both the presence of a survival benefit with one treatment versus another but may also identify the threshold of disease burden at which a benefit may be present. As shown in Figure 1, the amount of ischemia present identifies whether revascularization or medical therapy is associated with greater benefit and a potential threshold at which the benefit occurs. Similarly, interactions can identify unique relationships between 2 variables. For example, the presence of greater risk in female than in male diabetics but greater risk in male than in female nondiabetics is characterized by an interaction between sex and diabetes mellitus.

Similarly, the shape of a relationship, as revealed by demonstration of nonlinearity of a modeled variable, also yields significant insights. The likelihood of referral to revascularization after stress single photon emission computed tomography (SPECT) increases sharply in the lower range of ischemia but plateaus in the higher ranges, thereby permitting better understanding of physician utilization of stress SPECT results (Figure 2).

Sample size methods exist for both RCTs and observational study designs based on estimates of the number of events needed rather than total population size. Nonetheless, these power calculations may be challenging when little or no previous data exist. Overfitting (ie, fitting models with an excessive number of variables and complexity relative to the amount of data available) is ubiquitous because studies are often underpowered yet have numerous candidate variables for model entry. The minimum number of events necessary is expressed as the

Table. Assumptions Underlying Multivariable Survival Modeling

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Description</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additivity</td>
<td>The magnitude of effect of each variable does not depend on the value of other variables (x has an additive relationship with y); violation necessitates use of interactions</td>
<td>Interactions indicate that the effect of one variable depends on the level of another; failure to specify significant interactions results in compromised model fit and inferences; if valid, interactions can reveal clinically important relationships but complicate model interpretation and may represent spurious findings (especially with overfitting)</td>
</tr>
<tr>
<td>Linearity</td>
<td>Continuous variables are related to the outcome in a linear fashion; violation requires variable transformation</td>
<td>For example, when modeling death, if EF is linear, it indicates that a 20% change in EF has identical risk ramifications whether EF changes from 20% to 40% or from 50% to 70%; uncorrected nonlinearity results in compromised model fit and invalid hypothesis testing</td>
</tr>
<tr>
<td>Collinearity</td>
<td>The presence of independent variables that are highly interrelated (r&gt;0.80–85); the regression coefficient of one variable depends on which other predictors are in the model</td>
<td>Collinearity results in biased calculated regression coefficients and invalid hypothesis testing</td>
</tr>
<tr>
<td>Proportional hazards</td>
<td>CPH requires the ratio of risk in the presence and absence of a variable to remain constant throughout the observed time (proportionality of risk over time); violation requires the use of either extended CPH models or parametric survival analysis</td>
<td>Using CPH, the 2 survival curves of a binary variable should remain roughly parallel over time and not cross; violations may lead to data misinterpretation; eg, if men have a high risk early after test but women’s risk increases several years later, CPH may indicate no gender-related risk difference</td>
</tr>
</tbody>
</table>

EF indicates ejection fraction; CPH, Cox proportional hazards.

Figure 1. Relationship between percentage myocardium ischemic, treatment, and risk based on multivariable modeling. Rx indicates treatment; revasc, revascularization. Reprinted from Hachamovitch et al.14
events per variable ratio (EPV). (EPV actually refers to events per degree of freedom [rather than per variable] as the former increases with continuous variables, interactions, and nonlinearity terms.) Generally, an EPV >10:1 is strongly recommended; an EPV of 20:1 is preferred. Careful review often reveals that articles that claim to observe the 10:1 EPV may not.6

Modeling “sicker” or narrowly distributed cohorts increases error likelihoods and requires greater EPVs. Caution is recommended when overfitted models are interpreted because their reliability or calibration is questionable. When faced with excessive variables and insufficient events, data reduction techniques are available (eg, principal components analysis). These techniques can reduce complex, multidimensional data to lower dimension orders, yielding insights into the underlying structure of the data. Unlike variable elimination, these methods preserve information while permitting parsimonious modeling.7

Conversely, “underfitted” models do not consider variables of likely importance. For example, when the incremental value of imaging over pretest information is determined, it is assumed that an optimized model of pre-SPECT data will be developed. A suboptimal model is often used in that predictors of known prognostic value (eg, exercise capacity, heart rate reserve, resting ECG) are not included, resulting in overestimation of the added value of imaging.

Model Validation, Calibration, and Discrimination
After development, model performance is assessed. A model’s calibration (reliability) assesses agreement between observed and (model) predicted events across the range of predicted probabilities.7 Discrimination refers to a model’s ability to discern between patients with versus without the outcome. For example, when modeling risk, patients with events should have a high predicted probability, whereas patients without events should have a low predicted probability. Good discrimination and calibration do not necessarily coincide.7 Model validation is performed to ascertain whether the developed model will perform in populations other than those in whom it was developed. Model performance is often overestimated because models are derived and tested with the use of the same cohort.

Model Interpretation
Understanding the interpretation of multivariable models greatly enhances the amount of information these models yield. Separate statistics are generated for the overall model fit as well as for each variable in the model. Each variable also has a β coefficient. In Cox proportional hazards, the exponential (exp) yields the hazard ratio; in logistic regression, the exp yields the odds ratio. A change of >5% to 10% in the value of the β of the variable after the addition of another variable to a model indicates that the former variable is confounded by the latter.7

A model’s conclusions can be misinterpreted. A study examining prognostic implications of different tracers may find variable “tracer used” to be statistically insignificant after risk adjustment, but if each site used only 1 tracer, whether the variable “tracer used” represents the tracer used or intersite differences is unclear. Other sources of intersite variability (patient mix, referral patterns, interpretation methods) must also be considered and adjusted for.

Alternative Approaches
Although multivariable modeling has dominated statistical analyses, it is both limited and limiting, particularly in data sets with complex variables of unclear distribution and interrelationship.9 Several alternative approaches, although less commonly used, excel in classification problems such as clinical decision rules: artificial neural networks, Bayesian approaches, and decision tree–based approaches. Artificial neural networks have both advantages (ability to detect complex nonlinear relationships and all possible interactions) and disadvantages (“black box” character, greater computational burden, tendency for overfitting, empirical model development). These approaches are not universally accepted, and disagreement exists on whether they are equivalent, superior, or inferior to conventional techniques. Importantly, methodological flaws have compromised many studies utilizing artificial neural networks, suggesting that the enthusiasm surrounding this approach requires tempering.12
Assessing the Prognostic Value of Testing

Approaches to determining a test’s prognostic value have evolved significantly. Previously, demonstrations that imaging resulted in better predicted outcomes than clinical, historical, or stress testing data were adequate. The concept of incremental value—the obligation to show that a test predicted outcomes even after all other preimaging data were considered—changed this standard. This concept expanded through the 1990s, and a paradigm for validating testing emerged. This paradigm now includes the following components:

- Statistical incremental value: ascertaining the statistical added value of a test over pretest information (incremental value based on an increase in C index or $\chi^2$ of a model incorporating pretest data);
- Predictive incremental value: demonstrating further (enhanced) risk stratification on the basis of the test results after considering initial risk stratification by pretest data; alternatively, measurement of risk reclassification by test results beyond initial preclinical risk determination;
- Economic incremental value: the reduction in the cost of a testing strategy after the inclusion of the test of interest;
- Therapeutic incremental value: the ability to identify which patients may benefit with a specific posttest therapeutic approach.\textsuperscript{13–15}

Prognostic test validation is based on multiple criteria and is defined by a body of evidence consisting of multiple, well-powered studies verifying the presence of these various measures of clinical value in diverse patient groups.

Limitations of prognostic assessment are 3-fold: issues of the validity of the literature, issues of bias, and intrinsic limitations of risk as a clinical tool. A significant proportion of the supporting data for imaging is based on large databases from a limited number of centers whose data are predominantly from an era before the use of devices and medications currently considered standard of care (eg, stents, statins, clopidogrel), as well as the use of imaging techniques that may be outmoded. The generalizability of these data to other centers with different techniques, referral patterns, and expertise, using newer therapeutic approaches, is unclear. Importantly, the performance characteristics of imaging as performed in most private and community settings (the majority of studies performed) are undefined, suggesting the need for large-scale registries to assess the value of testing in practice.

Prognostic Posttest Referral Biases

Imaging results affect patient management, especially referral to revascularization.\textsuperscript{1} Because revascularization also affects risk, the association between test results and revascularization referral introduces a bias that lowers observed patients’ risk in proportion to their imaging results. To prevent underestimating risk, studies remove (censor) from prognostic analyses patients revascularized shortly after testing.\textsuperscript{2} Patients with revascularizations after this threshold are included in analyses (the revascularization likely results from worsening clinical status). Investigations of newer modalities (eg, CTA) must consider the impact of test results and patient management on outcomes and differentiate physician-driven outcomes (eg, revascularization) from risk-related outcomes (death).\textsuperscript{8}

Although censoring early revascularizations was well intended, selective removal of patients with greater test abnormalities results in relative underestimation of risk and flattening of the test abnormality–risk relationship\textsuperscript{2} in proportion to revascularization referral rates (treatment selection bias). Thus, for example, a greater observed risk reduction exists in patients with versus those without ischemia and in those with versus those without angina.

Interestingly, of 2 data elements reported by a test (eg, ischemia and ejection fraction), if one (ischemia) but not the other (ejection fraction) triggers revascularization referral, the prognostic value of the latter relative to the former is overestimated if censoring occurs (a differential treatment selection bias; Figure 3). Initial gated SPECT studies reported that the incremental value of ejection fraction over perfusion data was such that perfusion data were no longer predictive.\textsuperscript{16} Cardiac mortality increased from 2.2% in mild to moderately abnormal results to 3.6% in severely abnormal results. However, with reduced left ventricular ejection fraction (<45%), the latter were at lower risk than the former (5.7% versus 9.2%).\textsuperscript{16} This counterintuitive result was caused by referrers predominantly using perfusion but not left ventricular ejection fraction data for revascularization decisions, thereby reducing the predictive power of perfusion with minimal impact on the value of the ejection fraction.\textsuperscript{2,14} Consequently, analysis of medically treated patients underestimated the
value of ischemia relative to the value of the ejection fraction. More recently, a study modeling both medically treated and revascularized patients found that ejection fraction and perfusion added incrementally to each other for risk stratification. Therefore, inclusion of all patients in the analysis and modeling appears to at least partially correct this form of referral bias.

This differential treatment selection bias is probably common, occurring in any prognostic study of medically treated patients that compares variables with disparate associations with posttest revascularization; for example, even after risk adjustment, prognostic comparisons between patients with ischemic versus nonischemic abnormalities would be problematic because the former are more commonly revascularized than the latter, with a resulting impact on event rates. Similarly, comparing stress echocardiographic results (eg, left ventricular ejection fraction change) with pretest data (stress ECG response) will underestimate the incremental prognostic value of the former compared with the latter. Importantly, this bias will similarly impact analyses that include medically treated patients and exclude early revascularization patients.

**Estimating Posttest Risk**

Postimaging risk estimation is often suboptimal because of the failure to consider pretest information. For any imaging result, a range of posttest risk exists (Figure 4); for example, risk after moderate SPECT ischemia ranges from 2% to 10%, varying with patient characteristics. Similarly, after a coronary artery calcium score of zero, posttest risk varies with Framingham risk score (Figure 5). A coronary artery calcium value of zero with high Framingham risk score indicated greater risk than a low Framingham risk score with any coronary artery calcium value. Comparable risk variation after a normal SPECT occurs with underlying clinical risk, and both stress echocardiography and SPECT are associated with event rates >1% per year in higher-risk cohorts. Thus, postimaging risk is contextual, challenging the validity of a fixed risk threshold after a normal result as a benchmark in technology validation. Furthermore, meta-analyses comparing the prognostic value of these modalities must also adjust...
for cohort characteristics; for example, comparing risk after a normal result with 2 modalities with similar performance characteristics will reveal differing event rates if they are used in patients at different risks. Similarly, low risk after a normal result may occur with a mediocre test if a sufficiently low-risk cohort is tested. Alternative approaches to prognostically assess normal tests include indexing the event rates to the underlying risk of the overall population tested. A 0.5% per year event rate after a normal test has a different meaning in a cohort with a 3% per year overall risk and 60% abnormal test prevalence than a 2% per year overall event rate and 20% abnormal test prevalence. Expressing a relative risk of an abnormal versus normal imaging study may be helpful.2

Scores for Predicting Cardiovascular Risk and Potential Benefit

Postimaging risk estimation necessitates inclusion of preimaging data and therefore is challenging in practice and requires application of validated scores. A stress echocardiography score estimating cardiac mortality in medically treated patients exists,18 as does a stress SPECT score permitting estimates of risk with revascularization versus medical therapy (thereby also estimating potential patient benefit).2 Although these scores require extensive validation in a varying populations before clinical application, they may greatly enhance test generalizability and efficacy.

Risk-Based Versus Benefit-Based Testing

Although risk-based validation and application of imaging are accepted, identification of risk is no guarantee that intervention is necessary or beneficial. Whether imaging-based approaches to therapeutic decision making reduce mortality or hospitalization is an unproven hypothesis. Because new technologies are the major cause of increasing healthcare costs,19 pressure to test this hypothesis increases. A shift to benefit-based technology validation would “ensure that the use of new therapy and technology is tied to evidence of clinical benefit, resulting in a value-based health care system,”19 in turn permitting us to provide “services based on scientific knowledge to all who could benefit, and refraining from providing services to those not likely to benefit.”20

However, adoption of pay-for-performance programs, already widespread, necessitates fine-grained information linking imaging results, patient treatment, and subsequent outcomes in various populations and scenarios.21 Extensive evidence based on large databases and/or RCTs with extensive follow-up will be needed to identify when “benefit” is accrued, whether in the form of enhanced survival or improvements in quality of life or functional status. Thus, benefit-based validations are overdue, but significant supportive evidence will be needed.

The therapeutic incremental value of imaging has undergone limited evaluation. In stable post-MI patients, a RCT used stress SPECT–defined ischemia and ejection fraction as inclusion criteria in comparing invasive and conservative strategies, finding the two equivalent for ischemia suppression.3

The value of PET in patients with severe left ventricular dysfunction and suspected coronary artery disease has been examined by both a RCT and an observational study. Although patients randomized to viability evaluation with fluorodeoxyglucose PET versus conventional care had similar outcomes, revascularization was not mandated for PET-determined viability, and non-PET viability testing was permitted in the latter arm.4 In a propensity score–matched observational study, revascularization reduced mortality irrespective of PET viability and/or ischemia.19 Hence, despite previous findings in multiple small, nonrandomized studies and meta-analyses, no therapeutic benefit could be identified with the use of imaging in patients with reduced ejection fraction.

Two studies using propensity scores in observational data series investigated the relationship between stress-induced ischemia, revascularization, and survival in patients without prior coronary artery disease.13,14 These risk-adjusted studies found a survival benefit associated with revascularization only in patients with ischemia beyond a specific threshold.13 Conversely, ejection fraction was most closely associated with cardiac death but did not identify which patients benefited from revascularization.14 Thus, limited data indicate the potential for the therapeutic incremental value of imaging, and considerably more data will be needed.

Evaluation of Newer Modalities

The hurdles faced in validating “newer” modalities (cardiovascular magnetic resonance, CTA, PET) differ from those faced by the “older,” more widely validated modalities (SPECT, echocardiography). Previous validations were comparisons with clinical, historical, and exercise treadmill testing data, not other imaging techniques. Newer tests face a higher hurdle in that comparisons to other modalities (in addition to pretest data) will be required, necessitating more rigorous and complex study designs. As with all newer modalities, issues of study design and execution arise in initial studies but will likely improve with time, experience, and larger cohorts. However, valid or not, prior validations, even risk based, will be questioned because of the need to redefine tests in the context of enhancing patient outcomes. The acceptance of this new paradigm may “level the playing field” relative to the need for new validations.

Test Reliability

Reliability assessment is an underappreciated but important part of the validation process. The ability of a modality to yield consistent results with minimal interobserver and intraobserver variability is vital for both clinical and research applications and is a prerequisite for imaging to serve as a surrogate end point (eg, ischemic defect size in trials of anti-ischemic agents) or as a means to follow therapeutic success.

Methodological Challenges

Reliability refers to the ability of a test to give the same result on repeated application in an individual with a given level of disease.22 Reliability assessment utilizes several approaches. Continuous variables use a Pearson correlation coefficient, a measure of the extent that a relationship between 2 measures can be explained by a straight fitted regression line. The alternative is an intraclass correlation coefficient based on
repeated-measures ANOVA. These two are not equivalent; a Pearson correlation of 1.0 does not necessarily pass through (0,0) (strong correlation despite poor agreement), and an intraclass correlation coefficient of 1.0 passes through the origin. Hence, the latter is a more conservative measurement. With multiple observers, the intraclass correlation coefficient is advantageous because it yields a single result summarizing all readers or a pairwise comparison, whereas Pearson correlations generate only the latter.\(^\text{22}\)

Although the \(\kappa\) statistic is used to assess concordance, its use has several limitations. Various types of \(\kappa\) exist for different types of data and study designs. The \(\kappa\) statistic is influenced by the prevalence of agreement versus nonagreement and baseline rates; hence, a low \(\kappa\) can occur despite significant agreement and even though individual ratings are accurate.\(^\text{23}\) The \(\kappa\) statistic requires that 2 raters use the same categories; therefore, situations in which different categories exist require conversion to a common metric. Hence, \(\kappa\) values are seldom comparable across studies, procedures, or populations,\(^\text{23}\) and scales purported to categorize ranges of \(\kappa\) are inappropriate because of variability.

**Clinical Challenges and Implications**

Test reliability is sometimes limited and therefore affects modality selection and application. Dobutamine stress echocardiography interinstitutional agreement among 5 experienced centers revealed that \(\geq 4\) centers agreed on normal versus abnormal in only 73% of patients, with further variability related to image quality and extent of coronary artery disease.\(^\text{24}\) If simple dichotomous result characterization has this degree of variability, the reliability of defect size or severity interpretations is a concern.

Specific clinical situations are particularly susceptible to error, and therefore reliability rather than precision is important (eg, serial left ventricular ejection fraction measurements in valvular disease or perichemotherapy). Left ventricular mass and volume assessments in tracking left ventricular hypertrophy or remodeling depend on modality and observer reliability. Despite an excellent intraclass correlation coefficient, test-retest variability of an echocardiographically measured left ventricular mass had 95% confidence intervals wider than the average decrease in this measure in most antihypertensive regression studies.\(^\text{25}\) When the measurement error of a modality exceeds the thresholds for clinical recommendations, 2 alternatives exist. Automated, operator-independent software can minimize test-retest variability (with reader data checks). Left ventricular ejection fraction determined by various gated SPECT software has extremely high interobserver and intraobserver reproducibility (\(r=0.99\) to 1.00),\(^\text{26}\) far superior to human readers. For reliability-dependent applications, software, rather than human interpretation, may be preferable. However, accuracy-reliability trade-offs must be carefully considered.

Newer technology with superior resolution, such as cardiovascular magnetic resonance, is associated with dramatic reductions in interstudy reproducibility. Left ventricular assessment with the use of cardiovascular magnetic resonance reduces calculated sample sizes by 55% to 93% versus echocardiography to show changes in left ventricular dimensions and function.\(^\text{27}\) Thus, newer technologies aid in overcoming reliability issues, whether new software or modalities, emphasizing the need for reliability assessment as part of routine validation studies.

**Conclusions**

When one considers the enormous costs and impact of technology development and introduction, the cost restraints placed on the system, and the need to identify which patients may benefit from imaging, technology validation has an increasingly important role. This process requires an outcomes-based approach, preferably with an assessment of therapeutic benefit. Methodological rigor is necessary at all levels of these investigations. In light of the considerable data needed to achieve this, multicenter registries of new modalities will greatly enhance and accelerate this process and identify methods and questions for future clinical trials of imaging modalities.

**Disclosures**

The authors disclose grant support from Bracco Diagnostics, Astellas-Pharma, GE Healthcare, and Siemens Medical Solutions; material support from Vital Images; and participation in speakers’ bureaus for Astellas-Pharma and GE Healthcare. Dr Hachamovitch discloses participation in the speakers’ bureau for Lantheus Medical Imaging and consultancies (King Pharmaceuticals, Lantheus Medical Imaging). Dr Di Carli discloses participation in the speakers’ bureau for Bracco Diagnostics and advisory boards (Bracco Diagnostics, GE Healthcare, and Lantheus Medical Imaging).

**References**


KEY WORDS: epidemiology imaging prognosis statistics tests
Methods and Limitations of Assessing New Noninvasive Tests: Part II: Outcomes-Based Validation and Reliability Assessment of Noninvasive Testing
Rory Hachamovitch and Marcelo F. Di Carli

_Circulation_. 2008;117:2793-2801
doi: 10.1161/CIRCULATIONAHA.107.714006

Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2008 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/117/21/2793

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org/subscriptions/