Conceptualizing and Characterizing Accuracy in Assessments of Competence to Stand Trial

Douglas Mossman, MD

This article describes a mathematical framework for conceptualizing the accuracy of forensic experts' opinions on competence to stand trial (CST) and explains how an expert's expressed opinion about CST can be decomposed into four elements: (1) contextual requirements of the defendant (determined partly by the defendant's past actions) that lie outside the defendant's future control; (2) personal attributes of the defendant that are relevant to competence; (3) the expert's intrinsic ability to distinguish competent from incompetent defendants; and (4) the expert's wish to favor or avoid certain types of outcomes (e.g., a preference to avoid seeing an incompetent defendant stand trial for a serious charge). Because experts are imperfect and have varying levels of confidence in their opinions, one can describe the accuracy of CST assessments by using receiver operating characteristic (ROC) analysis. The article describes some types of insights one might derive from ROC analyses of CST assessments if experts, at least for research purposes, expressed opinions as graded levels of confidence. Although no satisfactory gold standard exists for establishing the truth about a defendant's competence, statistical methods developed over the past two decades may allow investigators to make inferences about the diagnostic accuracy of experts' CST assessments.

J Am Acad Psychiatry Law 36:340–51, 2008

In formulating opinions on psychologal matters such as criminal responsibility or adjudicative competence, forensic practitioners typically describe their decision-making as follows: they assimilate available psychological and medical evidence, consider that evidence in light of the relevant legal standard, and then determine whether the evaluee meets or does not meet the standard.[1–3] The intent of this process is to reach a clear opinion (one way or the other) with reasonable medical (or scientific) certainty and to be ready to articulate the rationale for the opinion. On occasion, forensic clinicians feel that they cannot formulate opinions with this level of certainty, or they offer opinions with the qualification that further information may change their conclusions. Usually, however, forensic clinicians simply believe that they can reach yes-or-no conclusions about psychologal questions and that if a colleague disagrees with an opinion, it must be because one of them (probably the colleague) is wrong. Often, researchers—along with clinicians who formulate opinions—assume that yes-or-no opinions are the only sort of judgments that forensic clinicians might make.[4]

On reflection, however, one realizes that this picture oversimplifies evaluees, forensic examinations, and the data evaluated by forensic practitioners. If one considers assessments of competence to stand trial (CST), for example, one immediately realizes that because criminal cases vary enormously in complexity and seriousness of charges, the capacity required to function as a competent defendant ranges from modest in some cases to substantial in others. More important, human beings vary greatly in the attributes and capacities that are needed to perform competently as a criminal defendant. Almost all the readers of this article would be highly competent to stand trial if they faced any criminal charge, and almost no preschooler would be competent, no matter how simple the case against him. But most actual CST evaluees are neither completely lacking in, nor perfectly endowed with, competence-related capacities and qualities. Cooperativeness, intelligence, rationality, and all the other human attributes that af-

Dr. Mossman is Director, Glenn M. Weaver Institute of Law and Psychiatry, and Volunteer Professor, Department of Psychiatry, University of Cincinnati College of Law, Cincinnati, OH. Address correspondence to: Douglas Mossman, MD, U.C. College of Law, Clifton Avenue & Calhoun Street, Cincinnati, OH 45221-0040. E-mail: douglas.mossman@uc.edu

fect adjudicative competence are dimensional rather than present-or-absent features of persons and their mental functioning. To the extent that these capacities and qualities are measurable or scalable, they represent spectra along which actual criminal defendants might be arrayed in near-continuous distributions. Forensic clinicians therefore can anticipate that a few accused persons will appear unambiguously capable or incapable of functioning as defendants. Most CST evaluees, however, will display competence-related capacities that place them somewhere between these extremes, by virtue of their comparative strengths and weaknesses along many emotional and cognitive dimensions.

This means that a forensic expert who evaluates adjudicative competence is not attempting to discover whether a defendant has or lacks something. Rather, the expert seeks to obtain information about where a defendant falls along several spectra of competence-related capacities and qualities. An expert's opinion concerning CST, therefore, reflects implicit conclusions about the defendant's positions along these spectra, coupled with the expert's understanding of what the specific case will require of the defendant.

In any given criminal case, the capabilities required of the defendant reflect factors that are beyond his future control and therefore are independent of him. Examples of such factors include what the defendant has done and cannot go back and change, what charges he actually faces, potential punishments, options, and choices about defenses, and case-specific behavioral demands (e.g., whether the defendant must testify and withstand cross-examination). In a sense, then, the case requirements are external or contextual factors fixed by the defendant's situation and predicament, and the expert's task is to evaluate internal or personal factors about the defendant to find out whether he can meet those requirements. Although experts may not think of the CST evaluation process in this way, the effort to arrive at a yes-or-no (i.e., competent or incompetent) conclusion is actually an effort to discern where a defendant stands in relation to the external, contextual demands placed on him and whether he falls on one or the other side of an implicit boundary between competence and incompetence. This means that, to the extent that forensic clinicians acknowledge that their opinions may be wrong, they may (or should) feel the least confident in cases where the data locate a defendant close to the competence-incompetence boundary.

In a recent article,[5] Buchanan notes that psychiatrists actually do make errors and have varying degrees of confidence in their opinions about CST. Because of this, he suggests that in offering opinions on CST, forensic clinicians should factor in three considerations: the defendant's mental functioning (presumably along all dimensions relevant to adjudicative competence), the contextual requirements of the specific case (including, for example, the amount and complexity of the information that the defendant must process), and the potential penalties. Concerning the last item, Buchanan posits an analogy between CST determinations and decisions concerning competence to consent to treatment, where a sliding scale is sometimes invoked to require a higher level of competence when a patient's decision could have grave consequences.[6,7] Although forensic clinicians cannot actually weigh risks and benefits on some sort of scale, Buchanan believes that experts can and should "take into account the seriousness of the charges"; greater potential penalties should incline experts to require "a greater level of confidence [in their opinions] before suggesting that a defendant is competent" (Ref. 5, p 463). The more serious the consequences of a conviction, the more forensic clinicians should err on the side of caution—that is, they should favor erroneous opinions that deem actually competent defendants incompetent over erroneous opinions that would allow actually incompetent defendants to proceed to trial. In other words, says Buchanan, "the implications of preferring some types of error to others" mean that "all other things being equal, the seriousness of the charge that a criminal defendant faces should affect the evidence a psychiatrist gives and the conclusion a court reaches" (Ref. 5, p 458).

Implicit in Buchanan's position is the idea that experts are imperfectly accurate and that, in evaluating adjudicative competence, they face inevitable tradeoffs between wrongly categorizing competent defendants as incompetent and incompetent defendants as competent. Thus, when experts offer their customary, binary, yes-or-no opinions about adjudicative competence, those opinions likely combine experts' awareness of the potential for errors with experts' feelings about how those errors should be balanced. For example, an expert may believe that evidence concerning a particular psychotic defendant facing a serious charge indicates, on balance, that the defendant might manage to stand trial successfully. Yet to avoid seeing an incompetent defendant face prosecution and punishment, the expert might state—in Buchanan's view, with moral justification—that the defendant lacked capacity to proceed with adjudication. Offering this opinion might be justified if the expert believed that treatment would cause the defendant's psychosis to abate and his fitness for trial to improve substantially.

This article agrees and begins with Buchanan's explicit position that forensic experts both have varying levels of confidence in their opinions and that experts make errors. But this article suggests that for scientific purposes, mental health experts should think about confidence in their opinions' accuracy as a matter distinct from the implications of potential errors in those opinions and distinct from their choices to favor one or another type of error based on the consequences. To the extent that forensic experts want to understand how well they perform, they need a framework for characterizing and quantifying the accuracy of their opinions that considers accuracy separate from their beliefs about the relative costs and benefits associated with incorrect and correct opinions.

Receiver operating characteristic (ROC) analysis is medicine's preferred method for describing diagnostic accuracy as a set of tradeoffs and for separating diagnostic accuracy from clinical decisions based on balancing of costs and benefits.[8,9] Akinkunmi[10] has shown how to use ROC methods to describe the accuracy of assessment tools, using senior psychiatrists' opinions as the gold standard for the truth about an evaluee's competence. This article explains how ROC analysis can be used to conceptualize and gauge the accuracy of experts themselves—the accuracy, that is, of experts' "diagnoses" concerning criminal defendants' adjudicative competence.

Confidence About Adjudicative Competence

Any valid characterization of diagnostic accuracy must rest on a clear understanding of both the condition being diagnosed and the diagnostic process. It lies far beyond the scope of this article to provide a thorough account of what competence to stand trial is or how evaluators assess it, but several sources (e.g., Refs. 11–13) provide detailed, thoughtful discussions of these topics. For present purposes, however, some summary comments will facilitate understanding of the quantitative perspective presented later.

In *Dusky v. United States*,[14] the U.S. Supreme Court articulated what has become the basic constitutional standard against which U.S. state and federal courts assess defendants' adjudicative competence. Under *Dusky*, the test for CST is whether a defendant "has sufficient present ability to consult with his lawyer with a reasonable degree of rational understanding—and whether he has a rational as well as factual understanding of the proceedings against him" (Ref. 14, p 402). In a subsequent case, *Drope v. Missouri*, the U.S. Supreme Court added that, beyond consulting with counsel, a criminal defendant must be able "to assist in preparing his defense" (Ref. 15, p 171).

For more than four decades, appellate courts, legislatures, legal scholars, and forensic clinicians have undertaken various efforts to elaborate and flesh out these sparsely worded requirements (e.g., Refs. 16–24). Often, these efforts take the form of lists intended to focus forensic evaluators' attention on the functional capacities that are directly relevant to adjudicative competence. The Utah Code[25] contains one of the most detailed lists developed by state legislatures; the relevant portion appears in Table 1.

The Utah statute points to mental capacities that forensic examiners in any U.S. jurisdiction would find relevant to adjudicative competence. These mental capacities reflect several dimensions of sophisticated social and interpersonal functioning, a partial description of which include: (1) capacities for comprehending social facts, (2) ability to project oneself into hypothetical situations (e.g., being convicted), (3) anticipating one's response to possible events, (4) establishing trusting relationships and recognizing whom to trust, (5) recognizing what things are relevant in a complex social situation, (6) communicating those things logically to other persons (e.g., to one's attorney), (7) understanding the role of social institutions, (8) entertaining and evaluating one's own beliefs and desires, (9) reasoning practically in light of rational beliefs and desires, (10) maintaining self-control in emotional situations, and (11) responding to interpersonal events and expressing emotion appropriately. This long but far from complete list serves as a reminder that adjudicative competence is an abstract, "open-textured" construct.[26] That is, CST is a "postulated attribute" (Ref. 27, p 283) of a criminal defendant; because it is a construct that is intended "to apply to an infinite

Table 1 Instructions to Experts Who Perform Court-Ordered Evaluations of Competence to Stand Trial (Utah Code 77-15-5(4))[25]

The experts shall consider . . . and address, in addition to any other factors determined to be relevant . . . :

(a) the defendant's present capacity to:

(i) comprehend and appreciate the charges or allegations against him;

(ii) disclose to counsel pertinent facts, events, and states of mind;

(iii) comprehend and appreciate the range and nature of possible penalties, if applicable, that may be imposed in the proceedings against him;

(iv) engage in reasoned choice of legal strategies and options;

(v) understand the adversary nature of the proceedings against him;

(vi) manifest appropriate courtroom behavior; and

(vii) testify relevantly, if applicable;

(b) the impact of the mental disorder, or mental retardation, if any, on the nature and quality of the defendant's relationship with counsel;

(c) if psychoactive medication is currently being administered:

(i) whether the medication is necessary to maintain the defendant's competency; and

(ii) the effect of the medication, if any, on the defendant's demeanor and affect and ability to participate in the proceedings.

number of fact situations," we cannot reduce its meaning "to an invariable group of rules about a set of facts" (Ref. 28, p 323).

Adjudicative competence is not an entity like blood pressure or hematocrit, which are things that one can measure for purposes of diagnosis and that reflect physical states of organisms or biological systems. We cannot fully define most qualities that contribute to adjudicative competence, because the matters that CST evaluators consider require varying descriptions and can take on a kaleidoscopic array of forms. At the same time, a condition of the possibility that evaluators can perform CST assessments with greater or lesser accuracy is the idea that those assessments seek to ascertain some real characteristic of defendants. Discourse about the "accuracy" of CST assessments presupposes that adjudicative competence refers to some actual (though hypostatized[29]) feature of defendants and is not an arbitrary legal status or social construct. Adjudicative competence, in this view, is an objective (though imperfectly defined and apprehended) property of individuals who face prosecution. In categorizing defendants as competent or not, then, courts are making decisions based on inferences about the degree to which defendants exhibit or possess this real quality. And when evaluators assess adjudicative competence, they are discovering information concerning the presence or absence of this real quality.

Coupled with considerations of defendants' various human capacities are considerations about the demands of the case. As was noted earlier, CST determinations are contextual: the examiner's focus is on whether the defendant, facing one or more specific charges, confronting specific alleged facts, and represented by a particular attorney, can understand the specific proceedings against him in his current criminal case and assist the attorney in preparing a defense.[30] A defendant who anticipates being tried for alleged income tax violations probably can anticipate courtroom demands far exceeding those required of a defendant who expects to plead guilty to misdemeanor assault. One might say that the former defendant faces legal, behavioral, and intellectual hurdles far higher than the latter. This metaphor is imperfect, however, because the demands faced by criminal defendants, though perhaps amenable to some rough ranking of difficulty, are not measurable (as are the heights of hurdles).

This means that during evaluations of adjudicative competence, a forensic examiner must mentally assess many of a defendant's intangible and hard-to-quantify personal qualities against the equally hard-to-quantify challenges inherent in the defendant's specific situation. Put this way, the task of assimilating data and formulating an opinion about CST sounds impossible.

Fortunately for courts and forensic experts, however, this is not the case, for two reasons. First, individual defendants are justifiably presumed competent to stand trial, because the average defendant is competent to stand trial,[28] most defendants are competent to stand trial, and almost any normal adult who does not have serious psychopathology or cognitive impairment would be competent to stand trial if charged with a criminal offense. Therefore, a major part of an examiner's assessment consists in determining whether a defendant-evaluee has a mental illness or disability that would put him at a substantial disadvantage compared with most criminal defendants.

Second, the training and experience of mental health professionals often allows them to develop a decent idea about whether a task lies within a defendant's ability, even when they cannot "measure" the ability upon which the defendant relies. For example, speaking coherently is an attribute important to adjudicative competence, and though verbal coherence is hard to measure numerically, its relative quality and adequacy are easily apprehended. Therefore, when a forensic examination detects signs of a serious psychiatric disorder or cognitive impairment, the examiner usually does not need to quantify the extent of the problem. The question, rather, is whether and how the deviation from mental normality prevents the defendant from doing what most individuals could do.

It is helpful to think about the duties of CST evaluators through an analogy: asking a sightless person attending a football game to say whether the ball is closer to the home team's or the visitors' end zone, based on hearing the quarterback call signals before the ball is snapped. Though perhaps unable to state the precise line of scrimmage, the sightless person could always give level-of-confidence statements based on his perceptions. In many instances, for example, he may say that he was "highly confident" that the ball was nearer either the home team's or the visitor's goal line, in other instances, that the ball was "probably" on one or the other side of the field, and in a few instances, that he was "unsure" about which side of the field contained the line of scrimmage.

Such level-of-confidence responses are similar to the rating categories that radiologists have used for decades in studies of diagnostic accuracy. In mammography, for example, the reader's task is to sort those films that represent cancer from those that represent benign findings. For purposes of assessing diagnostic accuracy, however, radiologists often rate mammograms on a five-category scale (1, normal; 2, benign finding; 3, probably benign finding; 4, suspicious abnormality; or 5, highly suggestive of malignancy) that signifies their beliefs about the likelihood that the imaged tissue contains a cancer.[31] As is true of forensic examiners, mammographers' confidence in their opinions is not a precisely measurable quantity, but this does not prevent their ratings from being useful. Both for clinical purposes and for quantifying of accuracy, it is sufficient that a mammographer's confidence levels embody consistent, meaningful rankings of the probability of malignancy.

In our hypothetical football stadium example, the "truth" about the football's location can be established by (for example) asking the referee. In mammography studies, true disease status is established through biopsy (if a lesion is observed) or the presence or absence of detection of a tumor (clinically or via a subsequent mammogram) within a year of the original diagnostic study.[32,33] In forensic contexts, however, establishing the truth is a problematic matter. We have no way to biopsy a defendant's competence to stand trial and obtain a tissue diagnosis, and beyond examiners' opinions, no gold standard criterion exists to establish a defendant's adjudicative competence.[34,35] For any given defendant, all we can hope to know are various individuals' opinions about the matter (accompanied in some instances by scores from assessment instruments). Although courts ultimately determine whether defendants proceed with criminal adjudication, judges' rulings on trial competence are not perfect; nor, of course, are the opinions of even the best forensic examiner.

We shall return later to the absent gold standard problem. For the moment, let us suppose that we have a way of establishing the truth about defendants' adjudicative competence. How might we characterize the accuracy of competence assessments so as to reflect levels of confidence in experts' determinations?

ROC Analysis of CST Data

The designers of the MacArthur Competence Assessment Tool–Criminal Adjudication (MacCAT-CAJ)[36] addressed the truth problem as follows. They presumed that 486 jail inmates (197 randomly selected, unscreened pretrial defendants, plus 249 inmates who were receiving psychiatric treatment unrelated to competence restoration) were actually competent to stand trial; they also presumed that 283 individuals who had been adjudicated incompetent and hospitalized less than 14 days for competence restoration were actually incompetent to stand trial (IST). The MacCAT-CA instructions state explicitly that evaluators should not regard or use the instrument as though it were an objective, diagnostic test for adjudicative competence. But for illustration purposes, let us treat the three MacCAT-CA scales, Understanding, Reasoning, and Appreciation, as though they were diagnostic tests. Doing so will help

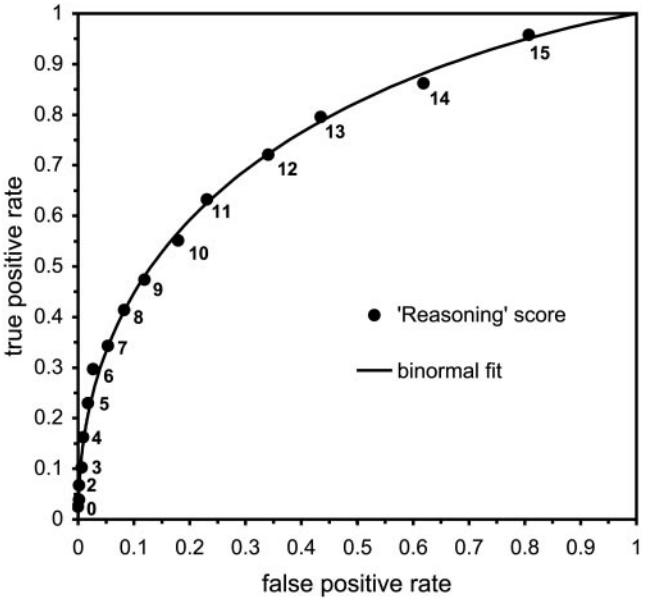

Figure 1. A ROC graph, based on data in Poythress *et al.*,[36] that treats the MacCAT-CA Reasoning scale as a diagnostic test for incompetence to stand trial. Numbered markers represent (FPR, TPR) pairs associated with scores on the Reasoning scale. A smooth, binormal ROC is fitted to the data.

us to understand how CST assessments involve tradeoffs between sensitivity and specificity and how the accuracy of assessments can be quantified. We shall also be able to sample the kinds of insights one might gain from thinking about CST assessments quantitatively.

Figure 1 shows a ROC graph (based on data in Table 5 of Ref. 36) that treats the MacCAT-CA Reasoning scale as though it were a diagnostic test for being IST (so that a positive test result indicates incompetence). As is customary in ROC graphs, the true-positive rate (TPR, equal to test sensitivity) is plotted along the vertical axis as a function of the false-positive rate (FPR, equal to 1 − test specificity), which is plotted on the horizontal axis. The numbered markers represent (FPR, TPR) pairs associated with scores on the Reasoning scale if each score were used as a cutoff to classify defendants as either incompetent or competent. On the Reasoning scale, defendant-evaluees can score between 0 and 16; the lower the score, the stronger the evidence of incompetence. Fitted to the markers is a smooth ROC curve based on the "binormal assumption," which means that on some monotonic transformation of the decision axis (here, the MacCAT-CA Reasoning score), the distribution of results will conform to two normal distributions with different means and variances. Binormal ROC fitting methods are robust to a variety of plausible data distributions.[37] They allow us to summarize a diagnostic system's performance throughout its entire range of outcomes using two indices, A and B, related as follows: $Z_{\mathrm{TPR}} = A + BZ_{\mathrm{FPR}}$, where Z_{FPR} and Z_{TPR} are the normal deviates, or Z-transforms, of FPR and TPR. In addition, the binormal fit offers the advantage of avoiding spurious inferences based on random data scatter, which for some of the interpretations offered later in the article is a valuable attribute. (For more detailed explanations of the binormal assumption in ROC curve fitting, see Refs. 38–40.)

As Figure 1 shows, the lower left portion of the ROC square contains (FPR, TPR) pairs that correspond to lower MacCAT-CA Reasoning scale scores, and the upper right portion contains (FPR, TPR) combinations associated with higher scores. If one used a lower Reasoning score as a diagnostic cutoff, one would expect to make fewer false-positive errors (that is, specificity is high because one makes relatively few incorrect declarations that an individual is IST). Because TPR would also be low, however, the scale would be insensitive and would fail to detect most IST defendants. Using higher Reasoning scores as cutoffs, one would classify a larger fraction of the IST defendants correctly (the test would be more sensitive); specificity would decrease, however, and one would misclassify more actually competent defendants as incompetent.

The area under the ROC curve (AUC) is a commonly used summary statistic of a diagnostic test's accuracy[41,42] which, in this context, would have this practical meaning: AUC equals the probability that a randomly chosen incompetent defendant would score lower on a MacCAT-CA scale than a randomly chosen competent defendant. Looking at Figure 1, one might guess that about three-fourths of the ROC square lies below the ROC curve for the Reasoning scale, and in fact the calculated area, $\mathrm{AUC}_{\mathrm{Reasoning}}$, is 0.763. This value is comparable to the AUCs reported by Akinkunmi for a smaller group of London pretrial detainees[10] and to AUCs for actuarial risk assessment instruments (ARAIs) (see, for example, Refs. 43, 44). The result suggests that the Reasoning scale ranks competent and incompetent defendant-evaluees as proficiently as ARAIs rank individuals' likelihood of future violence.

Figure 2 shows ROC curves for all three MacCAT-CA scales. A quick look at the curves suggests

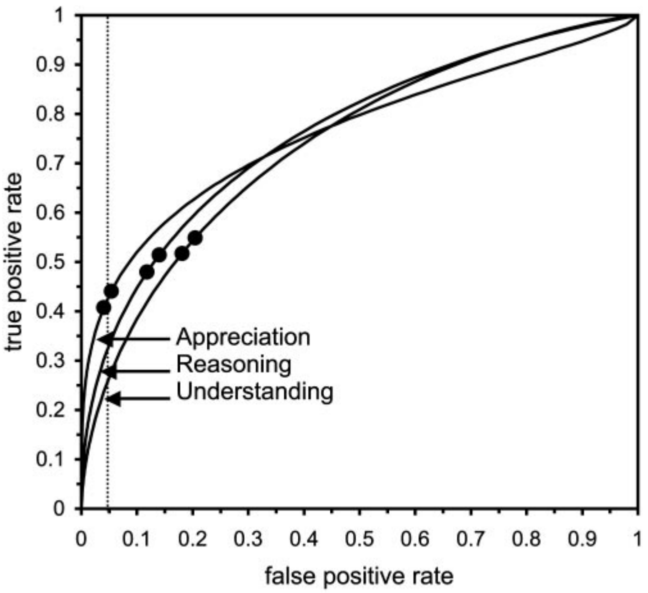

Figure 2. ROC curves for all three MacCAT-CA scales, based on data in Poythress *et al.*[36] The dotted line (at FPR = 0.05) suggests that when detecting incompetence with high specificity, the Appreciation scale has the highest sensitivity. The circular markers represent points that maximize diagnostic information when the base rate of incompetence is 0.16 (the bottom left circle on each curve) and 0.30 (top right circle on each curve).

that the Understanding and Appreciation scales have overall accuracies that are globally similar to those of the Reasoning scale, and the scales' AUCs bear this out ($\text{AUC}_{\text{Understanding}} = 0.741$; $\text{AUC}_{\text{Appreciation}} = 0.762$). But in the ROC square, the curves for each scale cross each other, and, looking more closely, one sees that the performances of the three scales may be quite different over certain score ranges.

Suppose, for example, that we want to investigate the performance of each scale when the specificity is high (or equivalently, when FPR is low). In the lower left corner of the ROC square, we see that the curve for the Appreciation scale lies above the curve for the Reasoning scale, which in turn lies above the curve for the Understanding scale. This means that, at any given value of FPR, the TPR will be highest for Appreciation and lowest for Understanding. For example, when FPR = 0.05 (see Fig. 2, dotted line), TPR for Understanding, Reasoning, and Appreciation are 0.266, 0.332, and 0.431, respectively. In a situation where one wishes to make judgments about adjudicative incompetence with high specificity, the MacCAT-CA's Appreciation scale may be the most helpful, because it has the highest sensitivity.

Setting FPR at 0.05 effectuates an arbitrary cutoff choice, however. A sophisticated way to select a non-arbitrary cutoff would be to find the operating points on each scale that provide the most diagnostic information. It turns out that each point in the ROC square is associated with a specific amount of diagnostic information.[45] To find the cutoff along a binormal ROC curve that yields maximum information, one needs only to know *A* and *B*, which are the two indices that define the binormal ROC curve's shape, and the prevalence or base rate (BR) of the condition to be detected.[46–48] Various reports[49–52] suggest that 16 to 30 percent of criminal defendants referred for CST evaluations are incompetent. The circular markers in Figure 2 represent the points along the ROC curves for each MacCAT-CA scale that maximize diagnostic information when BR = 0.16 (the lower left circle on each curve) and when BR = 0.30 (higher right circle on each curve).

Figure 2 indicates that information is maximized at values of FPR and TPR that differ for the three scales. Table 2 makes this point numerically: it lists the (FPR, TPR) pairs that maximize each scale's information, the amounts of information yielded by each scale (both in bits and as a percentage of the information that would be produced by a perfect test), the corresponding MacCAT-CA scale scores, and the MacCAT-CA normative interpretations of

Table 2 (FPR, TPR) Pairs that Maximize Each MacCAT-CA Scale's Information, Based on Data in Poythress et al.[36]

				Information			
MacCAT-CA Scale	BR	FPR	TPR	Bits	%max	Scale Score	Interpretation (Impairment Level)
Understanding	0.16	0.181	0.517	0.053	8.3	9	Mild
	0.30	0.204	0.548	0.081	9.2	10	Minimal
Reasoning	0.16	0.118	0.479	0.071	11.2	9	Mild
	0.30	0.140	0.514	0.106	12.0	9	Mild
Appreciation	0.16	0.040	0.407	0.105	16.5	7	Clinically significant
	0.30	0.054	0.441	0.148	16.8	8	Clinically significant

Also shown are the amounts of information yielded by each scale (both in bits and as a percentage of the information that would be obtained from a perfect test), corresponding scale scores, and the MacCAT-CA interpretations of these scores. BR = assumed base rate.

these scores. Table 2 suggests that if one seeks to optimize diagnostic information, then a clinically significant impairment in Appreciation provides the most information about a defendant-evaluee's adjudicative competence.

Problems With Ascertaining Accuracy and the Truth

The previous section provides examples of the type of insights we might derive from using ROC methods to characterize the accuracy of CST assessments. However, in the previous section, an evaluation tool was (mis)treated as a diagnostic instrument. What forensic clinicians and courts would really like to know is how accurate real assessments of CST are—that is, how accurately evaluators combine interview information, scores from assessment instruments, and other data about defendant-evaluees to form opinions concerning adjudicative competence.

Any effort to respond to this need would encounter two problems. First, if Buchanan is correct in his view that experts should take into account the seriousness of the charges that defendants face and should factor in the costs of diagnostic errors when formulating opinions about competence, then the actual, competent or incompetent opinions that evaluators provide to courts incorporate two kind of judgments: those about defendants' abilities, and those about the relative desirability of certain types of errors.

To understand this point, imagine that a CST examiner emulates the practice (described earlier) used by radiologists in diagnostic studies, so that, at the conclusion of each CST evaluation, the examiner assigns each defendant-evaluee to one of five categories: 1, very likely competent; 2, probably competent; 3, uncertain; 4, probably incompetent; and 5, very likely incompetent. Suppose now that the examiner considers the appropriate binary (competent or incompetent) opinion for defendants who face misdemeanor charges for which the likely consequences of a guilty plea would be time served. The examiner may feel that the legal presumption of competence, combined with the minimal adverse consequences of proceeding with adjudication in such cases despite being incompetent, favors conclusions that defendants in categories 1, 2, and 3 (and maybe even category 4) are competent. By contrast, if defendants face major felony charges or complex cases that will place high demands on their mental functioning, the same examiner may feel comfortable stating that defendants are competent only if they fall into category 1 (or perhaps, categories 1 and 2). Thus, using the expert's yes-or-no opinions about CST as the basis for quantifying accuracy would provide a muddied picture of the expert's actual powers of discernment.

The second problem in quantifying accuracy is our ultimate inability to ascertain the truth about a defendant's adjudicative competence. Several studies have shown that at least 80 percent of the time, pairs of evaluators agree about evaluees' competence.[28,49,53–56] This finding sounds impressive until one recalls that approximately 80 percent of defendants referred for CST evaluation are competent, and that someone who had no information and simply said all defendant-evaluees were competent would therefore agree with an examiner around 80 percent of the time.[56] Even more important, high reliability values do not mean that judgments are accurate: two evaluators using the same simple but foolish judgment rule (e.g., defendants are competent if and only if their last names start with vowels) would agree perfectly while being quite inaccurate.

How, then, could one establish the truth about a defendant's competence? As was noted earlier, there is no biopsy or other gold standard for establishing a diagnosis of adjudicative incompetence. The MacArthur studies[37,57] addressed this problem by using subgroups for whom the truth seemed reasonably clear (for example, forensic inpatients recently found incompetent to stand trial, and defendants who appear unquestionably competent to the investigators; Ref. 57, p 187). Of course, this approach depends on opinions of imperfect human evaluators; more important, it may yield a subject pool that does not represent the full spectrum of individuals who undergo CST evaluations. Specifically, this approach may exclude those pretrial defendants who appear merely probably competent or who seem possibly incompetent, the very defendants whose evaluations generate ambiguous results and who make one wonder how accurate evaluators are.

One might treat criminal courts' findings about each defendant's competence as correct because ultimately, these findings are the only ones that count. But usually, courts simply accept mental health experts' judgments about competence without any review of the evaluator's work.[55,58–60] When courts do hold contested hearings on adjudicative competence, judges learn information indirectly (from testimony

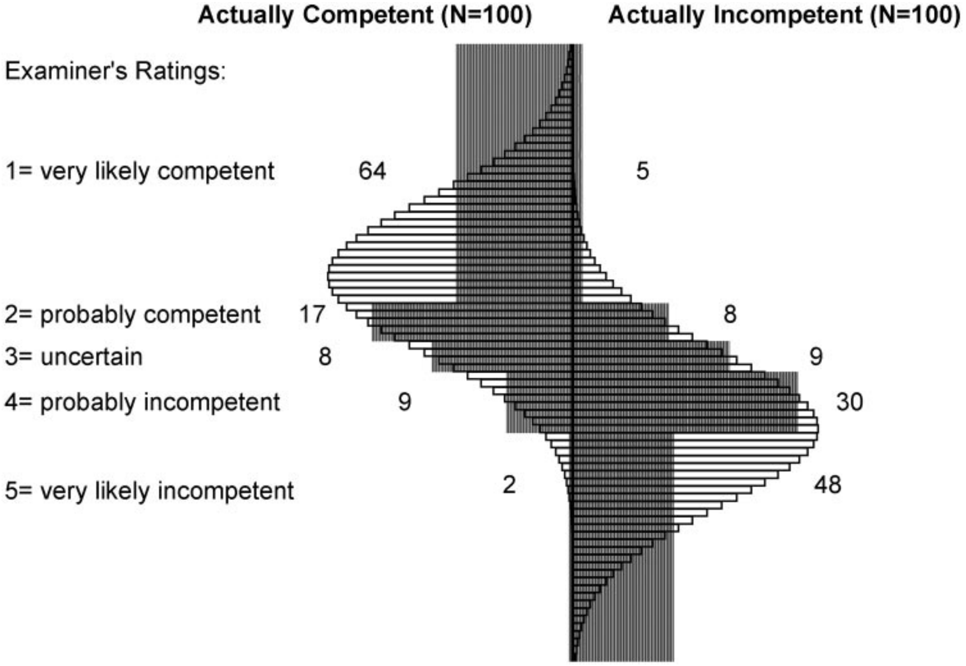

Figure 3. Hypothetical study results from a five-category rating of 200 defendants, half of whom are actually competent to stand trial. Gaussian distributions fit the rating data to the binormal assumption.

and written opinions) and do not conduct their own independent examinations of defendants. But even if judges interviewed defendants, questioned experts, and formed their own opinions in every case, the accuracy of courts' decisions would be limited by judges' human imperfections and inaccuracies. (As the existence of successful appeals demonstrates, judges disagree about various matters, and when they do, at least one judge must be in error.) Conceivably, one could conduct an experiment in which all defendants proceeded with adjudication (perhaps in a provisional trial[54]) irrespective of experts' opinions, and their performance was directly assessed in vitro, rather than inferred in advance. Even if this were practical, evaluators might not agree in some (and perhaps a substantial fraction of) cases as to whether defendants met competence criteria. This problem persists even when the judgment of a blue-ribbon panel of experts renders a decision,[28] because in cases of disagreement, the outvoted experts think they are right, and they well could be, assuming, as this article does, that CST is a real quality about which the majority could be mistaken. Of course, all uncertainty would be eliminated by having an Omniscient Being provide the correct judgment in each case, but unless the research team includes a universally recognized prophet who can transmit that judgment, investigators will remain unsure of the truth, or at least unsure that they can convince their colleagues.

A Hypothetical Study

The previous paragraphs help us understand why, to date, no study has reported on the accuracy of competence assessments. For the moment, however, let us suppose that a universally recognized prophet joined a research team and supplied the truth about each defendant's competence. Let us also suppose that examiners offered opinions about evaluees in the form of graded judgments about competence, rather than restricting themselves (as statutes usually require them to do) to providing binary, yes-or-no opinions. How, then, might we depict an examiner's accuracy?

Figures 3 and 4 show two formats for portraying results in this hypothetical study. Looking at Figure 3, we see that an examiner has evaluated 200 defendants, half of whom are actually competent and half of whom are actually incompetent to stand trial. The examiner assigned ratings of "5" to 48 actually incompetent defendants and to just 2 actually competent defendants. If the examiner were to use this category as the only one for which a "not competent" opinion was offered (recall the previous discussion of

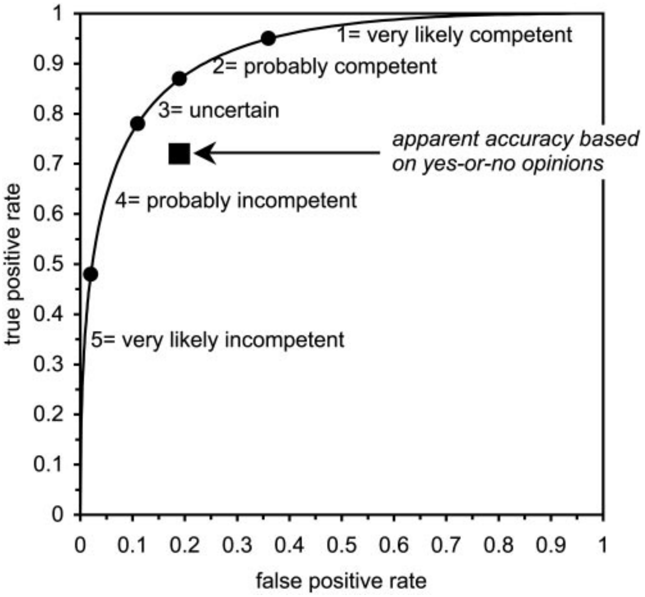

Figure 4. ROC graph based on hypothetical data in Figure 3. Segments of ROC curve areas labeled to indicate their sources in one of five rating categories. Four potential decision thresholds appear along the curve as circles that are fitted to a smooth curve according to the binormal assumption. The square labeled apparent accuracy based on yes-or-no opinions assumes that half the defendants rated 2, 3, and 4 are deemed competent and that these opinions are used to quantify accuracy.

misdemeanor defendants who face minimal adverse legal consequences), then FPR = 0.02 and TPR = 0.48. The examiner assigned ratings of "4" to 30 actually incompetent defendants and 9 actually competent defendants; if defendants in both categories 4 and 5 were included in those for whom a not-competent opinion was offered, then FPR = 0.11 and TPR = 0.78. (A comment: at incompetence base rates of 0.16 to 0.30, this decision threshold maximizes diagnostic information.) We can make similar calculations for the two other possible policies, which would produce binary opinions that favor avoiding having incompetent defendants stand trial. The Gaussian distributions in Figure 3 show the fit of the examiner's rating data under the binormal assumption. The notion is that because the number of rating categories is somewhat arbitrary, the binormal distributions lead us to a better appreciation of the examiner's underlying ability to distinguish competent and incompetent defendants.

Figure 4 depicts the hypothetical examiner's performance as a ROC graph. Each portion of the ROC curve is labeled to indicate its source in, or correspondence to, one of the five rating categories shown in Figure 3. The four potential decision thresholds appear along the curve as circles at these (FPR, TPR) pairs: (0.02, 0.48), (0.11, 0.78), (0.19, 0.87), and (0.36, 0.95). The smooth curve is fitted to these points based on the binormal assumption. (In Figure 4, AUC = 0.921, a larger value than is the case for the AUCs shown in Figs. 1 and 2. This result is consistent with the assumption that an examiner would be more accurate than a rating tool alone.)

Figure 4 also contains a square located at (0.19, 0.72) labeled apparent accuracy based on yes-or-no opinions. The square's location assumes that when expressing opinions as yes-or-no judgments about adjudicative competence, the hypothetical examiner considered consequences of errors for the 81 defendants rated 2, 3, and 4, and deemed half of these defendants competent and half incompetent. The examiner also said that all defendants rated 1 were competent, and all defendants rated 5 were incompetent. Notice that the square lies *below* the ROC curve fitted to the examiner's rating data. This illustrates how judging accuracy from examiners' binary opinions about CST, which incorporate their beliefs about relative costs of errors, may underestimate the examiners' actual ability to distinguish competent from incompetent defendants.

Conclusions

This article has conceptualized an expert's expressed opinion about a given defendant's CST as reflecting four separable factors: (1) the contextual demands on the defendant, determined by external factors beyond the defendant's future control (e.g., what he actually did, the complexity of the case, the charges filed, the evidence against him, and potential punishments); (2) the personal attributes of the defendant that are relevant to competence (e.g., intelligence, mental organization, social awareness, and cooperativeness); (3) the expert's accuracy, which depends on the expert's ability to obtain and assimilate information about (1) and (2) and to use that information to distinguish competent from incompetent defendants; and (4) the expert's decision to favor one sort of error over another based on perceptions of the consequences of those errors.

This article also adopts Buchanan's idea[5] that experts are imperfect and have varying levels of confidence in their opinions and suggests that because of this variability, we should conceive of experts' determinations as having a certain level of accuracy that might be characterized by ROC analytic methods.

The article has described some types of insights one might derive from ROC analyses of CST assessments if experts (at least for research purposes) were to express opinions about competence as graded levels of confidence instead of restricting themselves to binary (competent or not) options. For example, it is possible to quantify accuracy either as the area under the ROC curve or by using the cutoff for maximum information and to make judgments about relative merits and usefulness of measures by considering such data.

Most medical and scientific studies that attempt to quantify the accuracy of a detection or diagnostic technique use some recognized gold standard that establishes the truth about the phenomenon of interest. This article has explained why, where studies of accuracy in CST determinations are concerned, no satisfactory gold standard exists, especially if such studies are intended to evaluate experts themselves and to include tough or ambiguous cases, the very sorts of cases that make mental health experts aware that their assessments are not perfect.

Fortunately, however, statistical research over the past two decades has yielded methods that allow for inferences about diagnostic accuracy in the absence of a gold standard diagnosis.[61–64] Though description and examination of these statistical methods lie beyond the scope of this article, the author hopes that the preceding discussion will encourage those who investigate psycholegal topics to collect and evaluate data about CST determinations that are amenable to such methods. Such efforts would allow researchers and forensic experts to know and quantify, perhaps with considerable precision, the accuracy of assessments of adjudicative competence.

References

1. Resnick PJ, Noffsinger SG: Competence to stand trial and the insanity defense, in Textbook of Forensic Psychiatry: the Clinicians Guide to Assessment. Edited by Simon RL, Gold LH. Arlington, VA: American Psychiatric Publishing, 2004, pp 329–47.
2. Giorgi-Guarnieri D, Janofsky J, Keram E, *et al*: AAPL practice guideline for forensic psychiatric evaluation of defendants raising the insanity defense. J Am Acad Psychiatry Law 30(2 suppl):S3–40, 2002
3. Grisso T: Competence to stand trial. Evaluating Competencies: Forensic Assessments and Instruments (ed 2). New York: Kluwer Academic/Plenum Publishers, 2003, pp 69–148
4. Morris GH, Haroun AM, Naimark D: Assessing competency competently: toward a rational standard for competency-to-stand-trial assessments. J Am Acad Psychiatry Law 32:231–45, 2004
5. Buchanan A: Competency to stand trial and the seriousness of the charge. J Am Acad Psychiatry Law 34:458–65, 2006
6. Drane JF: Competency to give informed consent: a model for making clinical assessments. JAMA 252:925–7, 1984
7. Drane J: The many faces of competency. Hastings Cent Rep 15:17–21, 1985
8. Zweig MH, Campbell G: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–77, 1993
9. Obuchowski NA: Receiver operating characteristic curves and their use in radiology. Radiology 229:3–8, 2003
10. Akinkunmi AA: The MacArthur Competence Assessment Tool-Fitness to Plead: a preliminary evaluation of a research instrument for assessing fitness to plead in England and Wales. J Am Acad Psychiatry Law 30:476–82, 2002
11. Melton GB, Petrila J, Poythress NG, *et al*: Psychological Evaluations for the Courts: A Handbook for Mental Health Professionals and Lawyers (ed 3). New York: Guilford Press, 2007, pp 125–64
12. Miller RD: Criminal competence, in Principles and Practice of Forensic Psychiatry (ed 2). Edited by Rosner R. London: Arnold, 2002, pp 186–212
13. Appelbaum PS, Gutheil TG: Clinical Handbook of Psychiatry and the Law (ed 4). Philadelphia: Lippincott Williams & Wilkins, 2007, pp 225–59
14. Dusky v. United States, 362 U.S. 402 (1960)
15. Drope v. Missouri, 420 U.S. 162 (1975)
16. Robey A: Criteria for competency to stand trial: a checklist for psychiatrists. N Engl J Med 122:616–22, 1965
17. McGarry AL: Competency to Stand Trial and Mental Illness. Washington DC: National Institute of Mental Health, 1973
18. Group for the Advancement of Psychiatry: Misuse of Psychiatry in the Criminal Courts: Competency to Stand Trial (Formulated by the Committee on Psychiatry and Law). New York: GAP, 1974
19. Winick BJ: Reforming incompetency to stand trial and plead guilty: a restated proposal and a response to Professor Bonnie. J Crim Law Criminol 85:571–624, 1995
20. Burt MN, Philipsborn JT: Assessment of client competence: a suggested approach. Champion 22:26:55–8, 1998
21. Godinez v. Moran, 509 U.S. 389 (1993)
22. Cooper v. Oklahoma, 517 U.S. 348 (1996)
23. United States v. Duhon, 104 F.Supp.2d 663 (W.D. La. 2000)
24. Texas Code Crim. Proc. Ann. Art. 46B.024. Available at http://tlo2.tlc.state.tx.us/statutes/cr.toc.htm. Accessed September 9, 2007
25. Utah Code Ann. Available at http://le.utah.gov/%7Ecode/code.htm. Accessed August 10, 2007
26. Bonnie RJ: The competence of criminal defendants with mental retardation to participate in their own defense. J Crim Law Criminol 81:419–46, 1990
27. Cronbach LJ, Meehl PE: Construct validity in psychological tests. Psychol Bull 52:281–302, 1955
28. Golding SL, Roesch R, Schreiber J: Assessment and conceptualization of competency to stand trial: preliminary data on the interdisciplinary fitness interview. Law Hum Behav 8:321–34, 1984
29. Grisso T: Legally relevant assessments for legal competencies, in Evaluating Competencies: Forensic Assessments and Instruments (ed 2). New York: Kluwer Academic/Plenum Publishers, 2003, pp 21–40
30. Golding SL, Roesch R: Competency for adjudication: an international analysis, in Law and Mental Health: International Perspectives (vol 4). Edited by Weisstub DN. New York: Pergamon, 1988, pp 73–109

31. American College of Radiology: Illustrated Breast Imaging Reporting and Data System (BI-RADSJ) (ed 3). Reston, VA: American College of Radiology, 1998
32. Barlow WE, Lehman CD, Zheng Y, *et al*: Performance of diagnostic mammography in women with signs or symptoms of breast cancer. J Natl Cancer Inst 94:1151–9, 2002
33. Linver MN, Osuch JR, Brenner RJ, *et al*: The mammography audit: a primer for the mammography quality standards act (MQSA). Am J Roentgenol 165:19–25, 1995
34. Cooper VG, Zapf PA: Predictor variables in competency to stand trial decisions. Law Hum Behav 27:423–36, 2003
35. Mankad MV, Brakel SJ, Wilson RM: Commentary: incorporation of competence instruments into clinical practice. J Am Acad Psychiatry Law 30:483–5, 2002
36. Poythress NG, Nicholson R, Otto RK, *et al*: Professional manual for the MacArthur Competence Assessment Tool-Criminal Adjudication. Odessa, FL: Psychological Assessment Resources, 1999
37. Hanley JA: The robustness of the "binormal" assumptions used in fitting ROC curves. Med Decis Making 8:197–203, 1988
38. Somoza E, Mossman D: ROC curves and the binormal assumption. J Neuropsychiatry Clin Neurosci 3:436–9, 1991
39. Mossman D, Somoza E: Maximizing diagnostic information from the dexamethasone suppression test: an approach to criterion selection using receiver operating characteristic analysis. Arch Gen Psychiatry 46:653–60, 1989
40. Swets J: Measuring the accuracy of diagnostic systems. Science 240:1285–93, 1988
41. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29–36, 1982
42. Rice ME, Harris GT: Violent recidivism: assessing predictive validity. J Consulting Clin Psychol 63:737–48, 1995
43. Douglas KS, Ogloff JRP, Nicholls TL, *et al*: Assessing risk for violence among psychiatric patients: the HCR-20 risk assessment scheme and the Psychopathy Checklist: Screening Version. J Consult Clin Psychol 67:917–30, 1999
44. Metz CE, Goodenough DJ, Rossmann K: Evaluation of receiver operating characteristic curve data in terms of information theory with applications in radiology. Radiology 108:297–303, 1973
45. Somoza E, Mossman D: Comparing and optimizing diagnostic tests: an information-theoretical approach. Med Decis Making 12:179–88, 1992
46. Somoza E, Mossman D: Comparing diagnostic tests using information theory: the INFO-ROC technique. J Neuropsychiatry Clin Neurosci 4:214–19, 1992
47. Mossman D, Somoza E: Diagnostic tests and information theory. J Neuropsychiatry Clin Neurosci 4:95–98, 1992
48. Nicholson RA, Kugler KE: Competent and incompetent criminal defendants: a quantitative review of comparative research. Psychol Bull 109:355–370, 1991
49. Warren J, Murrie D, Stejskal W: Opinion formation in evaluating the adjudicative competence and restorability of criminal defendants: a review of 8000 evaluations. Behav Sci Law 24:113–32, 2006
50. Hubbard KL, Zapf PA, Ronan KA: Competency restoration: an examination of the differences between defendants predicted restorable and not restorable to competency. Law Hum Behav 27: 127–39, 2003
51. Warren JI, Rosenfeld B, Fitch WL: Beyond competence and sanity: the influence of pretrial evaluation on case disposition. Bull Am Acad Psychiatry Law 22:379–88, 1994
52. Goldstein RL, Stone M: When doctors disagree: differing views on competency. Bull Am Acad Psychiatry Law 5:90–7, 1977
53. Poythress NG, Stock HV: Competency to stand trial: a historical review and some new data. Psychiatry Law 8:131–46, 1980
54. Roesch R, Golding SL: Competency to Stand Trial. Urbana, IL: University of Illinois Press, 1980
55. Skeem JL, Golding SL, Cohn NB, *et al*: Logic and reliability of evaluations of competence to stand trial. Law Hum Behav 22: 519–47, 1998
56. Roesch R, Zapf PA, Golding SL, *et al*: Defining and assessing competency to stand trial, in Handbook of Forensic Psychology (ed 2). Edited by Weiner IB, Hess AK. New York: John Wiley & Sons, 1999, pp 327–49
57. Hoge SK, Poythress N, Bonnie R, *et al*: Mentally ill and nonmentally ill defendants' abilities to understand information relevant to adjudication: a preliminary study. Bull Am Acad Psychiatry Law 24:187–97, 1996
58. Hart SD, Hare RD: Predicting fitness to stand trial: the relative power of demographic, criminal, and clinical variables. Forensic Rep 5:53–65, 1992
59. Cox ML, Zapf PA: An investigation of discrepancies between mental health professionals and the courts in decisions about competency. Law Psychol Rev 28:108–31, 2004
60. Zapf PA, Hubbard KL, Cooper VG, *et al*: Have the courts abdicated their responsibility for determination of competency to stand trial to clinicians? J Forensic Psychol Pract 4:27–44, 2004
61. Henkelman RM, Kay I, Bronskill MJ: Receiver operator characteristic (ROC) analysis without truth. Med Decis Making 10: 24–9, 1990
62. Beiden SV, Campbell G, Meier KL, *et al*: On the problem of ROC analysis without truth: the EM algorithm and the information matrix. Proc SPIE 3981:126–34, 2000
63. Zhou XH, Castelluccio P, Zhou C: Nonparametric estimation of ROC curves in the absence of a gold standard. Biometrics 61: 600–9, 2005
64. Albert PS: Random effects modeling approaches for estimating ROC curves from repeated ordinal tests without a gold standard. Biometrics 63:593–602, 2007