Lecture 16: Measureing Accuracies of Diagnostic Tests

Xiao-Hua Andrew Zhou

Department of Biostatistics, University of Washington

Hierarchical model in assessment of the usefulness of a diagnostic test

- Level 1, at the bottom, is technical efficacy, as measured by such features as image resolution and sharpness for radiographic tests.
- Level 2 is diagnostic accuracy efficacy, i.e. sensitivity, specificity, and the ROC curve.
- Level 3 is diagnostic thinking efficacy; it can be measured, for example, by measuring the difference in the clinician's estimated probability of a diagnosis before vs. after the test results are known.
- Level 4 is therapeutic efficacy and can be measured by the percentage of times therapy, planned before the diagnostic test, is altered by the results of the test.
- Level 5 is patient outcome efficacy, as defined, for example, by the number of deaths avoided due to the test information, the change in the quality of life due to the test information, or the number of patients needed to be treated in order to present one event.
- item Level 6, the top level, is societal efficacy, which is often described by the cost-effectiveness of the test as measured from a societal perspective.

Hierarchical model in assessment of a diagnostic test, cont

- A key feature of the model is that in order for a diagnostic test to be efficacious at a higher level, it must be efficacious at all lower levels.
- The reverse is not true, i.e. a test can be efficacious at one level but it doesn’t guarantee that it will be efficacious at higher levels.
- In this talk we deal exclusively with the assessment of diagnostic accuracy efficacy (level 2), recognizing that it is only one step in the complete assessment of the usefulness of a diagnostic test.

Intrinsic Accuracy

- The intrinsic accuracy of a test is measured by comparing the test results to the true condition status.
- Assume true condition status is one of two mutually exclusive states: “the condition is present” or “the condition is absent”.
- We determine the true disease status by the mens of a gold standard.
- **Gold standard** is source of information, completely different from tests under evaluation, which tells true condition status of patient.
- Some of common examples of the gold standard are autopsy reports, surgery findings, pathology results from biopsy specimens, and the results of other diagnostic tests.
- Once a test is shown to have some level of intrinsic accuracy, we consider not only intrinsic accuracy of test but also prevalence and nature of disease, patient characteristics, and consequences of test’s misdiagnoses.

Sensitivity and specificity

- Two basic measures of diagnostic accuracy are sensitivity and specificity.
- Sensitivity: test's abilities to correctly detect condition when condition is actually present.
- Specificity: test's ability to correctly rule out condition when it is truly absent.

Basic 2x2 Count Table

	Test Result:		
True Condition Status:	*Positive* (T=1)	*Negative* (T=0)	*total*
Present (D=1)	s_1	s_0	n_1
Absent (D=0)	r_1	r_0	n_0
total	m_1	m_0	N

$$Se = P(T = 1 \mid D = 1) = s_1/n_1$$
$$Sp = P(T = 0 \mid D = 0) = r_0/n_0$$

Results of 30 Patients With and 30 Without Breast Cancer

A mammographer's diagnoses of 60 patients presenting for breast cancer screening (Powell et al, 1999). The study sample consisted of 30 patients with pathology-proven cancer and 30 patients with normal mammograms for two consecutive years.

	Test Result:		
Cancer Status:	*Positive*	*Negative*	*total*
Present	29	1	30
Absent	19	11	30
total	48	12	60

Need Clear Definitions

- The definition of “positive” and “negative” test results, as well as the condition of interest, must be clear.
- Example: in a study of lung disease (Remer et al, 1999), patients with detected adrenal adenomas were called “positive”, while patients with detected lung metastases were called “negative”.

Gap Measurements of 10 Patients With and 10 Without Fractured Heart Valve

- Many diagnostic tests yield numeric measurement as a result.
- Consider digital imaging algorithm to identify patients whose implanted artificial heart valve has fractured (Powell et al, 1996).

Fractured	Intact
0.58	0.13
0.41	0.13
0.18	0.07
0.15	0.05
0.15	0.03
0.10	0.03
0.07	0.03
0.07	0.00
0.05	0.00
0.03	0.00

Estimated Sens and Spec

Table 1: Estimates of Se and Sp From Heart Valve Imaging Study

Defn of + Test	Se	Sp	FNR	FPR
> 0.58	0.0	1.0	1.0	0.0
> 0.13	0.5	1.0	0.5	0.0
> 0.07	0.6	0.8	0.4	0.2
> 0.05	0.8	0.7	0.2	0.3
> 0.03	0.9	0.6	0.1	0.4
> 0.0	1.0	0.3	0.0	0.7
≥ 0.0	1.0	0.0	0.0	1.0

As Se increases, Sp decreases.

Types of Decision Thresholds

- Gap measurement is objective test result
- Other tests yield results that must be subjectively interpreted. Observer establishes decision threshold in his/her mind.
- Example: Ask the mammographer to use stricter decision threshold to increase his specificity. Reread 60 cases.

Confidence Scales

- Mammographer assigns confidence score to each case to reflect belief the patient has condition.
- Ordinal (rating) scale: the condition is “definitely not present”, “probably not present”, “possibly present”, “probably present”, and “definitely present”.
- Percent confidence scale: 0% to 100% scale.
- Certain tests have specialized scale. Mammography: “normal”, “benign”, “probably benign”, “suspicious”, and “malignant”.

Mammogram Results Using 5-Category Scale

	Test Result:					
Cancer Status:	*Normal*	*Benign*	*Probably Benign*	*Suspicious*	*Malignant*	*Total*
Present	1	0	6	11	12	30
Absent	9	2	11	8	0	30

Intrinsic properties

- Sensitivity and specificity are not affected by prevalence of condition because
  - sensitivity is computed from only the subjects with the condition,
  - whereas specificity is computed from the subsample of patients without the condition.
- This property of sensitivity and specificity is important; in practical terms, it means the sensitivity and specificity estimated from a study sample are applicable to other populations with different prevalence rates.

Spectrum of Disease

- Sensitivity and specificity are not affected by prevalence of condition.
- Sensitivity and specificity of some diagnostic tests are affected by **spectrum of disease**.
- Spectrum of a disease refers to disease's range of clinical severity or range of anatomic extent.
- For example, large, palpable breast cancer tumors are easier to detect than sparse, dispersed malignant calcifications; thus mammmography has greater sensitivity when it is applied to patients with advanced patients.
- Similarly, patient characteristics can affect the sensitivity and specificity of some diagnostic tests. Older women have fatty, less dense breasts than younger women, and mammography is better able to detect lesions in fatty breasts.

Combined Measures of Se and Sp

- Often useful to summarize accuracy of test by a single number. Example: when comparing two tests.
- Popular measure often referred to simply as “accuracy”. Really just probability of a correct test result: $(s_1 + r_0)/N$.
- $Se \times P(D = 1) + Sp \times P(D = 0)$.
- 1885 editorial by Gilbert about extremely high “accuracy” of fellow meteorologist in predicting tornadoes simply by calling for “no tornado” every day.
- Other limitations: Based on only one decision threshold
- Treats false positive and false negative results as if equally undesirable

Other Combined Measures- Odds Ratio

- $$Odds\ Ratio = \frac{Se/(1-Se)}{(1-Sp)/Sp} = \frac{Se \times Sp}{FNR \times FPR}.$$
- An odds ratio of 1 indicates the odds of likelihood of a positive test result is the same for patients with and without the condition.
- An odds ratio of greater than 1 indicates the odds of likelihood of a positive test result is greater for patients with the condition.
- An odds ratio of less than 1 indicates the odds of likelihood of a positive test result is greater for patients without the condition.

Other Combined Measures- Youden's Index

- **Youden's index**: Se+Sp-1, or, Se-FPR.
- It has a maximum value of 1.0 and a minimum value of 0.0 when the accuracy of the test is reasonable (e.g. ROC curve is a concave function).

Properties of odds ratio and Youden’s index

- They are not dependent on the prevalence of the condition in the sample
- They share the same limitation as the ‘accuracy”.

Receiver Operating Characteristic Curve

- Describes intrinsic accuracy of a test apart from decision thresholds
- Each point on graph generated by different decision threshold
- Use line segments to connect points from all possible decision thresholds; this forms **empirical ROC curve**.
- **Fitted ROC curves** (**smooth curves**) formed by fitting statistical model to test results. Binormal distribution (i.e. two Gaussian distributions)

ROC curve, cont

- Curves constructed from objective measurements of a test (e.g. gap value from digitized image of heart valve), objective evaluation of image features (e.g. attenuation coefficient from computed tomography), or subjective diagnostic interpretations.
- Essential assumption is that decision thresholds are the same for the subsamples of patients with and without the condition.

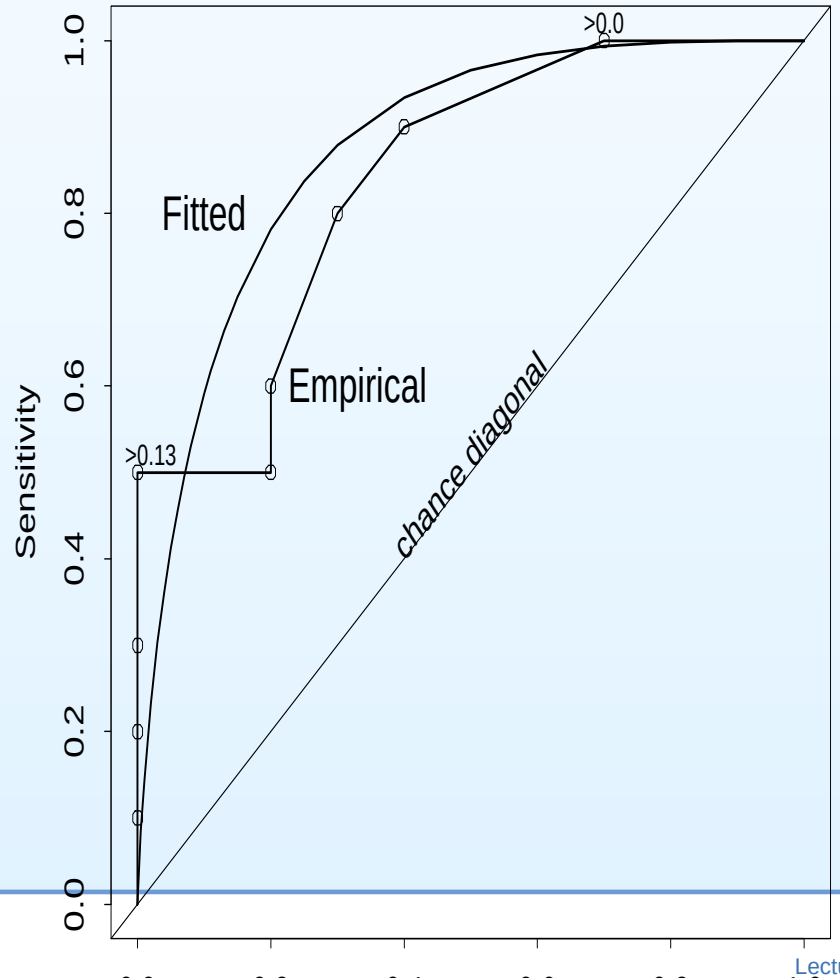
An example

- In a study, readers looked at 58 mammograms, 13 of whom had a malignant lesion in the right breast and 45 of whom did not.
- All diagnoses were confirmed by either biopsy or a follow-up of two year.
- Readers gave a BIRAD score.

Result of Mammography	Malignant	Normal or benign
1, normal	22	1
2, benign	8	0
3, probably benign	7	1
4, suspicious	8	11
5, malignant	0	0
Total	45	13

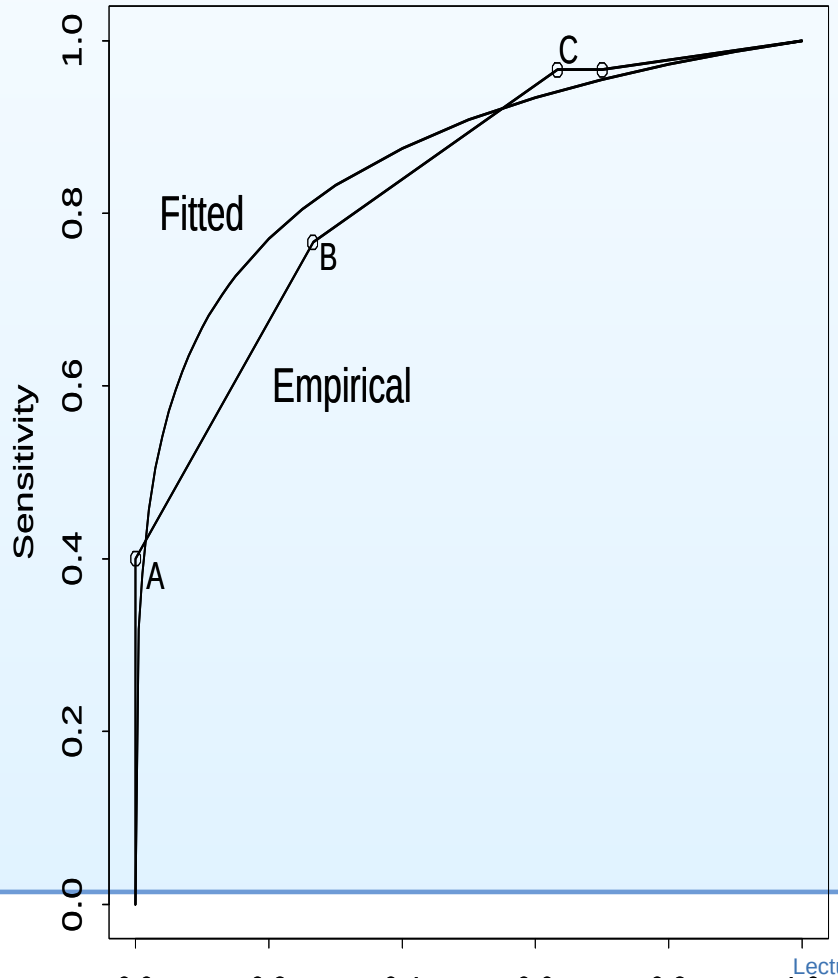
Empirical and Fitted ROC Curve for Heart Valve Imaging

Figure 2.2



Empirical and Fitted ROC Curve for Mammography

Figure 2.3



Advantages of ROC curve

- ROC curve is visual representation of accuracy data. Scales of curve are the basic measures of accuracy.
- Does not require selection of a particular decision threshold.
- Independent of prevalence. May be affected by spectrum of disease, as well as patient characteristics. Example: test for fetal pulmonary maturity; the ROC curve strongly affected by gestational age (Hunink et al, 1990).
- Does not depend on scale of test results. Empirical curve depends only on ranks of observations
- Provides direct visual comparison of two or more tests on common set of scales.

Relation between odds ratio and ROC curves

- HC Kraemer (2004). Reconsidering the odds ratio as a measure of 2x2 association in a population. Stat Med. 2004 Jan 30;23(2):257-70.
- The odds ratio (OR) is probably the most widely used measure of 2x2 association in epidemiology, but it often produces results that are puzzling or misleading.
- Receiver operating characteristic (ROC) methods are used to take a fresh look at the OR and show where and why such puzzling results arise.
- When researchers choose to report a summary measure of association, the OR is one of many measures of association that might be considered, not one that should be considered the ’gold standard’ of 2x2 measures of association.
- In a randomized clinical trial with binary outcome for success, either the success or failure rates in treatment and control groups might be reported separately or the number needed to treat to achieve one extra success, to emphasize the cost of unnecessary treatment needed to achieve a success.
- In studies assessing reliability or heritability, we recommend the intraclass kappa. In studies in which one binary variable is assessed against a binary criterion, we recommend the weighted kappa.

Area Under ROC Curve

- ROC area can take on values between 0.0 and 1.0 (practically, 0.5 to 1.0)
- several interpretations:
  - the average value of sensitivity for all possible values of specificity,
  - the average value of specificity for all possible values of sensitivity, and
  - the probability that a randomly selected patient with the condition has a test result indicating greater suspicion than a randomly chosen patient without the condition.

AUC

Bamber (1975) pointed out that area under empirical ROC curve is equivalent to quantity obtained when one performs the Mann-Whitney version of the two-sample rank-sum statistic of Wilcoxon.

Area Under ROC Curve for 2 examples

- Mammography Example: empirical curve area is 0.83 (fitted curve 0.86)
- GAP vs. OFFSET: fitted curves: 0.87 vs. 0.65.

Limitations of ROC curve Area

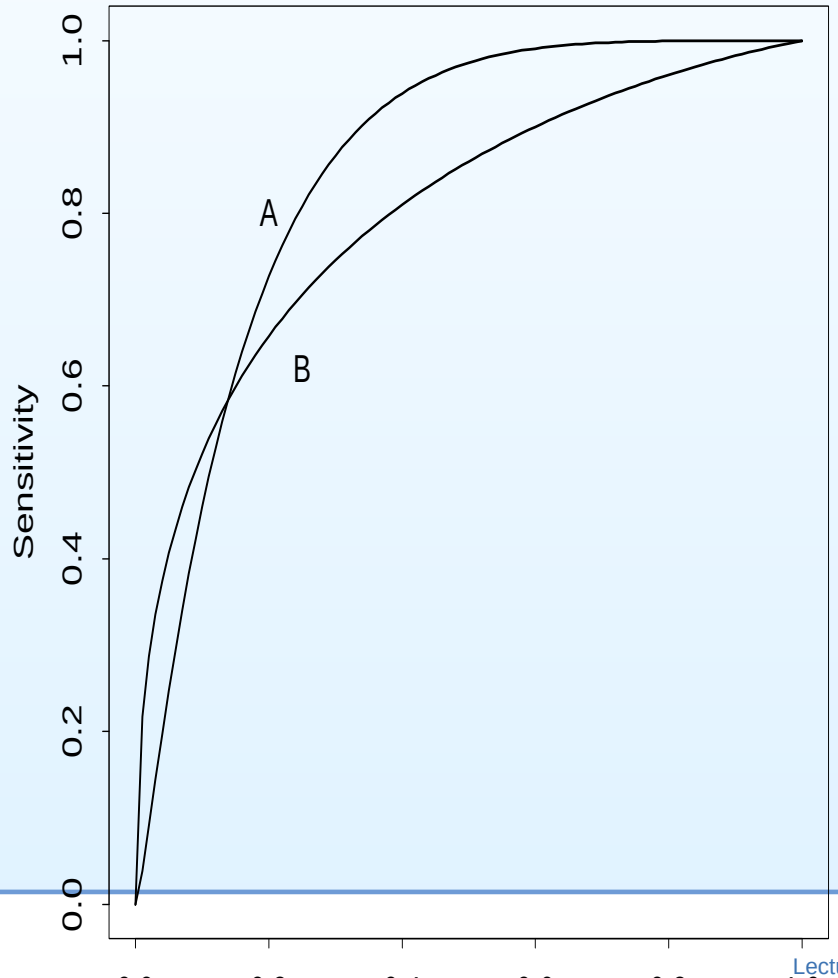
- Once test has been shown to distinguish well, its role for particular applications must be evaluated.
- Example: if we use heart valve imaging technique to screen asymptomatic patients, interested in the part of the ROC curve where the specificity is high.
- ROC area, because it is global measure of intrinsic accuracy, is not always relevant
- May be misleading when comparing the accuracy of two tests; when this is case, the study protocol should be expected to address this issue.

Limitations of ROC areas, continued

- Similarly, the ROC area may be misleading when comparing the accuracy of two tests.
- The ROC areas of two tests may be equal but the tests may differ in clinically important regions of the curve.
- Likewise, the ROC areas may differ but the tests may have the same area in the clinically relevant region of the curve.
- Figure 2.6 below illustrates two ROC curves that cross at a FPR of 0.14.
- The area under the A curve is greater than the area under the B curve (i.e. 0.85 vs. 0.80).
- If the clinically relevant region of the curve is at low FPRs, test B is preferrable to test A even though the ROC area is greater for A than B.

Two Tests With Crossing ROC Curves

Figure 2.6



Two alternative summary measures

- Next we present two alternative summary measures of intrinsic accuracy that focus on only a portion of the ROC curve, thus overcoming the main limitation of the area under the whole curve.
- Sensitivity at a fixed FPR and partial Area under ROC curves.

SENSITIVITY AT FIXED FPR

- An alternative summary measure of intrinsic accuracy is the **sensitivity at a fixed FPR**, or similarly the FPR at a fixed sensitivity.
- We write this $Se_{(FPR=e)}$ or $FPR_{(Se=e)}$. For a predetermined FPR of e (or predetermined sensitivity of e), the sensitivity (or FPR) is estimated from the ROC curve.
- The sensitivity at a fixed FPR is preferable to the ROC area when evaluating a test for a particular application. This measure also has a simple and clinically useful interpretation.
- One disadvantage to this measure is that reported sensitivities from other studies are often at different FPRs, thus comparisons with published literature can be problematic.
- A second limitation is that published reports are not always clear about whether the FPR was selected before the start of the study (as it should be) or after the data were examined (a practice which can introduce bias).
- Third, the statistical reliability of this measure is lower (i.e. the variance is larger) than that of the ROC area.

Partial Area Under ROC Curve

- Another summary measure of intrinsic accuracy is the **partial area under the ROC curve**. As its name implies, it is the area under a portion of the ROC curve. It is often defined as the area between two FPRs, e_1 and e_2. We write this: $A_{(e_1 \leq FPR \leq e_2)}$.
- If $e_1 = 0$ and $e_2 = 1$, then the area under the entire ROC curve is specified. If $e_1 = e_2$, then the sensitivity at a fixed FPR of e (or FPR at a fixed sensitivity of e) is given.
- The partial area measure is thus a compromise between the ROC area and the sensitivity at a fixed FPR.

Partial Area Under ROC Curve, cont

- Like the sensitivity at a fixed FPR index, the partial area allows one to focus on the portion of the ROC curve relevant to a particular clinical application.
- In Figure 2.4, if we restrict to a FPR range of 0.0-0.05, the partial area for offset is slightly larger than for gap, though not statistically significant, 0.0139 versus 0.0126.
- If we include larger FPRs, e.g. 0.0-0.20, then the partial area for gap (0.108) is larger than for offset (0.080).

Partial Area Under ROC Curve, cont

- To interpret the partial area we must consider its maximum possible value. The maximum area is equal to the width of the interval, i.e. $(e_2 - e_1)$.
- McClish (1989) and Jiang (1996) recommend standardizing the partial area by dividing by its maximum value. Jiang et al refer to this standardized partial area as the **partial area index**.
- The partial area index is interpreted as the average sensitivity for the range of specificities examined (or average specificity for the range of sensitivities examined).
- This interpretation is quite useful clinically. For the heart valve imaging example, the average sensitivities in the FPR range of 0.0-0.20 are 0.54 and 0.41 for gap and offset, respectively.

Partial Area Under ROC Curve, cont

- Dwyer (1997) offers a probabilistic interpretation of the partial area index when the partial area is defined for sensitivities greater than e_1 (i.e. $A_{(e_1 \leq TPR \leq 1.0)}$).
- The partial area index equals the probability that a randomly chosen patient without the condition will be correctly distinguished from a randomly chosen patient with the condition who tested negative for the criterion that corresponds to TPR=e_1. Note the similarities between this and the probabilistic interpretation of the ROC area.

Limitations of Partial ROC Areas

- A potential problem with the partial area measure is that the minimum possible value depends on the location along the ROC curve.
- The minimum partial area is equal to $(1/2)(e_2 - e_1)(e_2 + e_1)$ [**?**]. For example, the minimum value for $A_{(0 \leq FPR \leq 0.2)}$ is 0.02 (maximum value is 0.20) and the minimum value for $A_{(0.8 \leq FPR \leq 1.0)}$ is 0.18 (maximum value is 0.20).
- Suppose that we estimated a partial area of 0.19 for both of these FPR ranges; the partial area index is the same for both ranges: 0.95. However, we would probably not value these two areas the same.
- To remedy this problem, McClish offers a transformation of the partial area to values between 0.5 and one. The formula is

$$\frac{1}{2}[1 + \frac{A_{(e_1 \leq FPR \leq e_2)} - min}{max - min}] \quad (1)$$

  where min and max are the minimum and maximum possible values for the partial area.
- Continuing with this example, the partial area of 0.19 is transformed to 0.972 for the 0-0.2 FPR range and 0.75 for the 0.8-1.0 FPR range.

Limitations

- The partial area measure has similar limitations to the sensitivity at a fixed FPR.
- First, it is difficult to compare this measure with the published literature if different ranges are used.
- Second, the relevant range should be specified apriori; it is not always clear from published reports whether this occurred.
- Lastly, the statistical reliability of this measure is lower than that of the ROC area, but is greater than that of the sensitivity at a fixed FPR.

Localization and detection of multiple abnormalities

- Some diagnostic tasks are more complicated than simple detection of a single occurrence of the condition.
- For example, in mammography patients can have multiple lesions; these lesions must be correctly located prior to follow-up procedures like biopsy and surgery.
- Another example is the detection of infarcts in patients suspected of having a stroke. A patient can have multiple infarcts, and it is critical that they be detected and located in the correct brain hemisphere.

Time-dependent ROC curves

- T: time to an event
- Case at time t: $T < t$,
- Non-case at time t: $T > t$.
- M: biomarker.
- $Sens(c,t) = P(M \geq c \mid T < t)$,
  $Specs(c,t) = P(M < c \mid T > t)$.