Introduction to screening tests

Tim Hanson
Department of Statistics
University of South Carolina
April, 2011

Overview:

1. Estimating test accuracy: dichotomous tests.
2. Estimating test accuracy: continuous tests.
3. Does adding additional tests help?
4. Statistical research: modeling populations of nondiseased G_0 and diseased G_1 continuous test responses.

- We assume a state loosely termed **diseased** $D+$ or **not diseased** $D-$, but any event of interest works.
- **Examples**:
  - $D+ =$ cardiovascular disease
  - $D+ =$ hepatitis B
  - $D+ =$ Parkinson's disease
  - $D+ =$ recent use of illegal drugs
- Notice shades of gray and differences in these outcomes.
  - Cardiovascular disease is an umbrella term and can be tested for many different ways: exercise stress test, MRI, X-ray, Echocardiogram, CT scan, PET, SPECT, plus various blood tests. Usually diagnosis takes multiple tests into account.
  - Drug use is known to the person being tested!
  - Hepatitis B is either there or not.

Binary tests: result in one of two outcomes, either $T+$ or $T-$.

Examples:

- over the counter pregnancy tests
- rapid strep test
- cultures (either something grows or it doesn't)
- direct microscopic examination of body fluid (either see it or not)
- asking a potential employee if they've recently used illegal drugs

Continuous tests: result in a number Y. Typically as the number increases the likelihood of $D+$ increases.

Examples:

- Enzyme-Linked ImmunoSorbent Assay (ELISA) measures an inferred amount of antigen in a blood sample
- minutes of briskly walking on a treadmill before discomfort
- pathologist classifying a slide as (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma *in situ* (not metastasized), (4) invasive carcinoma (metastasized)

Often a continuous test is made into a binary one by *dichotomizing*:

$$T+ \Leftrightarrow Y > k \text{ and } T- \Leftrightarrow Y \leq k.$$

Binary tests

An individual from a population will fall into one of four categories:

$$(D+, T+), \; (D+, T-), \; (D-, T+), \text{ or } (D-, T-).$$

These are 'true positive', 'false negative', 'false positive', and 'true negative'.

Two common measures of *binary* test accuracy are sensitivity and specificity:

$$Se = P(T+|D+) \;\; Sp = P(T-|D-).$$

- How well does the test do identifying those that really are $D+$? The *sensitivity* of a test, denoted Se, is the probability that a diseased person tests positive.

- How well does the test do identifying those that really are $D-$? The test's *specificity* is the probability that a nondiseased person tests negative.

Note, *gold standard* tests have perfect sensitivity and specificity. For example, western blot test for HIV; culture for strep.

A measure for dichotomized tests that considers sensitivity and specificity over all possible cutoffs k will be discussed shortly.

Example: Rapid strep test

Sheeler et al. (2002) describe a modest prospective trial of $n = 232$ individuals complaining of sore throat who were given the rapid strep (*streptococcal pharyngitis*) test. Each individual was also given a gold standard test, a throat culture.

	D+	D−	Total
T+	44	4	48
T−	19	165	184
Total	63	169	232

	D+	D−	Total
T+	44	4	48
T−	19	165	184
Total	63	169	232

- An estimate of Se is $\widehat{Se} = \widehat{P}(T+|D+) = \frac{44}{63} = 0.70$.
- An estimate of Sp is $\widehat{Sp} = \widehat{P}(T-|D-) = \frac{165}{169} = 0.98$.
- The estimated prevalence of strep among those complaining of sore throat $P(D+)$ is $p = \widehat{P}(D+) = \frac{63}{232} = 0.27$.

If we have a sore throat, and test positive, we may be interested in the probability we have strep

$$\begin{aligned}
P(D+|T+) &= \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)} \\
&= \frac{Se \times p}{Se \times p + (1 - Sp) \times (1 - p)} \\
&\approx \frac{0.70 \times 0.27}{0.70 \times 0.26 + (1 - 0.98) \times (1 - 0.27)} \\
&= 0.92.
\end{aligned}$$

This is called the *predictive value positive* (PVP).

Similarly,

$$\begin{aligned} P(D-|T-) &= \frac{P(T-|D-)P(D-)}{P(T-|D-)P(D-)+P(T-|D+)P(D+)} \\ &= \frac{Sp \times (1-p)}{Sp \times (1-p) + (1-Se) \times p} \\ &\approx \frac{0.98 \times (1-0.27)}{0.98 \times (1-0.27) + (1-0.70) \times 0.27} \\ &= 0.90. \end{aligned}$$

This is called the *predictive value negative* (PVN).

- These four numbers summarize how useful a test T is: sensitivity $P(T+|D+)$, specificity $P(T-|D-)$, positive predictive value $P(D+|T+)$ and negative predictive value $P(D-|T-)$.
- PPV and NPV are tied to how prevalent $P(D+)$ the disease is in the population – useful to an individual.
- Se and Sp not tied to prevalence. Useful for picking a test in terms of cost of making a mistake.
- We ignored variability here and only reported *point estimates*. How reliable these estimates are depends on how many people were sampled. For example, $\widehat{Se} = 0.70$ but a 95% CI is $(0.57, 0.81)$; that's a large range. Similarly, $\widehat{Sp} = 0.97$ with 95% CI $(0.94, 0.99)$.

Comparing tests

Say we have two tests, T_1 and T_2, with:

$$Se_1 = 0.8,\ Sp_1 = 0.99,\ Se_2 = 0.99, Sp_2 = 0.8.$$

Which is better?

It depends which is worse: a false negative or a false positive.

- If a false positive is worse – perhaps resulting in unnecessary surgery or a regimen of pharmaceuticals with harmful side effects – then we want the false positive rate to be as small as possible $\Leftrightarrow$ want specificity to be high. Here we'd pick T_1.
- If a false negative is worse – perhaps letting a toxically diseased (think mad cow) proceed to slaughter, or a home pregnancy test – we want the false negative rate to be as small as possible $\Leftrightarrow$ want sensitivity to be high. Here's we'd pick T_2.

Evaluating continuous tests: ROC Curves

Recall that *dichotomizing* a continuous test Y makes a new binary test T:

$$Y > k \Rightarrow T+ \text{ and } Y \leq k \Rightarrow T-.$$

- Magnitude of the individual test scores ignored $\Rightarrow$ information loss
- Predictive probability of disease is same for *all* $T+$ (or $T-$) individuals regardless of actual test scores
- Subjects w/ very large scores Y are identical to those barely above the cutoff
- BUT, expect probability of disease to be an increasing function of Y...

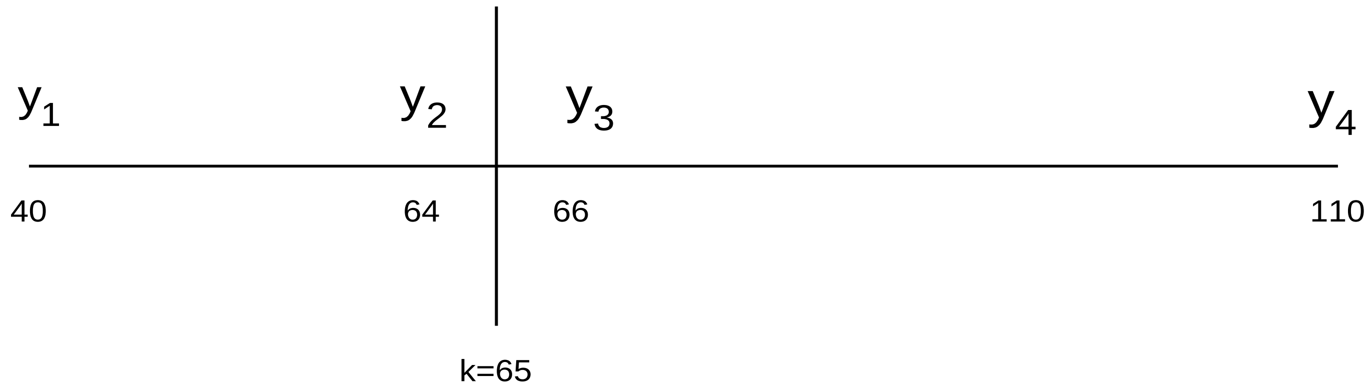

Figure 1: Four serology scores dichotomized using cutoff $k = 65$.

- Individuals 1 & 2 are $T-$; individuals 3 & 4 are $T+$.
- Individuals 1 and 2 $T-$, test scores differ by 24 units.
  Individuals 3 and 4 $T+$, test scores differ by 44 units.
- Individuals 2 and 3 different although differ by only 2 units.

Dichotomizing can oversimplify the analysis but gives easily interpretable parameters: Se, Sp, PVP, and PVN.

Let G_0 and G_1 be distribution of Y from non-diseased and diseased populations.

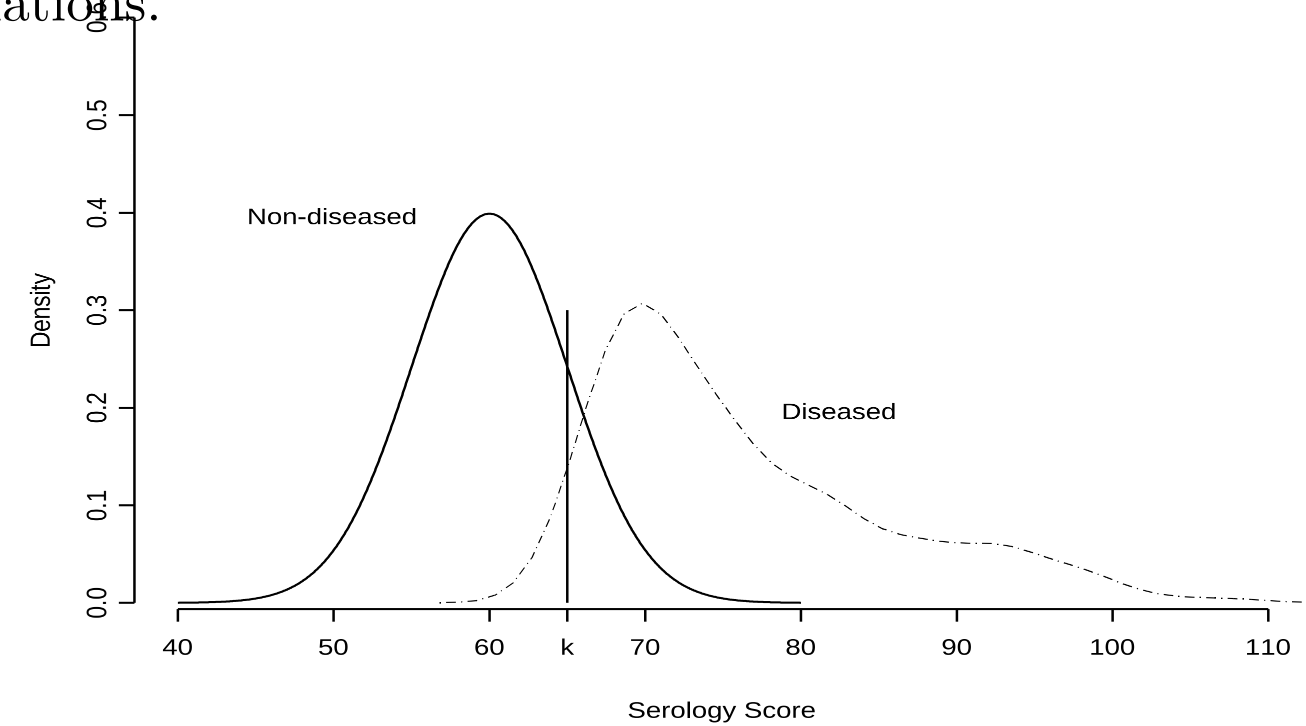

Figure 2: Cutoff $k = 65$ used to dichotomize continuous serology scores distributed according to G_0 (non-diseased) or G_1 (diseased).

The receiver operator characteristic (ROC) curve plots $((1 - Sp(k)), Se(k))$ for all cutoff values k.

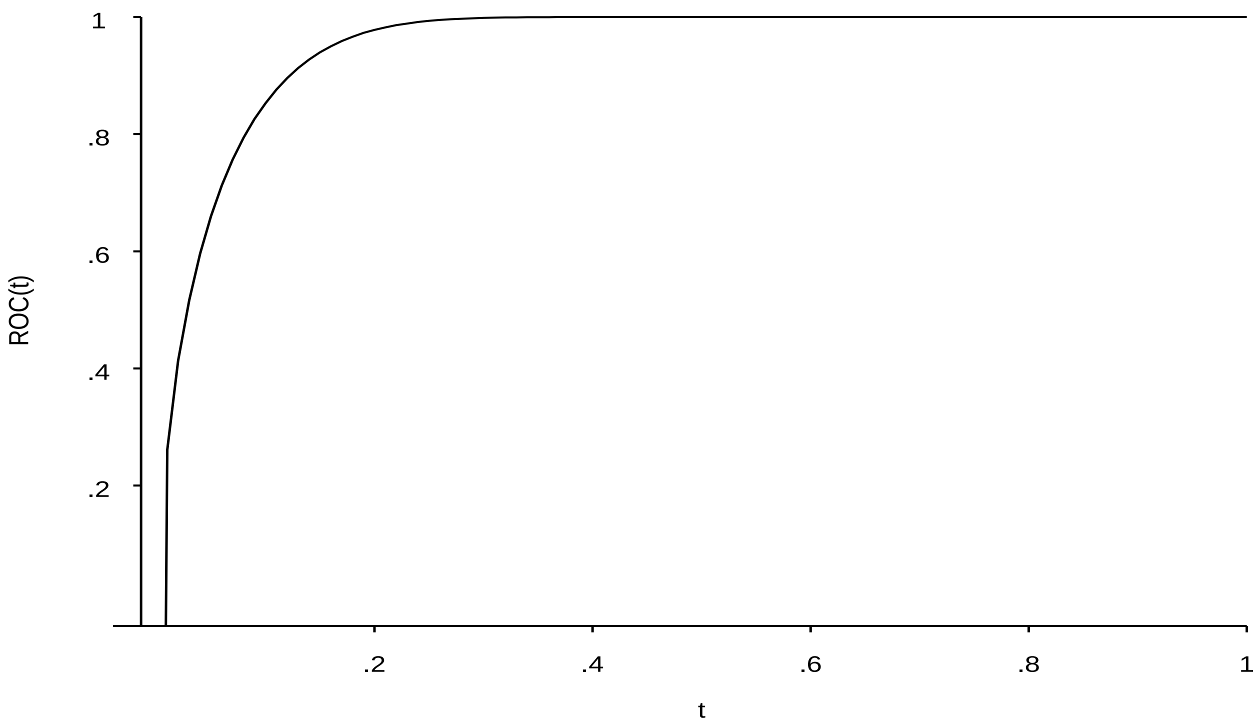

Figure 3: ROC curve corresponding to the distributions G_0 and G_1.

- ROC curve graphically illustrates a continuous test's X usefulness in terms of all error rates.
- Good tests have $Se(k)$ close to one and $1 - Sp(k)$ close to 0 for most k – translates into a concave curve with area underneath close to one.
- Area under the curve (AUC) is measure of tests overall diagnostic accuracy. Often reported in publications.
- The AUC is the probability of an infected having a larger Y than a non-infected – for reasonable tests, this should be larger than 0.5.

Example: A newly developed continuous measure $T_{1\rho}$ is derived from an MRI scan.

It is postulated that $T_{1\rho}$ is related to neuronal loss. This loss is focused in the substantia nigra part of the brain in Parkinson's disease (PD) patients.

- Case/control study looked at 9 PD patients (PD=1) and 10 controls (PD=0). $T_{1\rho}$ measured on all 19 subjects. (Other covariates also recorded: UPSIT (smell), age, etc.)
- Of interest is to determine if significant differences exist between the PD=0 and PD=1 groups. Let's look at a dotplot. $T_{1\rho}$ tends to be higher (more neuronal loss) in PD group.
- A t-test gives $p = 0.000$ on 18 df. We strongly reject $H_0 : \mu_0 = \mu_1$. That is, average $T_{1\rho}$ values are different in PD=0 and PD=1 groups.

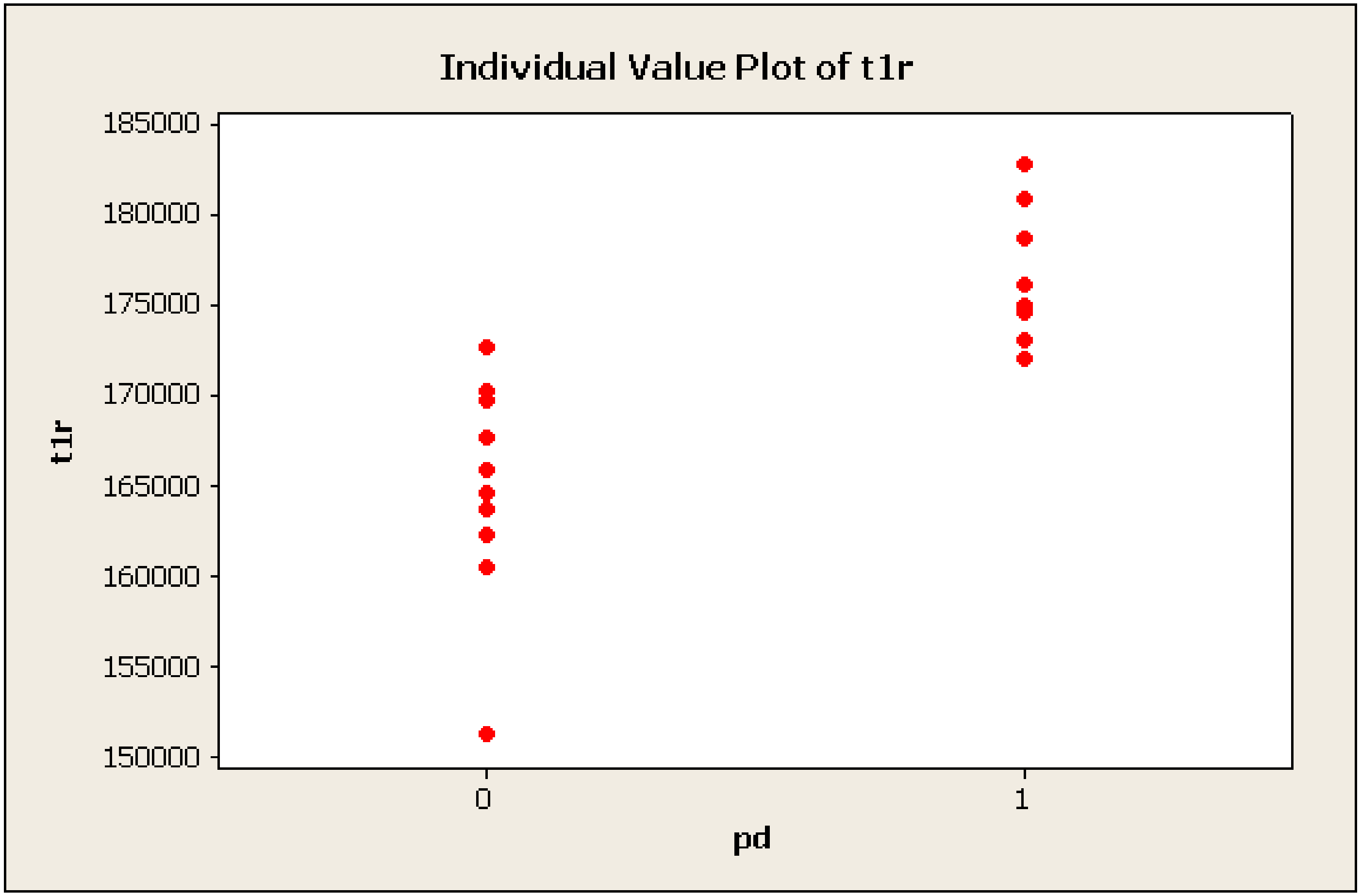

Figure 4: $T_{1\rho}$ values for Parkinson's and control.

Let's define a formal *binary* test based on $k = 172,500$.

	PD+	PD−	Total
$T_{1\rho}+$	7	1	8
$T_{1\rho}-$	1	9	10
Total	8	10	19

$k = 172,500 \Rightarrow \hat{Se} = 7/8 \approx 0.88$ and $\hat{Sp} = 9/10 = 0.90$.

If instead $k = 170,000$ we get

	PD+	PD−	Total
$T_{1\rho}+$	8	2	10
$T_{1\rho}-$	0	8	8
Total	8	10	19

Our estimates change to $\hat{Se} = 1.00$ and $\hat{Sp} = 0.80$.

These are small sample sizes! Variability? Se CI is $(0.72, 1.00)$, Sp $(0.52, 0.96)$ – almost useless.

Note that: estimates of Se and Sp change w/ k, written $Se(k)$ and $Sp(k)$.

Let's look at an estimated ROC curve. $\widehat{\text{AUC}} = 0.989$ w/ CI $(0.84, 1.00)$ from http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html

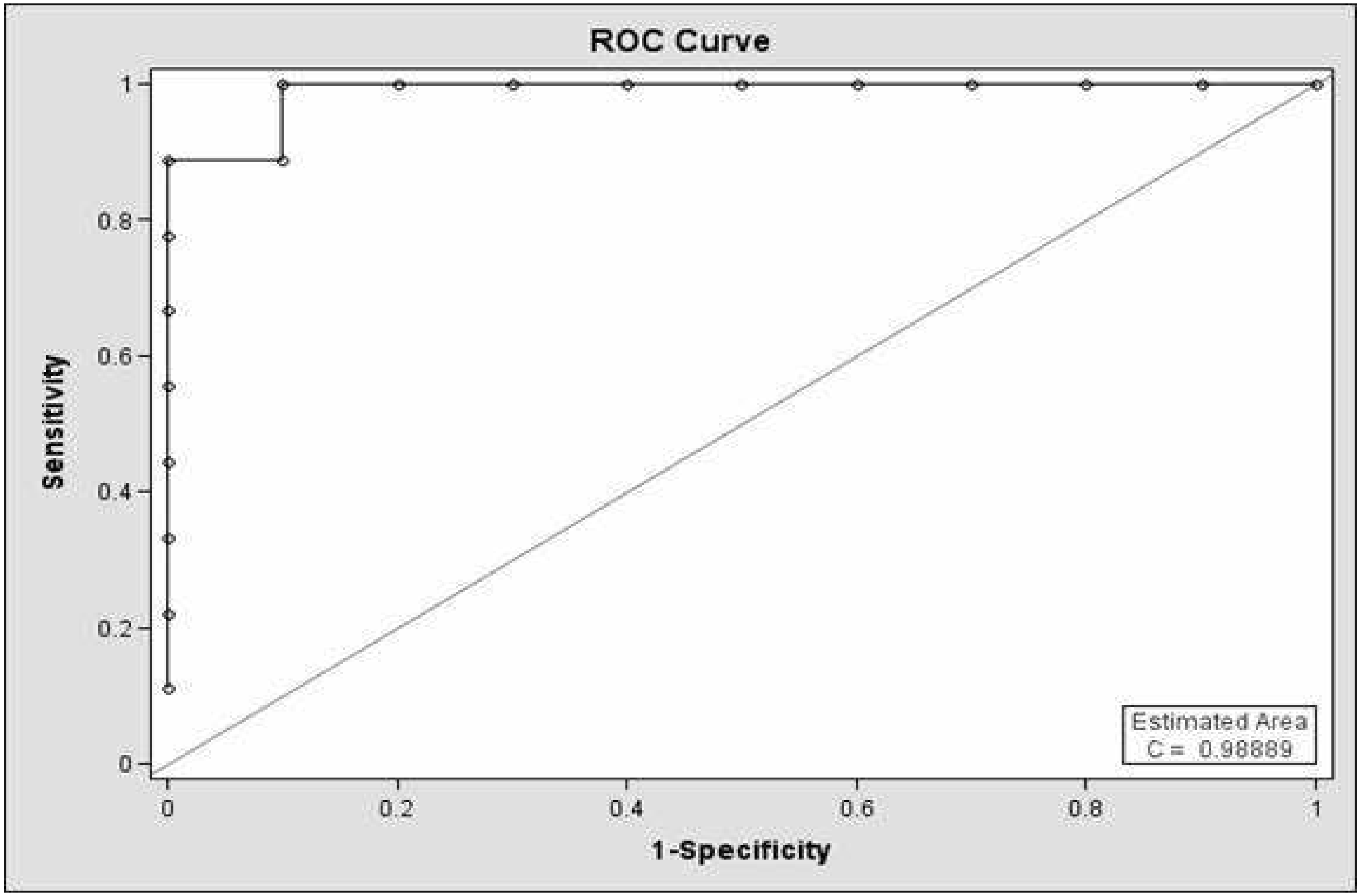

Figure 5: Empirical (nonparametric) ROC curve for $T_{1\rho}$.

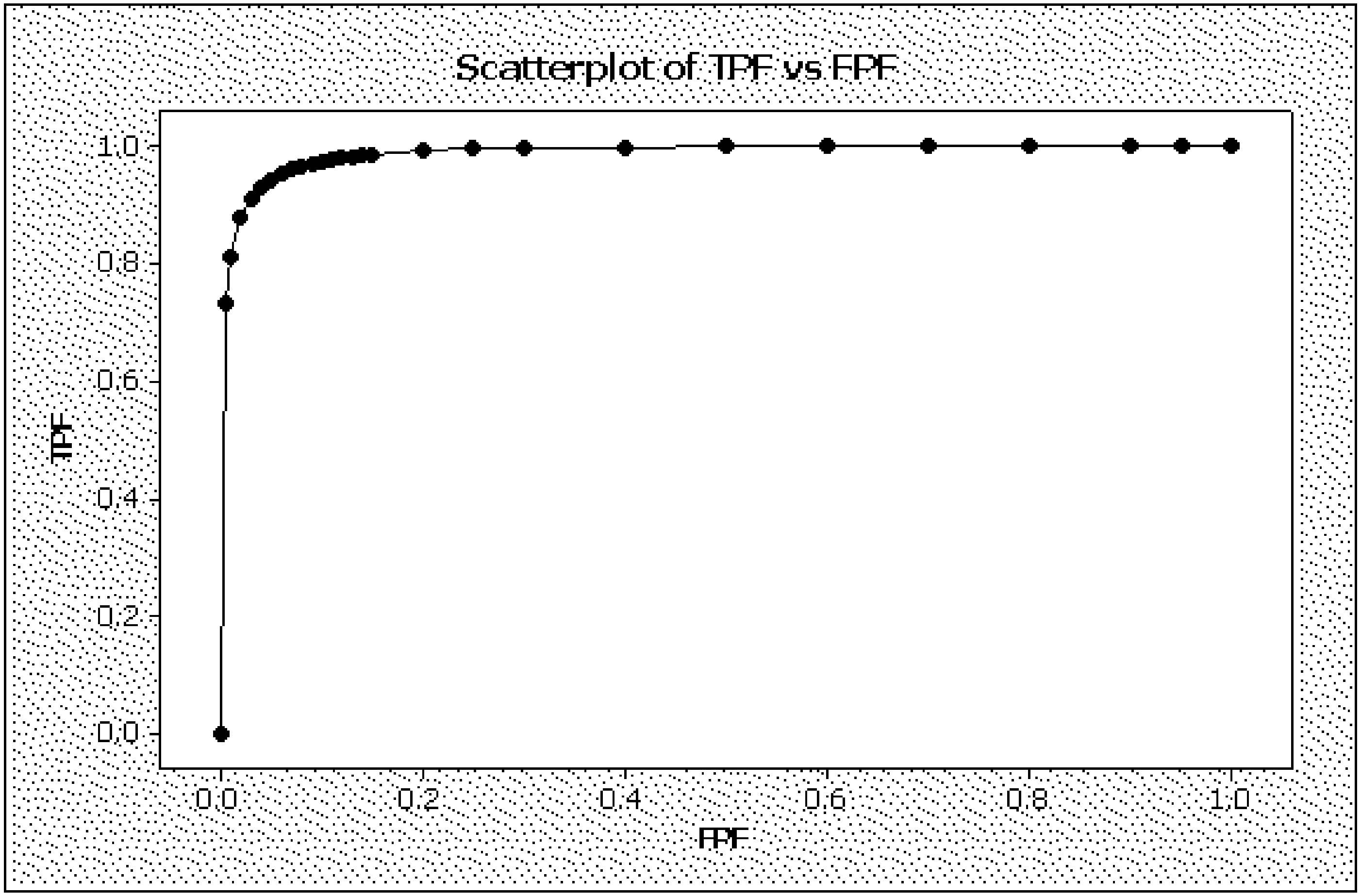

Figure 6: Parametric ROC curve for $T_{1\rho}$.

ROC curve interpretation:

- A test that perfectly discriminates between non-diseased and diseased individuals has ROC curve $\mathrm{ROC}(t) = 1$ w/ AUC $= 1$. G_0 and G_1 are completely separated.
- Diagnostic tests that are equivalent to a coin toss, and hence are worthless, have $\mathrm{ROC}(t) = t$ w/ AUC $= 0.5$. G_0 and G_1 are the same.
- Typically there's some overlap between G_0 and G_1 and the ROC will show this.
- The $T_{1\rho}$ has AUC close to one; based on limited information this measure shows real promise as a diagnostic tool.